Department for Education's generative artificial intelligence in education call for evidence: written submission

August 2023

About the Ada Lovelace Institute

The Ada Lovelace Institute was established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

The mission of the Ada Lovelace Institute is to ensure that data and AI work for people and society. We believe that a world where data and AI work for people and society is a world in which the opportunities, benefits and privileges generated by data and AI are justly and equitably distributed and experienced. Through research, policy and practice, we aim to ensure that the transformative power of data and AI is used and harnessed in ways that maximise social wellbeing and put technology at the service of humanity.

About the Nuffield Foundation

The Nuffield Foundation is an independent charitable trust with a mission to advance educational opportunity and social well-being. Each year the Foundation funds research projects with a total value of over £20 million on issues which inform social policy, primarily in the fields of education, welfare and justice. The Foundation also provides opportunities for young people to develop skills in confidence in science and research. The Foundation is a founder and co-funder of the Nuffield Council on Bioethics, the Nuffield Family Justice Observatory and the Ada Lovelace Institute.

We hope our response will be of interest to the Department for Education and other Government departments in navigating the challenging question of getting AI right for people and society, and would be happy to discuss these points in further detail.

For further information please contact Renate Samson at rsamson@adalovelaceinstitute.org or Kruakae Pothong at kpothong@adalovelaceinstitute.org.

Summary

The Ada Lovelace Institute (Ada) and the Nuffield Foundation welcome the Department for Education's call for evidence on generative artificial intelligence in education. This response provides an overview of potential opportunities, benefits, risks and concerns around uses of generative AI and foundation models.

It lays out the potential opportunities and risks surrounding these technologies in society more broadly, with reference to their current and potential future use in educational environments in the UK, and the legal and regulatory considerations affecting their deployment in educational contexts. For this reason we have provided our submission as a written response than as a survey response.

Ada has undertaken a wide range of research on the impact of AI, and where governance and regulation ought to be developed. Ada and the Nuffield Foundation are now jointly undertaking research into the role of AI and data-driven technologies in education including a landscape review. This will take in technologies used in education in UK state-funded primary and secondary schools, and relevant actors (from across the public, private and third sector).

We encourage the Department to read Ada's research in the field of AI (see 'Supplementary reading').

For the benefit of clarity, we have provided definitions for the types of AI we refer to in this response.

Definitions

Artificial Intelligence (AI) 'can be defined as the use of digital technologies to create systems capable of performing tasks commonly thought to require intelligence.'[1]

Other key terms related to AI:[2]

Foundation Models 'are AI models designed to produce a wide and general variety of outputs. They are capable of a range of possible tasks and applications, such as text, image, or audio generation. They can be standalone systems or can be used as a "base" for many other applications'.

Generative AI 'refers to AI systems that can generate content based on user inputs such as text prompts. The content types (also known as modalities) that can be generated include images, video, text and audio.'

Large language models (LLMs) 'are a type of AI system trained on text data that can generate natural language responses to inputs or prompts […] [They] are the basis for most of the foundation models we see today (though not all, as some are being trained on vision, robotics, or reasoning and search, for example), performing a wide range of text-based tasks such as question-answering, autocomplete, translation, summarisation, etc. in response to a wide range of inputs and prompts.'

In this response, we use foundation models as a catch-all term to refer to generative AI and large-language models specifically in relation to the definitions given above.

[1] Government Digital Service and Office for Artificial Intelligence, 'A Guide to Using Artificial Intelligence in the
Public Sector' (*GOV.UK*, 2019) <https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector> accessed 13 January 2020.

[2] 'Explainer: What Is a Foundation Model?' <https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/> accessed 1 August 2023.

General opportunities and benefits of foundation models

Ada's research has identified an interest among experts and relevant public-sector stakeholders in using foundation models for a wide range of purposes, and across a broad range of central and local government departments and potential public services.

It is notable that, despite the attention given over the past six months to the advancements in AI technologies, the opportunities for their use are yet to be clearly articulated, defined or put into practice. The most prominent use of these technologies is currently found in the integration of generative AI and large language models (LLMs) in products such as Microsoft's Bing, Google's Bard, and OpenAI's ChatGPT. The widespread availability of these technologies has led to their growing use in people's personal and professional lives, including by students in the education system.

However, the development of bespoke products and services using these technologies is mostly in the early stages. Forthcoming research by Ada has found that the anticipated early use cases for foundation models in the public sector will include document analysis, document synthesis, data analytics, decision-making support, drafting of policy briefings, chatbots and provision of public knowledge access.

The uses of foundation models, generative AI and large language models in education

In the classroom

AI-driven products and services are gaining ground in education, as seen through a growing number of products and investment which provide opportunities for new automated features, analytics and personalised learning.[3] To date, the uses of foundation models in education in the UK, be it for language learning,[4] personalised learning,[5] teaching and assignment preparations, and writing essays[6] so far appears experimental.

However, personalised learning is not a new technological development. AI has long been used to support forms of personalised learning; its development and evolution dates back to 2001.[7] There are currently over 30 viable AI-driven products in the education market, some of which rely on foundation models.[8]

To our knowledge, foundation models are not yet formally embedded in state-funded classrooms. However, Google Classroom will soon roll out a language model-powered personalised learning

[3] See 'Bromcom AI: The UK's First AI Powered MIS' (*Bromcom*, 14 March 2023) <https://bromcom.com/news/bromcom-ai> accessed 1 August 2023; Emma Whitford, 'ChatGPT And AI Will Fuel New EdTech Boom' (2023) <https://www.forbes.com/sites/emmawhitford/2023/01/18/chatgpt-and-ai-will-fuel-new-edtech-boom/> accessed 1 August 2023; Miguel A Cardona, Roberto J Rodríguez and Kristina Ishmael, 'Artificial Intelligence and the Future of Teaching and Learning' <https://www2.ed.gov/documents/ai-report/ai-report.pdf>.

[4] Duolingo Team, 'Introducing Duolingo Max, a Learning Experience Powered by GPT-4' (*Duolingo Blog*, 14 March 2023) https://blog.duolingo.com/duolingo-max/ accessed 26 March 2023.

[5] Anthony Spadafora, 'Google Classroom Is Using AI to Help Children Learn in a Whole New Way' (*TechRadar*, 2022) https://www.techradar.com/news/google-classroom-is-using-ai-to-help-children-learn-in-a-whole-new-way accessed 1 August 2023.

[6] Cardona, Rodríguez and Ishmael (n 3).

[7] Kam Cheong Li and Billy Tak-Ming Wong, 'Features and Trends of Personalised Learning: A Review of Journal Publications from 2001 to 2018' (2021) 29 Interactive Learning Environments 182.

[8] Wayne Holmes and others, 'Artificial Intelligence and Education: A Critical Viewthrough the Lens of Human Rights, Democracy and the Rule of Law.' (2022) <https://www.coe.int/en/web/education/-/new-isbn-publication-artificial-intelligence-and-education> accessed 11 August 2023.

feature to provide teachers with insights into areas needing extra instruction or support, while also providing students with instant feedback.[9]

Other uses of available LLM-powered products, such as ChatGPT, for education are also at an experimental stage. These experiments are often carried out informally by teachers and students. It has been noted by researchers that teachers have experimented with the likes of ChatGPT for lesson planning and personalising assignments while students have explored this technology for essay writing or completing assignments.[10]

For administration

The use of AI systems for schools' administration – for example organisation and analysis of student data for statutory reporting – is similarly in its infancy. Few management information systems[11] currently use AI of any form. Of the few that do, BROMCOM announced LLM-based features in March 2023. These features include chatbots that provide conversational engagement for school staff querying student data; insights about students' behaviours and progress; and translation to facilitate communication with parents whose first language is not English.[12]

Potential benefits of foundation models in education

Ada's current research in AI and formal education is identifying that AI may have the potential to improve education through technologies for 1) administrative tasks, such as student record keeping and progress tracking (e.g., management information system (MIS)), 2) accessibility and inclusion to address broader ranges of students' specific needs, and 3) teaching and learning, for example, through personalised learning.

While many AI systems focus on adaptive or personalised learning, not all of them are specifically designed for formal educational settings. Duolingo, for example, has recently launched new features to help language learners understand more about the answers given in a lesson and practice their skills through roleplay,[13] while Khan Academy is experimenting with a GPT-4-powered tool to facilitate adaptive or personalised learning for students and ease teachers' administrative tasks.[14] Crucially, research[15] shows that there is limited agreement on what counts as evidence for these improvements, confirming that opportunities promised by AI have not yet been realised.

The versatility of foundation models and applications built on top of them (such as ChatGPT and Bing Chat) makes it difficult to gauge the scope and scale of their impact on people and society. The ways

[9] Spadafora (n 12).

[10] Cardona, Rodríguez and Ishmael (n 3).

[11] MIS in the context of education is an information system used for organising data about students. Some systems come with built-in analytics features. State-funded schools in England rely on these systems to keep record data about students to fulfil their statutory reporting duty.

[12] 'Bromcom AI: The UK's First AI Powered MIS' (n 3).

[13] Duolingo Team, 'Introducing Duolingo Max, a Learning Experience Powered by GPT-4' (*Duolingo Blog*, 14 March 2023) https://blog.duolingo.com/duolingo-max/ accessed 26 March 2023[14] Sal Khan, 'Harnessing GPT-4 so That All Students Benefit. A Nonprofit Approach for Equal Access!' (*Khan Academy Blog*, 14 March 2023) <https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/> accessed 26 March 2023.

[14] Sal Khan, 'Harnessing GPT-4 so That All Students Benefit. A Nonprofit Approach for Equal Access!' (*Khan Academy Blog*, 14 March 2023) <https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/> accessed 26 March 2023.

[15] Beeban Kidron and others, 'A Blueprint for Education Data: Realising Children's Best Interests in Digitised Education' (Digital Futures Commission, 5Rights Foundation 2023) <https://digitalfuturescommission.org.uk/wp-content/uploads/2023/03/A-Blueprint-for-Education-Data-FINAL-Online.pdf>.

in which foundation models are used are still evolving and have not yet been systematically evaluated for their effectiveness. More research is needed to assess the impact on teaching practices and on the way students construct their knowledge.

General concerns and risks of foundation models

Foundation models rely on a large amount of training data, computing power (to process the training data) and commands or prompts from humans to perform required tasks. These key components set limits to the effectiveness and reliability of foundation models and characterise the systems' vulnerability to bias, misuse and abuse. We have identified four categories of risk:[16]

1. **Accidental harms arising from AI systems failing, or acting in unanticipated ways**, such as inaccurate or biased information mixed in with accurate, impartial, information about a particular subject. Such errors can undermine the quality and scope of knowledge constructed as part of students' learning.
2. **Harms arising from the misuse of AI systems**, such as the practice of malicious actors generating misinformation using generative AI applications such as ChatGPT and Midjourney. If foundation models were used to generate exam materials, they could be used nefariously to predict or leak both questions and answers prior to the exams.
3. **Structural harms arising from AI systems** altering the dynamics of social, political and economic systems, such as the potential for university access, or automated 'streaming'.[17] This type of harm could also manifest through perpetuating a digital divide (in terms of technology access and digital literacy) which limits the ability of already disadvantaged students to extract benefits from foundation models.
4. **Upstream harms arising further up the AI value chain**, such as negative environmental impacts, and the inappropriate collection or use of personal data or protected intellectual property.[18]

When foundation models are deployed, these harms can manifest in the following ways:[19]

- **Harmful content**: This includes misinformation, disinformation and 'hallucination'[20] due to bias and/or inaccuracies in the training data. Training data often comes from publicly available content such as text, images videos, which are not neutral and can contain misinformation, disinformation or hate speech. Models trained on this data will likely reflect and reproduce any patterns of biases in their outputs – which then reinforces stereotypes,

[16] See Box 4 in Ada Lovelace Institute, *Regulating AI in the UK* (2023) <https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/> accessed 1 August 2023.

[17] In education, 'streaming' refers to the practice of sorting students into groups or classes, according to their abilities and levels of achievements to allow students to learn, according to their abilities. This practice can have a negative effect on motivation and lock students into an underachieving learning path and career path.

[18] Upstream harms in the AI value chain refers to activities that feed into the development of an AI model, as part of a supply chain, that can cause harms, such as commercial exploitation of personal data in the data gathering phase to train AI and in the collection and usage of interaction data to further train AI, or labour abuse. (See Sabrina Küspert, Nicolas Moës and Connor Dunlop, 'The Value Chain of General-Purpose AI' (Ada Lovelace Institute, 10 February 2023) <https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/> accessed 27 March 2023.

[19] Ada Lovelace Institute (n 19).

[20] A situation in which the generative aspect of these models creates outputs that combine a mix of accurate and inaccurate information so seamlessly and convincingly that it is difficult to tell the truth from false information.

further spreads mis- and disinformation, and perpetuates discriminatory outcomes for already marginalised groups. The widespread use and abuse of foundation models could lock people in a bubble of harmful content, and/or skew their worldviews.

- **Automating disparity, injustice, unfair or discriminatory outcomes**: When these models are used to inform decisions – including automated decision-making – the same misinformation, disinformation, biases and prejudices inherent in the training dataset will likely reproduce discriminatory outcomes when it comes to, for example, job or university application screening. The use of these models to automate decisions and actions for efficiency savings could also have the unintended consequence of job losses.
- **Data protection and privacy breaches**: Foundation models introduce risks of unintended disclosure of sensitive information used to train them. This risk is pertinent to the use of internal government data to train these models. At their interfaces with users, these models can also infer protected characteristics or other sensitive information about people – accurately or inaccurately – and feed such inferences back to the datasets and the models' outputs. Inaccurate inferences that inform the models' outputs could result in negative or discriminatory outcomes for individuals.
- **Cybersecurity**: The generative capabilities of these models lend themselves to abuse by malicious actors. This functionality lowers the costs of spreading misinformation, disinformation and propaganda at scale, enables scammers and fraudsters to create even more convincing messages to fool victims, and facilitates the generation of malware that are more effective at evading detections.

Concerns and risks around foundation model usage in education

As noted in the section above 'The uses of foundation models, generative AI and large language models in education', various stakeholders (e.g., developers, teachers, awarding bodies, school administrators, students, parents) are already experimenting with or considering using foundation models, such as ChatGPT.

As uses of foundation models in education are still in the early stages of development and deployment, the challenge for policymakers, schools, teachers and students at this stage is limited evidence of the benefits of their use.

However, looking at the previous impacts of data-driven technologies and the deployment of algorithms within the education sector may help in identifying areas of risk and potential for harm. For example, the failure of the A-level grading algorithm in 2020,[21] and the data protection and data exploitation risks associated with popular commercial learning platforms such as Google Classroom and ClassDojo.[22]

These examples identify necessary safeguards that need to be put in place to ensure that the opportunities imagined or promised by foundation models are realised with minimal detrimental impacts on children and their development. Being clear about the purposes and ensuring careful

[21] Elliot Jones and Cansu Safak, 'Can Algorithms Ever Make the Grade?' (Ada Lovelace Institute, 18 August 2020) <https://www.adalovelaceinstitute.org/blog/can-algorithms-ever-make-the-grade/> accessed 1 August 2023.

[22] Louise Hooper, Sonia Livingstone and Kruakae Pothong, 'Problems with Data Governance in UK Schools: The Cases of Google Classroom and ClassDojo' (2022) <https://digitalfuturescommission.org.uk/wp-content/uploads/2022/08/Problems-with-data-governance-in-UK-schools.pdf> accessed 27 September 2022.

consideration of their benefits and risks will be vital to the safe deployment of foundation models within the education system.

The following are examples of common risks which have specific implications for schools, teachers and students in the context of education when foundation models are used.

Digital divide

The ability to use Foundation models will require basic digital infrastructure, such as a stable broadband connection and computers. The COVID-19 pandemic has highlighted the negative impact on learning caused by the digital divide – access (or lack thereof) to digital infrastructure and digital literacy (among teachers, parents and students) – particularly among disadvantaged children.[23]

Any improvements in access to digital infrastructure post-pandemic, following the Government's initiative to provide schools in England with broadband access by 2025,[24] have not been equally experienced. Research shows that schools still have unequal access to digital infrastructure, because of the reduction in school funding (particularly for non-salary costs)[25] and through inequalities in school funding. [26] Also contributing to the digital divide is the variation in teachers' digital skills and literacy – which are necessary for extracting value from data and data-driven technologies.[27]

Comparable to the context of health,[28] such disparities mean that the potential benefits of foundation models in education may be unequally distributed. To ensure even distribution of potential benefits from foundation models and other emerging technologies, Government should prioritise bridging the digital divide and lowering cost barriers for accessing digital infrastructures (and other emerging technologies).

Scope and quality of knowledge

If systems such as ChatGPT are used for lesson planning, risks relating to hallucination, misinformation, disinformation and accuracy and legitimacy of outputs[29] have the potential to undermine the quality of information and knowledge shared with students.

Bias in the training data has the potential to reinforce historical ways of thinking or dominant values.[30] It can pose risks such as narrowing the scope of knowledge and worldviews – for example in

[23] Alison Andrew and others, 'Inequalities in Children's Experiences of Home Learning during the COVID-19 Lockdown in England*' (2020) <https://onlinelibrary.wiley.com/doi/10.1111/1475-5890.12240> accessed 11 August 2023.

[24] DfE (Department for Education), 'All Schools to Have High Speed Internet by 2025' (*GOV.UK*) <https://www.gov.uk/government/news/all-schools-to-have-high-speed-internet-by-2025> accessed 11 August 2023.

[25] Elaine Drayton and others, 'Annual Report on Education Spending in England 2022' <https://www.nuffieldfoundation.org/wp-content/uploads/2019/11/Annual-report-on-education-spending-in-England-2022-Institute-for-Fiscal-Studies.pdf>.

[26] Luke Sibieta, 'School Spending in England: Trends over Time and Future Outlook' (2021) <https://ifs.org.uk/publications/15588> accessed 11 August 2023.

[27] Sarah Turner, Kruakae Pothong and Sonia Livingstone, 'Education Data Reality: The Challenges for Schools in Managing Children's Education Data.' (2022) <https://digitalfuturescommission.org.uk/beneficial-uses-of-education-data/>.

[28] Ada Lovelace Institute, *The Data Divide* (2023) <https://www.adalovelaceinstitute.org/report/the-data-divide/> accessed 11 August 2023.

[29] Ada Lovelace Institute (n 19).

[30] Torrey Trust, Jeromie Whalen and Chrystalla Mouza, 'Editorial: ChatGPT: Challenges, Opportunities, and Implications for Teacher Education' (2023) 23 Contemporary Issues in Technology and Teacher Education 1.

relation to teaching of historical events. As above, the impact of bias affects teaching preparation, learning materials and assessment development, as well students' completion of assignments.

A critical requirement for the use of these technologies is the need for outputs and generated content (be it visual, audio, or textual) to be checked for accuracy.

Child development and competencies

A key objective of education is to develop students' logical reasoning. Research shows that children gradually develop this ability from a young age and that students should master logical judgments at the conceptual level. This is seen through their thought processes and ability to infer and make judgments from information.[31]

The introduction of foundation models to classrooms, which base their outputs on the statistical sequencing of words,[32] has the potential to disrupt the way students can develop logical reasoning and critical thinking.

If students used foundation models to summarise information or to assist with research, they may be more efficient in completing assignments but could risk losing opportunities to develop logical thinking and other critical skills.

The lack of references to information sources used by the models also makes it harder for students and teachers to critically assess the quality of outputs from the models – which they would otherwise be able to do if they were to do the research themselves.

The development and potential use of foundation models in an educational setting will therefore need careful consideration of the potential impact on the critical thinking and logical reasoning capabilities of students (e.g., the ability to make inferences from given information).[33]

Serious consideration needs to be given to the long-term cognitive impact of potentially growing reliance and overreliance on generative AI and other models for educational purposes. More research – especially longitudinal research – is needed to measure the impact of the these technologies on children's cognitive development and academic competencies.

Assessment of student competencies and behaviours

There are expectations to use foundation models to generate exam materials[34] and mark students' work and exams.[35] However, foundation models remain limited in their capacity to master logical

[31] MA Kuchkarova and S Ganiyeva, 'Features of Logical Thinking' (2023) 4 Web of Scientist: International Scientific Journal <file:///C:/Users/KruakaePothong/Downloads/1463-Article%20Text-2491-1-10-20230328.pdf>.

[32] Luciano Floridi and Massimo Chiriatti, 'GPT-3: Its Nature, Scope, Limits, and Consequences' (2020) 30 Minds and Machines 681.

[33] Ibid.

[34] Yuhu Shang and others, 'Reinforcement Learning Guided Multi-Objective Exam Paper Generation', Proceedings of the 2023 SIAM International Conference on Data Mining (SDM) (Society for Industrial and Applied Mathematics 2023) <https://epubs.siam.org/doi/abs/10.1137/1.9781611977653.ch93> accessed 15 August 2023.

[35] Filippa Nilsson and Jonatan Tuvstedt, GPT-4 as an Automatic Grader : The Accuracy of Grades Set by GPT-4 on Introductory Programming Assignments (2023) <https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-330993> accessed 15 August 2023.

reasoning, making logical connections between sentences and paragraphs, or identifying information necessary to evaluate arguments.[36]

This limitation undermines the effectiveness of the models to generate exam materials that can measure student's logical reasoning. Models are also likely to have limited efficacy in marking work, for example in assessing argumentation.

More research is consequently needed to improve the capabilities of these models. One way to achieve this is to harmonise computational sciences with learning theories and benchmarks for students' progress (such as key stage assessments).

The reasons for harmonisation of computational sciences with learning theories and progress benchmarks are two -fold. The first is to enable fine-tuning of models to improve the quality of generated exam materials and marking. The second is to enable the evaluation of these technologies as part the procurement process.

[36] Weihao Yu and others, 'ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning' (arXiv, 22 August 2020) <http://arxiv.org/abs/2002.04326> accessed 1 August 2023.

Recommended approach to foundation models deployment and adoption in education

The development, deployment and adoption of foundation models is new across all aspects of society. The risks and the benefits are yet to be experienced or evidenced. While enthusiasm is natural, caution is also required. The adoption of foundation models within education therefore needs to be conscientious and fully aware of the risks. A measured approach to the adoption of foundation models in education will be needed.

As schools are currently adapting to an abundance of education technologies (EdTech), a standardised evidence base and evaluation framework – potentially offered by Education Endowment Foundation (EEF) – could provide support for schools to navigate procurement of these technologies. It could also encourage responsible innovation, safe deployment and adoption of EdTech and AI models in education.

Government should ensure it prioritises research on the benefits and risks of foundation models before promoting their use in schools. Equally important is for Government, together with schools, to develop a shared vision of the specific purposes that technologies could serve in education. This should be based on evidence of the technology's impact, effectiveness and limitations. This vision should, in turn, inform the development of EdTech – rather than the other way around.

Legal and regulatory considerations

The Department for Education recently issued a position statement, reflecting the government's pro-innovation approach to AI regulation, requiring education institutions to 'take reasonable steps […] to prevent malpractice involving the uses of generative AI and other emerging technologies', comply with data protection laws and protect students from harmful content.[37]

While it is welcome that the Department is aware of the potential harms, the lack of any AI-specific regulation or statutory guidance makes adhering to the position statement a complex ask for the people and organisations developing, procuring and using AI technologies.

At present the lack of specific legislation and statutory guidance means anyone developing, procuring, or using the technologies has to comply with a fragmented network of rules. This includes 'horizontal' cross-cutting frameworks, such as human rights, equality and data protection laws, and 'vertical' domain-specific regulation. Guidance on how to interpret these areas of law and regulation in the context of AI, including foundation models, is slowly emerging,[38] but the regulatory landscape remains complex and lacks coherence.

The reference to data protection law in the position statement is welcome. The UK GDPR plays a critical role in the lawfulness of data processing and the legality of developing foundation models

[37] DfE (Department for Education), 'Generative Artificial Intelligence in Education' (*GOV.UK*) https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education accessed 11 August 2023.

[38] See for example: ICO (Information Commissioner's Office), 'Artificial Intelligence' (19 May 2023) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/ accessed 1 August 2023; Google, 'Google AI Principles' (*Google AI*) <https://ai.google/responsibility/principles/ accessed 1 August 2023; TUC, 'Work and the AI Revolution' (25 March 2021) <https://www.tuc.org.uk/AImanifesto> accessed 1 August 2023; Equity, 'Equity AI Toolkit' (*Equity*) <https://www.equity.org.uk/advice-and-support/know-your-rights/ai-toolkit> accessed 1 August 2023; Cabinet Office, 'Guidance to Civil Servants on Use of Generative AI' (*GOV.UK*, 2023) <https://www.gov.uk/government/publications/guidance-to-civil-servants-on-use-of-generative-ai/guidance-to-civil-servants-on-use-of-generative-ai> accessed 1 August 2023.

such as large language models (LLMs) like ChatGPT. This is through rules around the use of data – sometimes personal or copyrighted data – that is used in large quantities to train a foundation model, or the collection of user data (including personal, sensitive, behavioural data) of those using the models.

The UK GDPR enshrines fairness as one of the core data protection principles.[39] Fairness is particularly important when considering the processing of personal data of children, and is of profound relevance to data-driven technologies for the classroom or other education setting.

Fairness is articulated in practical terms through the 15 standards of the UK Age Appropriate Design Code.[40] Fair processing begins at the design process and requires a balancing act between the impacts of data processing for individuals and the interests and purposes of the data processors.[41] It manifests through transparency about what data is being processed, how it is being processed and why, at a minimum. This information should be provided in an easily accessible, open, and honest way to individuals before the processing takes place.

Fair processing means there are no misleading of, or detrimental effects for individuals whose data is processed. This critical layer of transparency and adherence to data protection requirements is difficult to ensure given the current rapid pace of development of foundation models such as ChatGPT.

With that in mind we outline below what will be required from developers, deployers and users of AI foundation models to comply with the UK GDPR and to demonstrate responsible innovation and compliance with data protection. These points are fundamental to the safe and responsible development, deployment and usage of AI systems, especially in primary and secondary education.

Ensuring safe and responsible development, deployment and usage of AI

At the design and development phase

Data protection and privacy- by-design

- Developers must ensure that the collection and processing of data, including but not limited to personal data, is based on an appropriate lawful basis, purpose-specific, relevant and limited to what is necessary, and accurate.
- They are also required to demonstrate how they adhere to the principles of integrity, accountability, confidentiality, and storage limitation of the data, along with embedded default protections.

Impact assessment

- In education, any use of digital technologies has an impact on children's learning experiences and outcomes. Developers must assess the positive and negative impacts of the technologies and data processing on children and children's rights.

[39] Article 5(1)(a) of the UK GDPR prescribes that "personal data shall be processed lawfully, fairly and in a transparent manner in relation".

[40] ICO (Information Commissioner's Office), 'Age Appropriate Design: A Code of Practice for Online Services' <https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/age-appropriate-design-a-code-of-practice-for-online-services-2-1.pdf>.

[41] ICO (Information Commissioner's Office), 'Principle (a): Lawfulness, Fairness and Transparency' (19 May 2023) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/the-principles/lawfulness-fairness-and-transparency/> accessed 2 August 2023.

- Foundation models, generative AI and LLMs fit the definition of 'innovative technology' which is classified by the Information Commissioner's Office (ICO) as resulting in 'high-risk' processing and therefore requires Data Protection Impact Assessment (DPIA).[42]
- By nature, the training data of these models – as is the case with ChatGPT – likely originates from the collection of personal data 'from a source other than the individual without providing them with a privacy notice ("invisible processing")' – which requires a DPIA.
- Given the uses of these models in education (e.g., for lesson planning, preparing materials, personalised learning and assessments), Child Rights Impact Assessments (CRIA)[43] should also be used. This is because the uses of these models in education can interfere with[44]children's rights enshrined in the United Nation Convention on the Rights of the Child.[45]. For example, the right to freedom of thought (Article 14), access to information (Article 17), and goals of education (Article 29).

Safeguards

- DPIAs and CRIAs will help developers identify both positive and negative impacts of foundation models. Developers also have responsibilities to devise appropriate measures to mitigate these risks within the context of education. Their safeguards should be competent enough to address the issues of bias; misinformation; hallucination; generation and curation of harmful content; the (narrowing of) knowledge; fairness in assessment; privacy; and security breaches.
- These safeguards should also include measures to anonymise and filter personal data from the training data; ensure accurate and unbiased datasets; and include measures for retraining the models to improve accuracy, as well as mechanisms for managing Application Programming Interface (API) use.

Post-deployment

Audit and redress

- Developers should continuously monitor foundation models' operation and usage to identify and address any inaccuracy, biases, misinformation and other types of harmful content being generated by the models.
- Developers should also provide options for independent scrutiny for users, researchers, and other external experts to troubleshoot problems with and identify vulnerabilities in the models, as well as offer appropriate levels of redress.

Security

- Developers should clearly state which aspect of the service (e.g., premium, free or both) will receive security updates and for how long.

Application Programming Interface (API)

[42] See in ICO (Information Commissioner's Office), 'Accountability and Governance: Data Protection Impact Assessments (DPIAs)', p.21 <https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias-1-0.pdf>.

[43] UNICEF, 'MO-CRIA: Guide to Using the Child Rights Impact Self-Assessment Tool for Mobile Operators' (2016) <https://sites.unicef.org/csr/files/Guide_to_the_MOCRIA_English.pdf>.

[44] DfE (Department for Education), 'United Nations Convention on the Rights of the Child (UNCRC): How Legislation Underpins Implementation in England' (*GOV.UK*, 2010) <https://www.gov.uk/government/publications/united-nations-convention-on-the-rights-of-the-child-uncrc-how-legislation-underpins-implementation-in-england> accessed 1 August 2023.

[45] *Convention on the Rights of the Child, 20 November 1989, Treaty Series, vol 1577*.

- Developers should install effective mechanisms to monitor third-party uses of the API and ensure on a demonstrable best effort basis the application of the equivalent privacy and data protection policies for third party uses of the API.

These considerations apply irrespective of the type of data collected (personal, anonymised or non-personal, sensitive, synthetic etc). A DPIA and the CRIA are strategic tools to embed these principles in the development and deployment of foundation models in education. Applying DPIAs in combination with CRIAs will prompt developers to think about the purposes of their models in education, anticipate their impact and become more aware of children's diverse requirements -- thus addressing data protection and other risks by design.

Accountability measures

Regulatory backstops are also necessary to ensure safe development, deployment and use of data-driven technologies. The Italian Data Protection Authority suspended Open AI's activity in Italy[46] for reasons including: absence of an appropriate lawful basis for collection and processing of personal data to train the algorithm underpinning ChatGPT; inaccurate results of personal data processing; lack of transparency; and the absence of appropriate age assurance mechanisms and safeguards against exposure of children (under 13) to harmful content. This decision identified and acknowledged the opacity of the model and the failings of OpenAI to adhere to data protection law.

These grounds for suspending Open AI's activity in Italy are likely also applicable to Open AI's operation in the UK and how the model is being used by teachers and students in the context of education. These grounds set precedence for the ICO to carry out its own investigation into products of the same nature, including that from Open AI.

Government should introduce mandatory reporting and transparency requirements for developers of foundation models operating in the UK. This is in response to the data protection enforcement challenges posed by foundation models due to their opacity and the datasets used to train them. These requirements could include regulatory access to the data used to train models and the key parameters defining the model's performance.

The Data Protection and Digital Information (DPDI) Bill

Having a strong data protection regime is critical, however, we note that the regulation is currently subject to potential change under the Data Protection and Digital Information (DPDI) Bill. The Bill intends to reduce the burden on businesses of complying with data protection law by expanding the legal bases for data collection and processing; removing requirements such as the obligation to carry out DPIAs when high-risk processing is being carried out; and weakening protections currently enjoyed by individuals against solely automated decision-making.

The Ada Lovelace Institute believes that these changes, taken collectively, risk undermining the safe deployment of AI in the UK, and that the Government should reconsider them. These changes may have a direct impact on the development of systems, products and tools used within education.

The broader development of regulation, legislation and statutory guidance for AI and foundation models is nascent and will need to be closely monitored by the Department for Education. That will include monitoring the development of ideas from the Government's policy paper, 'A pro-innovation approach to AI regulation', published in March which is intended to begin the development of a clearer and more coherent set of rules for those developing, deploying and using AI in different parts

[46] 'Provvedimento del 30 marzo 2023 [9870832]' <https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9870832> accessed 2 August 2023.

of the economy – and which may lead to the introduction of non-statutory guidance, duties or codes of conduct which could impact the education sector.

The Ada Lovelace Institute welcomed these proposals as a sign of the UK's engagement with the difficult regulatory challenge of governing AI, but our research has also identified significant gaps that will require action from Government to fix. Our recent research report *Regulating AI in the UK*[47] provides further detail on these gaps, and makes recommendations for how they can be addressed which the Department may find beneficial for wider consideration.

[47] Ada Lovelace Institute, Box 4 (n 19).

Future predictions and enabling use

While there is an enthusiasm for the development and deployment of foundation models, consideration will need to be given to the many unintended consequences of these large-scale and powerful technologies. They will redefine the scope of knowledge accessible to society as a whole, including for children and therefore pedagogy. The emerging and experimental uses of these models will also shape teaching, learning and assessment practices.

In order to enable meaningful, accurate and beneficial use of these technologies, it will be vital to deploy them only when they are proven to be of benefit within an educational setting. Research[48] shows that the majority of these technologies have not been purposefully built based on teachers' and students' requirements, nor the holistic understanding of the education systems, as a result of a combination of policy push[49] and schools' adoption of digital technologies.[50]

Given both the potential and limitations of these models, Government should prioritise a cautious approach to these technologies' integration, particularly in primary and secondary education, due to the evolving capacities of students as well as their vulnerabilities.

Procurement of these models for education should strictly be determined by the suitable purposes they can serve, as they remain nebulous in their specific development for education. There is also a lack of evidence to prove tangible benefits to students, yet they have profound impact on the scope and quality of knowledge, pedagogy and privacy.

More research is required to build the evidence of tangible benefits and on the latent effect and long-term impact of the use of these technologies on knowledge, pedagogy, broader society, and cultural and human exchanges.

Any policies to encourage uptake of these technologies in schools should seriously consider the costs of access to these technologies and the infrastructure that supports them.

ENDS

Supplementary reading

Explainer: 'What is a foundation model?' – A simplified definition of foundation models, also known as 'general-purpose artificial intelligence' (GPAI).

Regulating AI in the UK – An approach to AI regulation and recommendations for the government and the Foundation Model Taskforce.

Regulate to innovate - A route to regulation that reflects the ambition of the UK AI Strategy.

Algorithmic accountability for the public sector - Learning from the first wave of policy implementation.

[48] Rebecca Eynon, 'The Future Trajectory of the AIED Community: Defining the "Knowledge Tradition" in Critical Times' (2023) International Journal of Artificial Intelligence in Education <https://doi.org/10.1007/s40593-023-00354-1>

[49] DfE, 'Realising the Potential of Technology in Education' (Department for Education, 2019) <https://www.gov.uk/government/publications/realising-the-potential-of-technology-in-education> accessed 5 June 2021.

[50] Sarah Turner, Kruakae Pothong and Sonia Livingstone, 'Education Data Reality: The Challenges for Schools in Managing Children's Education Data.' (2022) https://digitalfuturescommission.org.uk/beneficial-uses-of-education-data/.

Regulatory inspection of algorithmic systems - Establishing mechanisms and methods for regulatory inspection of algorithmic systems, sometimes known as 'algorithmic audit'.

Accountability of algorithmic decision-making systems - Developing foundational tools to enable accountability of public administration algorithmic decision-making systems.

Mapping AI and data ethics - Mapping the AI and data ethics field to understand the actors, issues and perspectives that constitute the space.

Supporting AI research ethics committees - Exploring solutions to the unique ethical risks that are emerging in association with data science and AI research.

Algorithmic impact assessment in healthcare - A research partnership with NHS AI Lab exploring the potential for algorithmic impact assessments (AIAs) in an AI imaging case study

AI and genomics futures - A joint project with the Nuffield Council on Bioethics exploring how AI is transforming the capabilities and practice of genomic science.

Rethinking data and rebalancing digital power - Providing a map of four cross-cutting interventions that challenge entrenched systems of digital power and create a digital ecosystem that works for people and society.