JAISD JOURNAL of ARTIFICIAL INTELLIGENCE for SUSTAINABLE DEVELOPMENT

Journal of Artificial Intelligence for Sustainable Development

Vol. 1, No. 1 (2024), Article e93643
https://doi.org/10.69828/4d4kqf

POSITION PAPER

Exploring the Ethics of Generative AI, Education, and Democracy

Wayne Holmes[1,*] and Sopio Zhgenti[2,†]

[1]UCL Knowledge Lab, IOE (UCL's Faculty of Education and Society), University College London, London (UK) and [2]British Georgian Academy, Tbilisi (Georgia)

[*]wayne.holmes@ucl.ac.uk
[†]sopho.zhghenti@bga.ge

Abstract

Generative Artificial Intelligence (GenAI) is increasingly being used in educational contexts, with mostly unsubstantiated claims that it will personalise learning and save teacher time. In fact, the supposed benefits of such outcomes remain both uncertain (do they happen?) and unclear (are they worth having?). Accordingly, in this paper, we explore the impact of GenAI on education, arguing that it poses significant threats to ethical values and democratic principles. Key concerns include the opacity and biases inherent in GenAI systems, the challenges to teacher agency and student learning, and the threats to values such as inclusivity, equity, human rights, transparency and accountability. We note that GenAI tools also make it easier to generate misinformation that can mislead students and teachers, potentially eroding trust in information altogether. We begin by unpacking the core constructs of GenAI, education, democracy, and ethics, highlighting how standard understandings often overlook details that are fundamental for effective education in democratic societies. We then analyse the impacts of GenAI on education and democracy through the lens of an ethical framework. We conclude by arguing that the growing impact of GenAI on education demands a comprehensive and critical AI literacy that prioritizes the human dimension - understanding the ethical considerations, social implications, and human impacts of these technologies. Only then can educators make informed decisions about whether and how to responsibly use GenAI, and how to align its use with democratic ideals.

Keywords: Generative AI; education; ethics; democracy; critical AI literacy

Introduction

Generative Artificial Intelligence (GenAI)[1] is increasingly being used in educational contexts (Alasadi and Baiz; 2023; Baytak; 2023; Chiu; 2023), raising significant ethical challenges that warrant careful examination, particularly regarding their implications for democracy (Allen and Weyl; 2024). Accordingly, this paper explores through a critical studies lens how GenAI might both reinforce and challenge ethical values and democratic principles, focusing on issues such as inclusivity, equity, transparency, human rights, and civic engagement (Miao and Holmes; 2023). In order to support an informed and reflective dialogue among educators, policymakers, and developers, we examine the power structures, biases, and ethical considerations embedded in the deployment of GenAI in education; we explore how GenAI is increasingly shaping the landscape of education, influencing pedagogy, curriculum design, and assessment practices; and we

1 Note that we capitalise Artificial Intelligence to clearly identify it as a research and development domain, rather than an intelligence that is artificial (Holmes and Tuomi; 2022).

discuss issues of fairness, informed citizenship, governance, and accountability. Throughout, the aim is to help ensure that the use of GenAI in education is ethical by design, empowers learners and teachers, promotes social justice, and aligns with democratic values.

The Constructs

Although each of the key constructs (education, GenAI, democracy, and ethics) may seem straightforward, they first need to be specified.

Education

We begin with education, which might appear the most straightforward of our constructs. However, standard understandings of education (as implemented in almost all AI systems developed for use in education) tend to be based on personal experiences of schooling, and do not account for the rich history of education philosophy and research (beginning with the ancient Greeks and including the work of Rousseau, Dewey, Illich, and others). Typically, in such understandings, 'education' is wholly equated with 'learning', the process of transferring knowledge or skills from the teacher's head into the head of the student. This has been criticised as the 'banking' model of education (Freire; 1970). In fact, transfer of knowledge and skills is only one of education's core functions. Education is also about 'socialisation', which "has to do with the many ways in which, through education, we become part of particular social, cultural and political 'orders'" (Biesta; 2011, p. 20), and 'individuation', the process "that allow[s] those educated to become more autonomous and independent in their thinking and acting" (Biesta; 2011, p. 21). Neither of these two additional functions are addressed in any serious way by any AI system designed for use in education.

GenAI

GenAI is also a well-known although often poorly understood construct, which is typically taken to mean 'only' a type of technology, ignoring its socioeconomic origins, the values it embodies, and the roles of the key players (including the developers, funders, and policymakers). Even at a technical level, GenAI is actually an assemblage, of technologies enabled by Artificial Intelligence (AI), designed to automatically create content in response to human-written natural language prompts (Miao and Holmes; 2023). Dramatically brought to the world's attention by the launch of ChatGPT in November 2022 (the novelty was making the technology easily accessible to the public rather than any groundbreaking technical development), most of the world's BigTech companies plus numerous smaller enterprises now offer their own variations, while thousands of other organisations have launched products built on top of one of the existing models.

GenAI works by statistically analysing distributions of words, pixels, or other elements in the data that it has been trained on, identifying and replicating common patterns, such as word associations. However, despite its impressive abilities, and that it is already "a fundamental feature of the digital ecosystem, seamlessly deployed into the physical and digital infrastructure of the internet, social media and smartphones" (Schick; 2023, p. 40), GenAI has severe limitations. Although its output may look inventive, GenAI cannot generate novel ideas or solutions for real-world challenges due to its lack of understanding of real-world objects and social relations inherent in language. In addition, it is only capable of regurgitating in different ways material from its training data, along with all its biases and errors. As such, its accuracy can never be guaranteed. Even the developer of ChatGPT acknowledges this key limitation, stating that while it can generate seemingly reasonable answers, it should not be relied upon (OpenAI; 2021). However, the errors themselves are only the start of the problem. A larger issue is that, because the output looks so credible, the errors often go unnoticed (unless the user happens to be knowledgeable of the topic at hand, which is rarely the case for students). It is this disconnect between the appearance of accuracy and understanding, and the reality of inaccuracy and lack of understanding, that must be addressed or at least recognised.

Democracy

Our third construct, 'democracy' (derived from the Greek 'demos', meaning people, and 'kratos', meaning rule or power), is broadly understood as:

> a universally recognised ideal as well as a goal, which is based on common values shared by peoples … irrespective of cultural, political, social and economic differences. It is thus a basic right of citizenship to be exercised under conditions of freedom, equality, transparency and responsibility, with due respect for the plurality of views.… As an ideal, democracy aims essentially to preserve and promote the dignity and fundamental rights of the individual, to achieve social justice, foster the economic and social development of the community, strengthen the cohesion of society and enhance national tranquillity. (Inter-Parliamentary Union, 1997, Universal Declaration on Democracy, Principles 1 and 2)

This version of democracy is that practised in places such as Europe and North America, which is often referred to as 'liberal' or 'representative democracy (for reasons of space, the many alternatives practised around the world are not addressed in this paper). From 1945, representative democracy as a form of government spread rapidly across the world, with many countries transitioning from authoritarian regimes to 'democratic' ones (Herre and Roser; 2023). However, in the past seventeen years, there has been a significant decline in global democracy, due largely to the erosion of democratic norms, the rise of authoritarianism, growing economic inequality, and threats to electoral integrity (Economist Intelligence Unit; 2022; Freedom House; 2023; Kolvani et al.; 2021). Importantly, the world is also backsliding on democracy in more subtle ways and for more subtle causes. The Inter-Parliamentary Union (2022) identifies critical issues such as disengagement (democracy is weakened when citizens cannot or do not get involved in governance), the lack of youth in politics (50% of the world's population is under 30 years old, while only 2.6% of parliamentary representatives worldwide are in that age range), gender imbalance (around the world, women hold only around a quarter of parliamentary seats), the climate crisis (which is exacerbating inequality, increasing poverty and food scarcity, displacing populations and negatively impacting democratic aims), and the use of technology (such as GenAI) to disseminate fake news and drive polarisation.

Ethics

While ethics is a popular topic of discussion for many (but not all) involved in Artificial Intelligence, the complexities are often missed (Holmes and Porayska-Pomsta; 2023c). For example, are there universal ethical principles that should guide behaviour and, if so, in what should such principles be grounded? Much Western ethics discourse over the past 2000 years has focused on the contrasting normative ethics models of deontology, consequentialism, and virtue ethics (ibid.). In Artificial Intelligence, elements of these approaches have been 'unified' into a framework of five principles (the '5P Framework', Floridi and Cowls; 2019) comprising: (i) beneficence ('doing good'); (ii) non-maleficence ('not doing harm'); (iii) autonomy ('freedom of choice'); (iv) justice (avoiding unfairness); and (v) explicability (transparency and explainability). However, these five principles do not address:

> the fact that the domains in which AI is applied by definition differ from one another, and raise ethical issues specific to them. Although there may be common issues and overlaps, discussions around, for example, the ethics of AI in healthcare and the ethics of AI in autonomous vehicles must consider issues specific to those domains (e.g. patient choice in healthcare, and assigning responsibility for crashes in autonomous vehicles). (Holmes and Porayska-Pomsta; 2023c, p. 8).

Accordingly, we propose a sixth 'p', centred on the domain in which the AI system is being applied (e.g., in education, raising such issues as the ethics of choice of pedagogy – e.g., didactic instructionism or guided constructivism).

While all these principles are clear, they are inevitably high-level, and there are few guidelines for how they might be actioned in the context of any particular application of AI. Actioning the principles is only made more difficult by their relativistic nature, that they tend to depend on socio-cultural norms, the circumstances and needs of individuals, and the context. One pragmatic approach is to ground the principles in human rights, as agreed by the almost two hundred United Nations' member states (United Nations; 1948). This is not to suggest that human rights are themselves always uncontroversial but to provide a starting point for actionable ethics policies. In education, there are the additional child rights that also need to be addressed (United Nations; 1989). For example, this includes a child's right to be "protected from economic exploitation" (ibid. Article 32) – which is something rarely considered when an AI company exploits data extracted from a child's classroom engagement with an AI system to enhance both its AI and business models (Holmes et al.; 2022a).

In addition to the 'how' question, there is also the 'when' question: at what point in the AI pipeline (i.e., task definition, data construction, model definition, model training, evaluation, deployment, and feedback, Cramer et al., 2019) should ethical questions be asked and addressed? One answer to the 'when' question is to consider an 'ethical by design' approach, which includes "examining the way we conceive of", as well as develop and deploy, AI (Nurock et al.; 2022, p. 173). Put another way, this involves ensuring that we 'do ethical things'" not just 'do things ethically' (Holmes et al.; 2021), which in education leads to the question of the purpose of the AI: "Is it to prepare students to pass examinations or to help them self-actualise? Is it to address issues identified in computer science departments or problems identified by educators in classrooms worldwide? Is it to replace teacher functions or to empower teachers?" (Holmes and Porayska-Pomsta; 2023c, p. 9).

GenAI and Education

As noted in the introduction, GenAI is already being widely adopted throughout education – from students using it to generate text to teachers using it to generate lesson plans, and much more besides. For many, GenAI is, when used 'carefully', nothing but a good thing. Excited headlines are common. For example, it has been predicted that GenAI will enhance productivity in education by $180m (although the authors don't share their sources, calculations, or timescale, nor do they explain what they mean in this context by 'productivity', McKinsey, 2024). In any case, public enthusiasm for GenAI in education (e.g., Alasadi and Baiz; 2023) has given rise to multiple unsubstantiated promises for swift resolutions to some of the key education 'wicked issues' (Bore and Wright; 2009). These promises include the creation of so-called personalised learning opportunities to address teacher scarcity, automated assessments to alleviate teachers' time constraints, and the rapid development of teaching and learning materials to mitigate planning challenges. However, for none of these claims is there robust evidence (Holmes; 2023b).

Following the launch of ChatGPT, educators around the world have been forced to reconsider their approaches to assessment. In fact, the earliest responses to the arrival of ChatGPT comprised major concerns that GenAI would enable students to cheat in exams and other assessments (SBS News; 2023). However, while it is unclear how students might use ChatGPT to cheat in exams, when they usually have no access to technology beyond a pencil, it is true that GenAI might be used by students when writing outside of school (such as for homework). Taking a different perspective, if machines can – to a credible extent – complete traditional assessments (Katz et al.; 2023), perhaps it is time to rethink both what we assess and how we assess it? Instead, a common response to the concerns about students cheating with GenAI was to use GenAI to perpetuate existing approaches, by automatically generating the assessment tasks and marking rubrics (Nordmann, E.; 2023). However, this overestimates the capability of GenAI (GenAI tools appear to draw impartially on a large pool of ideas, when in fact they by design prioritise those ideas that are most common in the training data, thus missing out on many novel possibilities), and they ignore the fact that many of the more easily available GenAI tools do not meet the standards of existing regulations (e.g., around data and privacy) such as GDPR (Hancock et al.; 2023).

In any case, there is an absence of robust evidence demonstrating the efficacy or safety of GenAI systems in resolving any problems in education globally (Miao and Holmes; 2023). On the contrary, the adoption of GenAI systems, without rigorous critical evaluation, has the potential to exacerbate rather than alleviate existing challenges. One notable concern is the risk of devaluing essential aspects of education, particularly "the role and the tasks of teachers as educators, incrementing the mistrust in their capacity to teach in a digital and AI learning environment" (Council of Europe; 2023) or undermining learner agency (Miao and Holmes; 2023).

For all these reasons, there have been many calls (see, for example, (European Commission; 2022) and (Council of Europe; 2024)) and many initiatives (see, for example, (AI4K12; 2025) and (Elements of AI; 2025)) for the development of what has been called 'AI literacy' – although what comprises AI literacy often remains unclear. Existing initiatives tend to focus on how AI works (what may be summarised as the 'technological dimension' of AI) and sometimes on how to write a GenAI prompt (what may be summarised as the 'practical dimension' of AI). Rarely do they substantively address the impact of AI on humans, on human rights, on democracy, or on the rule of law, as required by the Council of Europe's forty-six Ministers of Education (what may be summarised as the human dimension of AI) (Council of Europe;

2023). It is likely that any worthwhile AI literacy should comprise a judicious balance of all three dimensions of AI, with the somewhat overlooked human dimension promoted to the centre. In other words, AI literacy should be 'critical', addressing issues such as the impact of AI on human well-being, gender, dignity, inclusion, trust, and the digital divide; the implications of AI for human agency, autonomy, privacy, equity, diversity, and discrimination; 'fake' news, the ghost workers of AI, surveillance, election interference, and the impact on jobs; and the implications for sustainable development and the impact on the environment.

Democracy and Education

> Although our institutions may be solid, they will only function in a truly democratic manner if our citizens are fully aware not only of their voting rights, but also of the values our institutions embody. Our education systems and schools need to prepare young people to become active, participative and responsible individuals.… And at the dawn of quantum computing and artificial intelligence it is all the more important that our children should be equipped with the values, attitudes, skills, knowledge and critical understanding that will enable them to make responsible decisions about their future. (Council of Europe; 2018, p. 9).

Education has long been known to be essential for the development and maintenance of democratic societies, for cultivating democratic principles and encouraging citizen involvement in the democratic processes (i.e., education for democracy). Yet, the role of education goes beyond the physical: "We naturally associate democracy, to be sure, with freedom of action, but freedom of action without freed capacity of thought behind it is only chaos" (Dewey; 1903, p. 193). In other words, education plays a key role in the ethical and moral development of individuals, guiding them towards responsible and ethical behaviour in their civic duties. In particular, education provides individuals, especially young people, with key knowledge and a proper understanding of their democratic rights and democratic responsibilities, as well as the workings of democratic systems. It might also be argued that educational institutions and their agents (i.e., teachers, school leaderships, and policymakers) share a common duty to help nurture and promote democratic values such as justice, equality, and respect for others, contributing to the development of responsible and ethical citizens, and thus enhancing society for all. However, around the world education is also under threat, characterised by a lack of financial and human resources, especially in low-income countries (United Nations; 2023).

GenAI, Education and Democracy

In this penultimate section, we consider some of the ethical and human impacts of GenAI on education and democracy (space prevents us being fully comprehensive) through the lens of the 5P Ethical Framework (Floridi and Cowls; 2019), as summarised earlier, augmented by the sixth principle for the specific domain of application (education).

Principles (i) and (ii): Beneficence and non-maleficence

As has been noted previously, "it is almost certainly the case that all members of the Artificial Intelligence in Education (AIED) research community are motivated by ethical concerns, such as improving students' learning outcomes and their lifelong opportunities" (Holmes et al.; 2021, p. 504). In other words, AIED almost definitely aims to be beneficent (good) and there is no suggestion of it being intentionally or explicitly maleficent (bad). However, as that earlier paper continues, "as has been seen in other domains of AI application, ethical intentions are not by themselves sufficient, as good intentions do not always result in ethical designs or ethical deployments" (ibid.).

The fact is that the AIED research community have no control over how others use the technologies that they have developed (although it remains an open question whether they have any responsibility for how their technologies are exploited, Holmes, 2023a). Nor do they usually have control when students use GenAI. Given that GenAI can easily be used to generate fake or factually incorrect information, malevolent actors, or ill-informed actors with good intentions, might exploit GenAI's capabilities and deliberately or unintentionally, but easily and quickly, produce large volumes of malicious or incorrect content, including fake news articles and images (including deepfakes that can be impossible to distinguish from reality), social media posts, and educational content. Such unhelpful output has been flooding the Internet, often with particular narratives aiming to influence the public discourse and hence democracy which relies on an informed citizenry (Milner; 2002).

For at least three reasons, (i) the excess of information (a phenomenon identified as 'censorship through noise', Schick, 2023, p. 42), (ii) GenAI's output looking so human-like and credible, and (iii) GenAI often giving believable explanations that justify its false outputs (Azaria; 2022), it becomes even harder to distinguish between fact and fiction. However, "democratic systems work on core assumptions, including that the state can distinguish citizens from noncitizens, and that citizens can form coherent views based on a 'marketplace of ideas'" (Allen and Weyl; 2024, p. 147). This problem was dramatically illustrated by a deepfake video of Ukraine's president appearing to ask his soldiers, in the early stages of the Russian invasion of his country, to lay down their arms (Simonite; 2022). This particular deepfake was quickly 'defeated', partly because of its primitive quality but also because it was high profile. However, deepfake or simply inaccurate educational content are more likely to appear to be of high quality and less likely to attract critical attention.

The easy availability today of GenAI tools makes it even easier to generate, intentionally or otherwise, potential misinformation that might mislead students and teachers because they do not yet possess the knowledge or skills needed to critically assess and evaluate GenAI outputs, hence the need for teaching and learning critical AI literacy involving the technological and practical dimensions of AI and prioritising the human dimension of AI (Holmes; 2023a). In the face of the misinformation tsunami, teachers and students are all too likely to uncritically accept the information they are provided, increasing the risk that they teach and learn things that are just not true. Ultimately, and underscoring the importance of addressing the responsible use of GenAI in educational environments, students and teachers might lose trust in information altogether (Braw; 2023), which could have devastating implications for democracy.

Principle (iii): Autonomy

As GenAI increasingly impacts on education, a key concern centres on its potential to lead to teacher disempowerment and diminished student agency. The ambition to reduce teacher workload frequently involves the importing of technology, which today increasingly means GenAI, into classrooms: "Teacher workload is an important issue and we are committed to helping teachers spend less time on non-pupil facing activities. We are working with the education sector and with experts to identify opportunities to improve education and reduce workload using generative AI." (UK Department for Education; 2023)

As we have noted, this "reducing teacher workload" narrative involves teachers using GenAI tools to generate teaching materials, tests, and assessment rubrics. However, while automating tasks associated with curriculum development, assessment, and pedagogy is positioned as a time-saving measure, it usually results only in the displacement of teacher time towards content generation through technology: teachers "will find themselves becoming even further overwhelmed with substitutive work tasks" (Watters; 2021, p. 11). This is because the creation of AI-generated material demands a particular set of skills (the so-called "prompt engineering" skills, for which additional professional development is needed), while the GenAI outputs often require additional editing and always require critique.

In addition, the more that teachers use GenAI to generate curriculum materials, potentially the more they will become dependent upon it, the more it will pose a risk to some of their essential skills (including crafting teaching materials, diversifying resources, and adapting teaching approaches), and the more it will affect their role in pedagogical decision-making (especially for teachers whose first language is not one of the world's dominant languages). Indications of over-reliance and de-professionalisation are already being seen in medicine (Aoun and Sandhu; 2019). In short, the shift to GenAI-supported teaching may compromise teacher autonomy, potentially transforming the educational experience from being human-guided to machine-guided. This can also have unintended consequences for students, as the teaching process may become influenced by the hidden biases (particularly standardised opinions at the expense of diversity of opinions) present in AI-generated content. In summary:

> Teachers need to be afforded opportunities that respect their agency, allowing them to make decisions that align with their professional expertise and the specific needs of their students. Meanwhile, students need opportunities to develop their critical thinking, self-regulation and metacognitive skills, and to develop their intentionality, their autonomy, their adaptability, and their responsibility – either with or without the use of appropriate, effective and safe AI-enabled technologies. (Holmes; 2024, p. 142).

Principle (iv): Justice

A threat to justice or fairness centres on the concentration of GenAI power in 'WEIRD' (Western, Educated, Industrialized, Rich, and Democratic) countries (Pinkwart; 2016, p. 780). The most advanced GenAI systems require an enormous amount of computational resources (and, hence, money: GPT-4 is rumoured to have cost $68 million to train), which only the wealthiest companies and nations can afford (Zhgenti and Holmes; 2023). The inevitable concentration of power by those tech giants that build and control the algorithms that shape our lives, in processes usually hidden from public view, create a landscape fraught with challenges to democracy (Hao; 2020). It tilts the balance of influence and decision-making towards a small number of corporations in a small number of countries, thus diminishing the diversity of perspectives and hindering the democratic distribution of power.

The "increasing concentration of power and profitability in a small number of international technology superpowers, across just a few countries" (Miao and Holmes; 2023, p. 21), has other important consequences. Low-income countries rarely have access to the necessary human, technical and financial resources, thus exacerbating the existing digital divide. In particular, such countries also have limited access to technologies used for educational purposes. Taken together, this on the one hand hinders equitable access to high-quality education (which is a long-established child's human right), on the other, it results in limited data from low-income countries being used to train the GenAI models, thus under-accounting for minority (non-WEIRD) views, and further impacting the diversity of outputs generated by these systems. All of this is contrary to the argument that AI systems should foster diversity and be made accessible for everyone (European Commission; 2022).

Intriguingly, the digital divide (between those who are advantaged because they have access to technology and those who are disadvantaged because they do not, Afzal et al., 2023) might actually flip, with children from high socioeconomic contexts being privileged by having access to human teachers, while those from lower socioeconomic groups have to 'make do' with GenAI-enabled robo-teachers (Weller; 2017), fundamentally undermining children's democratic right to a quality education.

Some questions of justice actually arise before anyone has even had an opportunity to use GenAI. For example, it has been extensively reported that the first iterations of OpenAI's GPT generated outputs that were full of hate speech, sexist tropes, homophobic narratives, offensive language and so on – all due to the GPT's training data having been scraped from the Internet, which is well-known to be saturated with such content. To establish 'guard rails' that would prevent these objectionable materials appearing in the generated outputs seen by the public, OpenAI took an approach typical of AI BigTech. They employed poorly paid AI 'ghost' workers in Africa, here tasking them with identifying the unacceptable GPT outputs so that they might be programmatically filtered out. The consequences for those workers have frequently been devastating, including mental health trauma and relationship breakdowns (Perrigo; 2023). However, ensuring safe and humane working conditions is fundamental for democratic societies. In any case, if the filtering process is not transparent and accountable, certain voices or perspectives may be suppressed, affecting freedom of expression. Transparency and accountability are also essential for public trust, and the opacity of GenAI systems can lead to concerns about the democratic governance of emerging technologies.

It is also noteworthy, that despite its guardrails, GPT outputs are still full of "conscious or unconscious algorithmic biases" (Miao and Holmes; 2023, p. 20), and errors, and the tools frequently make things up generating what are inaccurately and unhelpfully anthropomorphised as 'hallucinations' but are more accurately called 'bullshit', (Holmes; 2023a). The biases, embedded in the algorithms during the training process, can perpetuate and even exacerbate existing societal prejudices, thus undermining democracy, because the training data by definition reflect historical inequalities and prejudices present in society. In education, this can lead to discriminatory outcomes in assessments, perpetuating and exacerbating existing societal inequalities, contrary to democratic values of equality and fairness.

For education, a further problem is that teachers and students are rarely properly apprised of the biases, errors or GenAI's habit of making things up. So, when GenAI outputs are relied upon, there is the growing danger that hidden prejudices become accepted as norms.

In any case, civic values that democratic societies strive for (including, but not limited to, tolerance, respect for diversity, and a sense of social responsibility) might be inadvertently undermined by teaching and learning materials generated by GenAI. As few educators using GenAI to support their teaching are fully aware of the ethical and other human impact implications, there is a need for them to develop a critical AI literacy that goes beyond just how to use these tools: a literacy that, as we have noted, prioritises the human dimension of AI, that involves a comprehensive understanding of the ethical considerations, social implications, labour practices and other human impacts connected with these technologies. Only such a literacy will enable educators to make informed decisions about whether to integrate GenAI into education and, if so, how to use them responsibly as well as effectively.

Educators need to be able to trust the technologies they use. This begins with the tools and technologies themselves (which critically includes the human developers and corporations behind those tools) being trustworthy (literally, worthy of our trust): "trust is only valuable when directed at agents or things that are worthy of it: that is, those that are 'trustworthy'; as when the untrustworthy are naïvely trusted, the results can be ruinous" (Jones et al.; 2023, p. 504). It is therefore unclear why so much effort has been put into encouraging teachers to trust AI (e.g., (Nazaretsky et al.; 2022)) without even considering whether the AIED tools and their developers and providers are worthy of that trust. Despite the more than forty-five years of AIED research, the hundreds of small-scale studies, and the hundreds of academic papers that have appeared over the recent months on the use of GenAI in education, currently there is almost no independent evidence at scale for the efficacy, safety and positive impact of AI-enabled tools (including GenAI-enabled tools) on education. Some international organisations and a few governments are beginning to demand tougher standards. In particular, the Council of Europe is currently developing a legal instrument to regulate the use of AI-enabled systems in education to promote and to protect human rights, democracy and the rule of law. This legal instrument is likely to draw on the medical model (which requires robust efficacy and safety testing before medicines can be given to humans) and if agreed will apply to all 46 member states.

Principle (v): Explicability

The opacity of AI systems such as GenAI poses multiple challenges, such as lack of accountability, difficulty in verifying decisions, and the disempowerment of users. For GenAI, while how the system generates its outputs is broadly understood, why it generates any particular output is not explainable (Miao and Holmes; 2023). In education, although not with such immediate consequences as AI deployed in autonomous vehicles (where it might lead to the death of road users) or in the penal system (where it might lead to certain people being more likely to be incarcerated due to the colour of their skin), transparency and explainability are still important. The design of AI systems used in education should enable teachers and students to understand the decision-making processes (for example, why a particular task is being recommended), as well as the limitations of the data and its potential biases, and the limitations of the technology. However, achieving transparency and explainability for GenAI poses a significant technical challenge (Schneider; 2024). Nonetheless, genuine explicability could re-empower teachers and students, enabling them to make better-informed decisions. In particular, teachers would be able to challenge or overturn the AI system's recommendations and maintain control (rather than just be 'in the loop') over decision-making in classrooms.

Principle (vi): Domain (education)

Finally, as we have noted, GenAI is rapidly transforming aspects of the educational landscape, prompting a re-evaluation of traditional approaches to learning, teaching and assessment. The swift integration of GenAI tools pressurises educators to reconsider not only the methodologies employed but also our fundamental understanding of education, its purpose, dynamics, and impact. There are many examples (e.g., the potential of student over-reliance on GenAI, and issues around student privacy and intellectual property). Here we have space only to explore three typical examples: assessment, writing, and 'personalised learning'.

The effect of GenAI on assessments has yet to be fully researched. For example, it has been noted that incorporating AI in assessment "undervalued the human effort of students" (Byrne et al.; 2010, p. 33). Moreover, "even if AI was capable of fair and accurate marking of free text, implementing such a system would also ignore how much a teacher learns about their learners when they read what the learner has written — insights that no dashboard [or AI tool] will ever give" (Holmes et al.; 2022a, p. 22). There is also a lack of evidence regarding the potential emotional impact on students if their assessments are conducted by AI, and whether they might think it to be fair. Furthermore, the influence of GenAI on a student's writing approach remains unexplored, as the dynamic in which human evaluators are replaced by AI counterparts has yet to be thoroughly examined. The potential impact of GenAI assessments (whether the generation or marking of assessment materials) might reduce opportunities for dialogue, discussion, conflict resolution, and compromise that authentic assessments can provide, thus compromising democratic principles of quality education. This is not to argue that such issues will occur but to acknowledge that as yet we simply do not know.

In fact, the challenges posed by GenAI in education transcends those around assessment. As we have noted, students and teachers are also using GenAI tools for creating teaching and learning materials and accomplishing numerous other everyday classroom tasks. However, as GenAI tools are increasingly integrated into education, how might they influence student skills' acquisition? For example, it has been convincingly suggested that writing is a process of thinking, and that 'writing' with GenAI is therefore writing without thinking:

> If writing is thinking, ordering one's ideas, generating text with A.I. may be a way to avoid thinking. What is writing without thinking? As I tried to incorporate [GenAI] into my writing process, I felt a little like a gambler pulling a slot-machine lever over and over, in hope of finding the lucky combination of phrases that communicated something like what I wanted to say. (Chayka; 2023).

If this is anywhere near true, what will also be the impact on children's intellectual development, as thinkers and as writers, and as contributors to democratic discourse? Central to this discussion is the exploration of whether GenAI will alter students' ability to independently find solutions without relying heavily on machine assistance (Ko; 2023). In addition, the overreliance on GenAI tools "in education may have profound effects on the development of human capacities such as critical thinking skills and creativity" (Miao and Holmes; 2023, p. 23). The point is, again, we just do not know.

Finally, we turn to 'personalised learning'. Earlier, we noted that one of the core functions of education is 'individuation', enabling those who are educated to become more autonomous and independent. The problems is that some might confuse 'individuation' with

'personalisation', which has been a key goal of much of the 45+ years of AIED research. From its beginnings, the AIED community have focused on developing AI systems that automate one-to-one tuition, in order to leverage Bloom's 2-sigma effect (the claim that one-to-one tuition leads to improved learning outcomes when compared with standard group tuition). However, in so doing, the personalisation is almost always instantiated as individual pathways to standardised outcomes, which is quite different to Biesta's 'individuation' and also effectively undermines 'socialisation'. In any case, it is unclear whether AIED personalisation, even if it is possible, is worthwhile.

Yet, personalised learning is often presented as a remedy for numerous educational challenges, such as student disengagement, motivation deficits, and achievement disparities (e.g., (Baidoo-Anu and Owusu Ansah; 2023; Rudolph et al.; 2010)). In fact, the ambition has been around for almost 100 years (Watters; 2021), recently re-emerging from Silicon Valley. This version adopts a heavily technology-centric worldview, placing individualism above communal values, posing a significant challenge to the democratic intentions of teaching and learning. In particular, so-called personalised learning undermines crucial interactions between teachers and students, which can destabilise trust, motivation, and engagement – all fundamental components for effective education in democratic societies.

> "Imagining education in terms of 'personalised learning' leaves little room for the teacher to play much of a role. It doesn't so much save teachers' time as circumvent its necessity, relocating responsibility for education to the individual learner who goes at their own pace. The obviousness with which the benefits of 'personalised learning' are treated in much research literature on the topic suggests that the main obstacle to its realisation is its implementation, a move which conveniently relocates the problem from its conceptual justification to the backwardness of working digital immigrants. Teachers are once again blamed for the failure of technology to fully realise its supposed potential." (Pelletier; 2023)

In summary, the reductionist nature of personalised learning, towards the dubious aim of 'efficient learning', constrains education to a set of skills, competencies, and measurable outcomes, and risks neglecting the holistic development of students ('individuation') and their broader engagement with the world ('socialisation'). Moreover, technology-based personalised learning systems, such as those based on GenAI, may perpetuate existing disparities and power imbalances, exacerbating socio-economic and cultural inequalities, further challenging democratic ideals.

Conclusion

While those who advocate the use of GenAI in education are no doubt motivated by ethical concerns there is currently little or no evidence for the efficacy, safety or ethical impact of GenAI in educational settings. Nor is it clear whether GenAI will ultimately contribute to or undermine democracy. Instead, GenAI's capability to generate misleading information poses risks, with potential for deliberate or unintentional production of a flood of misinformation, challenging teachers' and students' ability to distinguish fact from fiction, and undermining pluralist dialogue and collaboration. All of this only reinforces the need for all citizens to develop an appropriate level of critical literacy in the human dimension of Artificial Intelligence (the ethical questions and the impact of AI on humans), not just in the technological dimension (how it works) or the practical dimension (how to use it).

Meanwhile, efforts to alleviate teacher workload by means of GenAI often involve automating tasks like generating teaching materials and assessments. Yet, this shift may actually overwhelm teachers with new tasks, requiring specific skills and additional professional development, while breaking the covenant of trust between teachers and students. Dependence on GenAI also risks compromising essential teaching skills and teacher autonomy, while reducing student opportunities for critical thinking and self-regulation. The point is again that, while GenAI is being rushed into education, we still do not know.

In any case, the concentration of GenAI power in wealthier countries raises additional concerns about democracy, power and equity, all of which bring serious ethical implications. Wealthy nations and tech giants control advanced GenAI systems, creating a power imbalance that diminishes diversity and hinders democratic distribution. The digital divide widens, limiting access for low-income countries and underrepresented perspectives, conflicting with the goal of fostering diversity. Meanwhile, GenAI's biases, errors, and opaque filtering raise further ethical issues, perpetuating societal prejudices and potentially undermining democratic values.

Finally, GenAI's lack of transparency raises accountability issues and may disempower users. In education, transparency and explainability are crucial for teachers and students to comprehend decision-making processes and address potential biases. An unreflective reliance on GenAI, in other words without proper awareness of the biases and other issues, risks normalising hidden prejudices and compromising democratic values. This is another reason why educators need a comprehensive and critical AI literacy, but one that prioritises the human dimension of AI, that includes ethical considerations, social implications, and human impacts to enable educators to make informed, responsible decisions about whether as well as how to integrate GenAI into education.

Democracy, although complex and rarely pure, is widely recognised as an ideal governance based on shared values and is underpinned by human rights and universal education. However, as we have illustrated, democracy and education are always under threat, with the arrival of GenAI presenting the latest set of challenges. In this paper, we have only raised multiple questions. The task now is to investigate further the connections between AI, education, ethics and democracy, to ensure that education in this age of GenAI empowers teachers, enhances student agency, and is genuinely for the common good.

Funding

The authors received no external funding for this work.

Authors' Contributions

Holmes conceptualised the work, undertook the literature review, and led the writing of the paper. Zhgenti contributed to refining the arguments and assisted with the writing and revisions.

Acknowledgements

The authors acknowledge that this work was completed independently but would like to thank the reviewers for their constructive comments on an earlier version.

References

AI4K12 (2025). The Artificial Intelligence for K-12 initiative.
URL: *https://ai4k12.org*

Alasadi, E. A. and Baiz, C. R. (2023). Generative AI in education and research: Opportunities, concerns, and solutions, *Journal of Chemical Education* **100(8)**: 2965–2971.

Allen, D. and Weyl, E. G. (2024). The real dangers of generative AI, *Journal of Democracy* **35(1)**: 147–162.

Aoun, M. and Sandhu, A. K. (2019). Understanding the impact of AI-driven automation on the workflow of radiologists in emergency care settings, *Journal of Intelligent Connectivity and Emerging Technologies* **4(6)**.

Azaria (2022). ChatGPT usage and limitations.
URL: *https://doi.org/10.13140/RG.2.2.26616.11526*

Baidoo-Anu, D. and Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning (SSRN Scholarly Paper No. 4337484).
URL: *https://doi.org/10.2139/ssrn.4337484*

Baytak, A. (2023). The acceptance and diffusion of generative artificial intelligence in education: A literature review, *Current Perspectives in Educational Research* **6(1)**: 7–18.

Biesta, G. J. J. (2011). *Good Education in an Age of Measurement: Ethics, Politics, Democracy*, Vol. 1st edition, Paradigm Publishers.

Bore, A. and Wright, N. (2009). The wicked and complex in education: Developing a transdisciplinary perspective for policy formulation, implementation and professional practice.

Braw, E. (2023). AI and gray-zone aggression: Risks and opportunities.
URL: *http://www.jstor.org/stable/resrep52271*

Byrne, R., Tang, M., Tranduc, J. and Tang, M. (2010). eGrader, a software application that automatically scores student essays: With a postscript on ethical complexities, *Journal of Systemics, Cybernetics and Informatics* **8(6)**: 30–35.

Chayka, K. (2023). My A.I. writing robot.
URL: *https://www.newyorker.com/culture/infinite-scroll/my-ai-writing-robot*

Chiu, T. K. F. (2023). The impact of generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney, *Interactive Learning Environments* pp. 1–17.

Council of Europe (2018). Context, concepts and model (vol. 1).
URL: *https://www.newyorker.com/culture/infinite-scroll/my-ai-writing-robot*

Council of Europe (2023). Council of Europe Standing Conference of Ministers of Education. Regulating Artificial Intelligence in Education.
URL: *https://rm.coe.int/regulating-artificial-intelligence-in-education-26th-session-council-o/1680ac9b7c*

Council of Europe (2024). Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CM(2024)52-final).
URL: *https://www.refworld.org/legal/agreements/coeministers/2024/en/148016*

Dewey, J. (1903). Democracy in education, *The Elementary School Teacher* **4(4)**: 193–204.

Economist Intelligence Unit (2022). Frontline democracy and the battle for Ukraine.
URL: *https://pages.eiu.com/rs/753-RIQ-438/images/DI-final-version-report.pdf*

Elements of AI (2025). The Artificial Intelligence for K-12 initiative.
URL: *https://www.elementsofai.com*

European Commission (2022). Ethics guidelines for trustworthy AI.
URL: *https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai*

Floridi, L. and Cowls, J. (2019). A unified framework of five principles for AI in society, *Harvard Data Science Review* **1(1)**.

Freedom House (2023). Freedom in the world.
URL: *https://freedomhouse.org/sites/default/files/2023-03/FIW_World_2023_DigtalPDF.pdf*

Freire, P. (1970). *Pedagogy of the Oppressed (M. B. Ramos, Trans.; 2nd Revised Edition)*, Penguin.

Hancock, R., Badri, S., Suha, A.and Mezei, S., Aas, M. B. and Gijsbertsen, B. (2023). Reconsidering education policy in the era of generative AI (G20 Policy Area: Digital Governance, Security, and Connectivity).
URL: *https://www.research.pitt.edu/sites/default/files/Reconsidering%20Education%20Policy%20in%20the%20Era%20of%20Generative%20AI.pdf*

Hao, K. (2020). OpenAI is giving Microsoft exclusive access to its GPT-3 language model.
URL: *https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/*

Herre, B. and Roser, M. (2023). 200 years ago, everyone lacked democratic rights. Now, billions of people have them.
URL: *https://ourworldindata.org/democratic-rights*

Holmes, W. (2023a). AIED—coming of age?, *International Journal of Artificial Intelligence in Education* .

Holmes, W. (2023b). The Unintended Consequences of Artificial Intelligence and Education, *Education International* .
URL: *https://www.ei-ie.org/en/item/28115:the-unintended-consequences-of-artificial-intelligence-and-education*

Holmes, W. (2024). *AI, AIED and Human Agency*, In C. de la Higuera & J. Iyer (Eds.), AI for Teachers, an Open Textbook. AI4T.
URL: *https://www.ai4t.eu/textbook/*

Holmes, W., Persson, J., Chounta, I.-A., Wasson, B. and Dimitrova, V. (2022a). Artificial Intelligence and Education: A Critical View Through the Lens of Human Rights, Democracy and the Rule of Law.
URL: *https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd*

Holmes, W. and Porayska-Pomsta, K. (2023c). Introduction, *In W. Holmes and K. Porayska-Pomsta (Eds.), The Ethics of AI in Education. Practices, Challenges, and Debates* pp. 1–19.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. M. T., Cukorova, M., Bittencourt, I. I. and Koedinger, K. (2021). Ethics of AI in Education: Towards a community-wide framework, *International Journal of Artificial Intelligence in Education* .

Holmes, W. and Tuomi, I. (2022). State of the art and practice in AI in education, *European Journal of Education: Research, Development and Policies* **57(4)**: 542––570.

Inter-Parliamentary Union (2022). How to mitigate these five threats to democracy.
URL: *https://www.ipu.org/news/news-in-brief/2022-09/how-mitigate-these-five-threats-democracy*

Jones, C., Thornton, J. and Wyatt, J. C. (2023). Artificial intelligence and Clinical Decision Support: Clinicians' perspectives on trust, trustworthiness, and liability, *Medical Law Review* **31(4)**: 501–520.

Katz, D. M.and Bommarito, M. J., Gao, S. and Arredondo, P. (2023). GPT-4 Passes the Bar Exam, *(SSRN Scholarly Paper No. 4389233)* .

Ko, A., J. (2023). More than calculators: Why large language models threaten learning, teaching, and education.
URL: *https://medium.com/bits-and-behavior/more-than-calculators-why-large-language-models-threaten-public-education-480dd5300939*

Kolvani, P., Lundstedt, M., Edgell, A. B. and Lachapelle, J. (2021). Pandemic backsliding: A year of violations and advances in response to Covid-19, *V-Dem Policy Brief* **32**.

Miao, F. and Holmes, W. (2023). Guidance for generative AI in Education and Research.
URL: *https://unesdoc.unesco.org/ark:/48223/pf0000386693*

Milner, H. (2002). *Civic literacy: How informed citizens make democracy work*, London: Henry Milner.

Nazaretsky, T., Ariely, M., Cukurova, M. and Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it, *British Journal of Educational Technology* **53(4)**: 914–931.

Nordmann, E. (2023). Using chatgpt to create teaching materials: Marking criteria rubrics.
URL: *https://www.emilynordmann.com/post/using-chatgpt-for-teaching-marking-criteria-rubrics/*

Nurock, V., Chatila, R. and Parizeau, M.-H. (2022). What does "ethical by design" mean?, *In B. Braunschweig & M. Ghallab (Eds.), Reflections on Artificial Intelligence for Humanity* pp. 171–190.

OpenAI (2021). ChatGPT.
URL: *https://chatgpt.com*

Pelletier, C. (2023). Against personalised learning, *International Journal of Artificial Intelligence in Education* .

Perrigo, B. (2023). Exclusive: Openai used Kenyan workers on less than $2 per hour to make ChatGPT less toxic.
URL: *https://time.com/6247678/openai-chatgpt-kenya-workers/*

Pinkwart, N. (2016). Another 25 Years of AIED? Challenges and Opportunities for Intelligent Educational Technologies of the Future, *International Journal of Artificial Intelligence in Education* **26(2)**: 771––783.

Rudolph, J., Tan, S. and Tan, S. (2010). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?, *Journal of Applied Learning and Teaching* **6(1)**.

SBS News (2023). Cheating with ChatGPT? controversial ai tool banned in these schools in australian first.
URL: *https://www.sbs.com.au/news/article/cheating-with-chatgpt-controversial-ai-tool-banned-in-these-schools-in-australian-first/8170dtv6e*

Schick, N. (2023). Faking it, *RSA Journal* **169(2(5593))**: 40–43.

Schneider, J. (2024). Explainable generative AI (GenXAI): A survey, conceptualization, and research agenda, *Artificial Intelligence Review* **57(11)**: 289.

Simonite, T. (2022). A Zelensky deepfake was quickly defeated. The next one might not be.
URL: *https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/*

UK Department for Education (2023). Generative artificial intelligence (AI) in education.
URL: *https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education/generative-artificial-intelligence-ai-in-education*

United Nations (1948). Universal declaration of Human Rights.
URL: *https://www.un.org/en/about-us/universal-declaration-of-human-rights*

United Nations (1989). Convention on the rights of the child.
URL: *https://www.ohchr.org/EN/ProfessionalInterest/Pages/CRC.aspx*

United Nations (2023). Sustainable development report 2023: Times of crisis, times of change: Science for accelerating transformations to sustainable development.
URL: *https://desapublications.un.org/publications/global-sustainable-development-report-2023*

Watters, A. (2021). *Teaching Machines: The History of Personalized Learning*, MIT Press.

Weller, C. (2017). The largest Internet company in 2030? This prediction will probably surprise you.
URL: *https://www.weforum.org/stories/2017/01/the-largest-internet-company-in-2030-this-prediction-will-probably-surprise-you*

Zhgenti, S. and Holmes, W. (2023). Generative AI and Education: Adopting a critical approach.
URL: *https://botpopuli.net/generative-ai-and-education-adopting-a-critical-approach/*