CAMBRIDGE

Cambridge Papers in English Language Education

# Ethical AI for language learning and assessment

Carla Pastorino-Campos
and Nick Saville

# What will we cover in this paper?

**AI has the power to profoundly enhance human lives – but it must be used responsibly. To unlock its full potential, especially in education, AI technologies must be safe, trustworthy and genuinely beneficial to the people they're designed to support. Put simply, providers have a fundamental responsibility to deliver AI ethically.**

This is more important than ever in language learning and assessment.

Education providers must understand not just how AI can be used, but when it should be used – and be aware of the risks. As Cambridge continues to explore how AI can support learners, educators, and institutions, we've defined six guiding principles to ensure AI is used ethically and effectively in language education.

## Our six guiding principles:

### 1 AI must consistently meet human examiners' standards

AI systems must accurately assess the right language skills and deliver results that people can trust. The technology should enhance the integrity of what's being measured and not be used to cut corners. This is essential when AI is used for high-stakes English tests for admissions or immigration purposes. We urge test providers to collect robust evidence to show how AI scores meet the same standards as highly skilled and experienced human examiners.

### 2 Fairness isn't optional – it's foundational

AI-based language learning and assessment systems must be trained on inclusive data to ensure they are fair and free from bias. Along with using diverse data sets it's essential to continuously monitor for bias and involve a wide range of stakeholders throughout the design process. Equal access to AI tools must also be prioritised, so that all learners – regardless of location or resources – can benefit from the technology.

### 3 Data privacy and consent are non-negotiable

By ethically collecting and leveraging data, we can improve the learning and assessment tools we offer. All parties must be clearly informed about what data is collected, how it's stored, and what it's used for – and they must actively give consent. Behind the scenes, this means implementing robust encryption, secure storage protocols, and safeguards against hacking. This robust approach helps us to develop quality AI language learning and assessment tools that users can trust.

## 4 Transparency and explainability are key

Learners need to know when and how AI was used to determine their results. It is also important that the integrity of the test is maintained. To do this, AI systems must be developed and deployed transparently, to ensure oversight and governance. Providers must be able to clearly articulate the role AI plays, as well as the frameworks that are in place to ensure test accuracy.

## 5 Language learning must remain a human endeavour

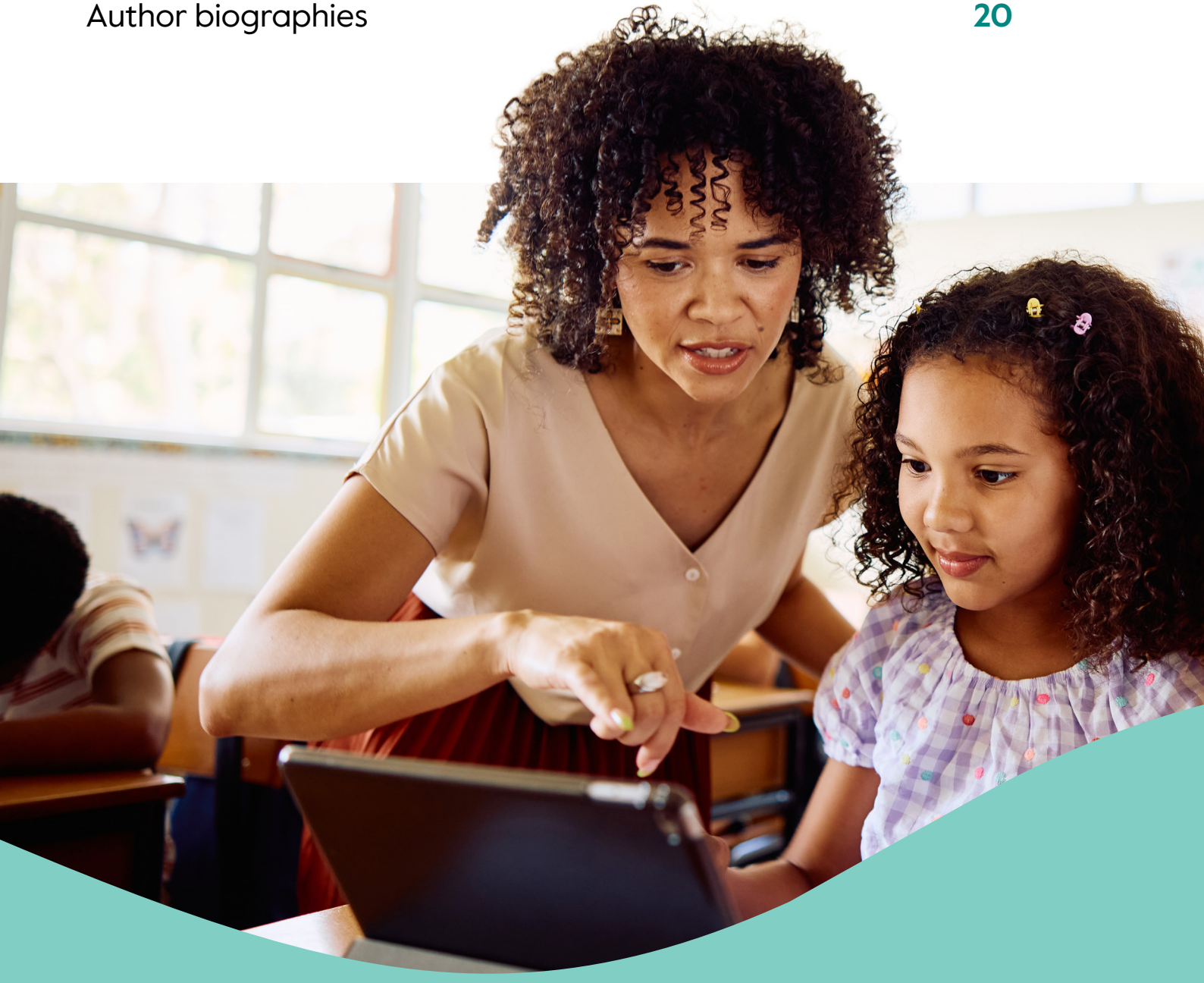
While AI can enhance learning, it cannot replace the uniquely human experience of acquiring and using language. Ethical AI in education must support and empower learners, not overshadow the human touch that makes language meaningful. AI-based assessment must always keep a human in the loop. This helps to establish accountability on the part of test providers, and allows a human to step in where oversight, clarity or a correction is needed for quality control.

## 6 Sustainability is an ethical issue

AI isn't just a digital tool – it's a physical one, with real-world environmental costs. AI systems crunch vast amounts of data and energy, which places a big responsibility on everyone including language providers. This must be kept in mind when choosing which of the different types of AI should be developed or used. It's important to ask: is this AI system necessary, or are there ecologically friendly and more sustainable options available?

# Contents


| | Page |
|---|---|
| Introduction | 5 |
| Terminology | 6 |
| Concerns about the use of AI in education | 8 |
| Ethical principles in language learning and assessment | 11 |
| Implementation | 12 |
| Conclusion | 18 |
| References | 19 |
| Author biographies | 20 |

# Introduction

Everyone is talking about artificial intelligence (AI) and with good reason. AI technologies have been around for a while, but they have now become common in our day-to-day interactions. They can give us shopping recommendations, help us navigate new cities or unlock our phones with only our faces. And they are not just relegated to working in the background; as the widespread access to **generative AI** applications increases, so does our ability to make personal use of the technology, in real time.

Since 2022 and the arrival of generative models of AI such as ChatGPT, the pace of change and the related disruption has accelerated in many spheres of life. The Bletchley Declaration, a statement created and endorsed by representatives of the 29 countries attending the first AI Safety Summit in November 2023, is one of the many documents that recognises that the use of AI presents enormous opportunities for global wellbeing, peace and prosperity while at the same time posing very significant risks at a planetary scale, both expected and unforeseen. This is what is sometimes called 'the **Great AI Trade Off**': AI technologies have the potential to fundamentally transform human lives for the better, and could be used ever more extensively, but they suffer from key vulnerabilities that give humans pause for thought (Mitchell, 2020).

This is why, in order to be **safe and trustworthy**, developers and users of AI-based systems need to take time to reflect on the pros and cons of using them. After all, the social, economic and ethical consequences are not just a matter of concern for the more technically minded among us; these are discussions on issues that may affect every single one of our lives and that can certainly benefit from a variety of voices expressing their views. For the purposes of this paper, we are particularly concerned with those voices in the field of language education.

# Terminology

Most of us by now have developed some kind of understanding of what AI is by repeated exposure to AI-enabled systems on our digital devices, such as Siri and Alexa, and most recently, many people have started using OpenAI's ChatGPT (or other similar applications) for their own purposes, including in language education. But in order to establish common ground for the discussion about ethical uses of AI, it is helpful to refer to some accepted definitions (Galaczi & Luckin, 2024).

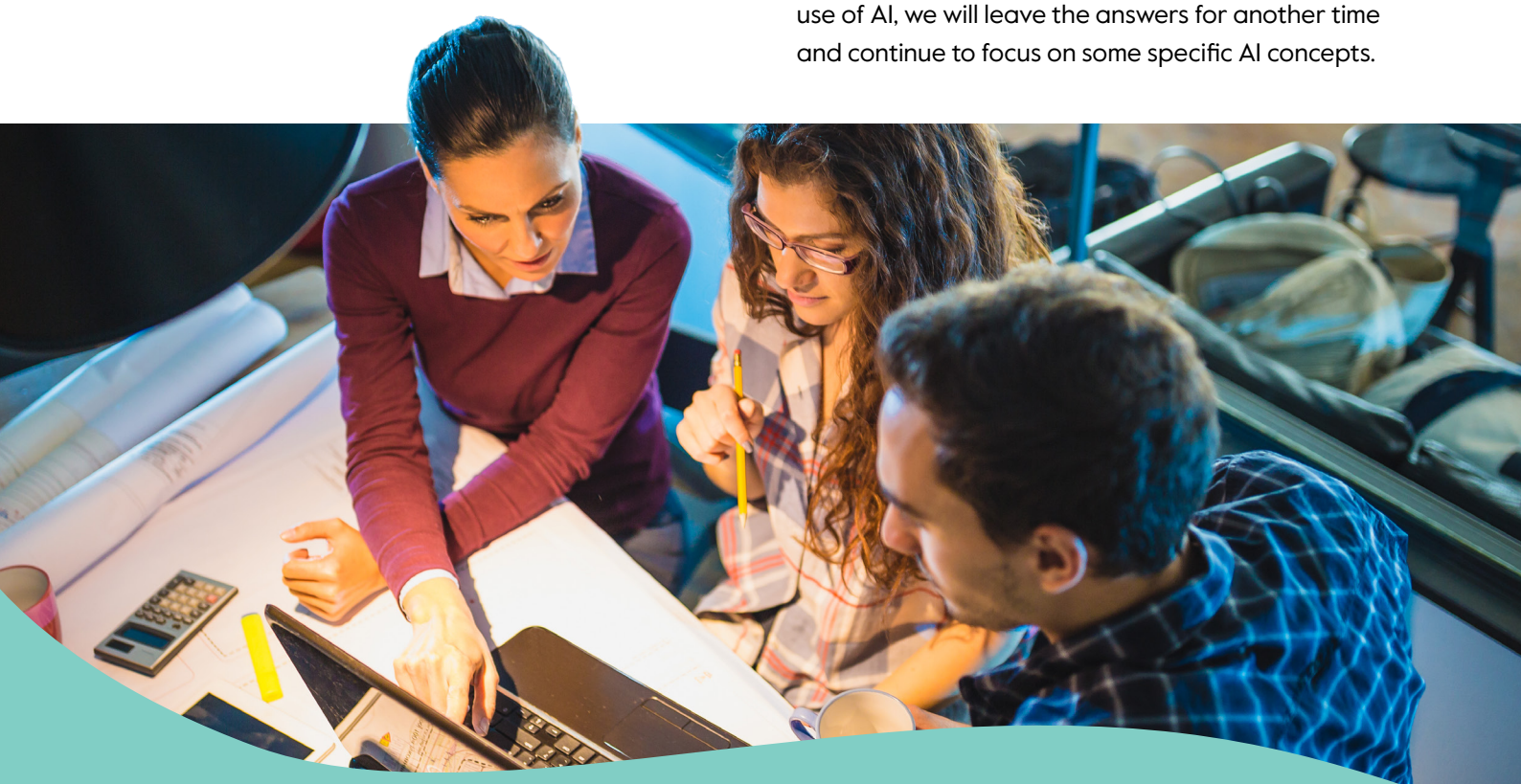
Put simply, AI systems interact with the world through **capabilities** and **behaviours** we think of as human. This is reflected in ChatGPT 's own self-definition:

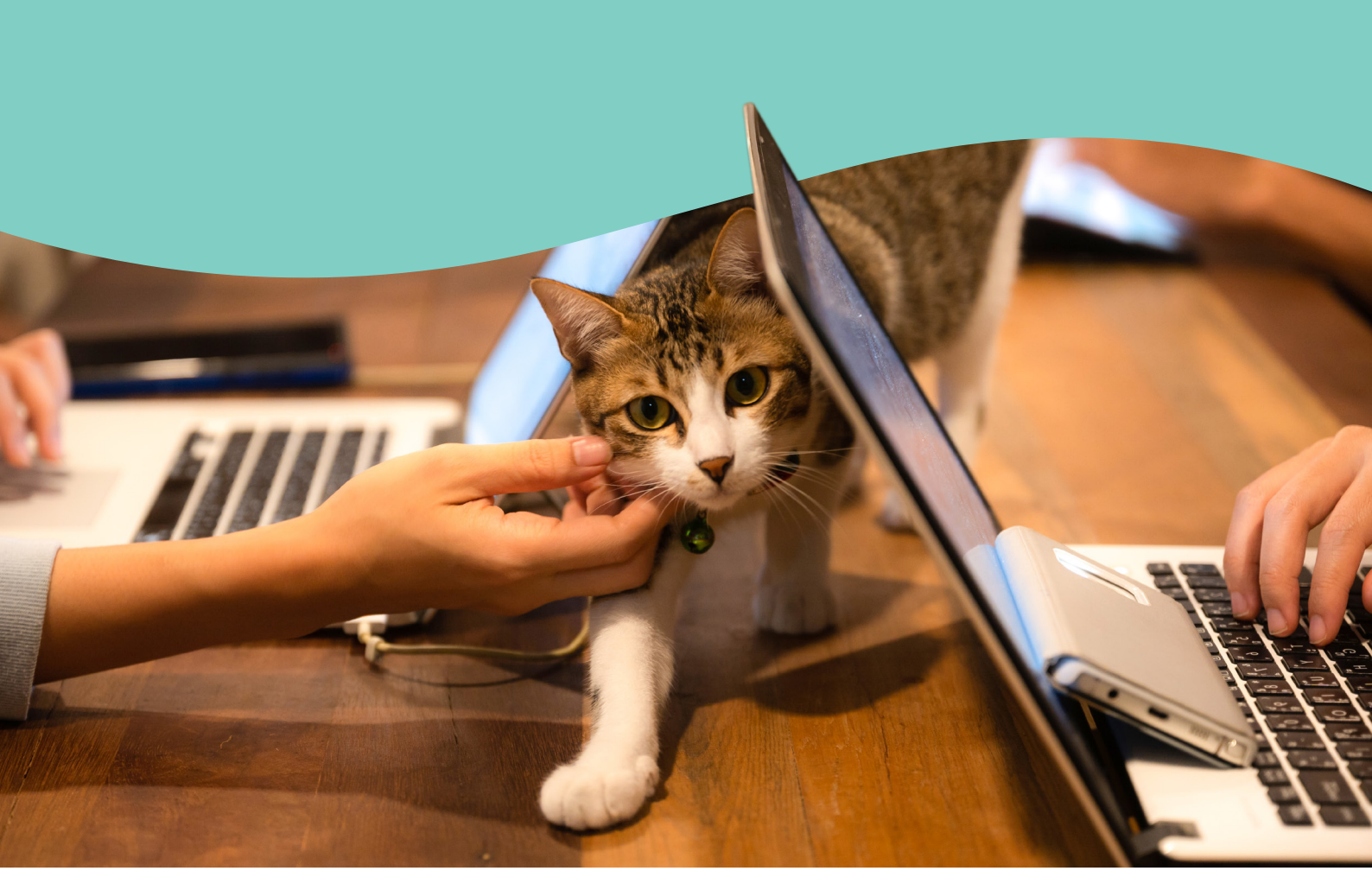
> AI systems are designed to **mimic certain aspects of human intelligence**, allowing them to **perform specific tasks** or **make decisions** autonomously rather than relying on a set of explicit instructions for every step.
>
> **(from GPT-4o by OpenAI) [emphasis our own]**

The European Commission proposes the following more technical definition:

> Artificial intelligence (AI) refers to **systems designed by humans** that, given a complex goal, act in the physical or digital world by **perceiving** their environment, **interpreting** the collected structured or unstructured data, **reasoning** on the knowledge derived from this data and **deciding the best action(s)** to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to **learn** to adapt their behaviour by analysing how the environment is affected by their previous actions.
>
> **(High-level Expert Group on AI, 2018) [emphasis our own]**

There are many outstanding questions surrounding these definitions, but for the purposes of our exploration of the ethical concerns surrounding the use of AI, we will leave the answers for another time and continue to focus on some specific AI concepts.





When people hear the term 'AI' nowadays, they may think of prompt-based systems such as ChatGPT. In a very short time these applications using **generative AI (genAI)** have exploded in popularity, but they are not the only type of AI-based technology currently impacting our everyday lives.

**GenAI** refers specifically to a type of AI that creates new content (e.g., text, images, sound, video) based on large datasets. Humans can interact with genAI systems by 'making requests' to the AI through carefully constructed natural language prompts, as in the case of ChatGPT.

Other types of AI include **predictive** models that classify (e.g., is this a picture of a cat?), predict (e.g., if I like a picture of this cat, I am likely to like this other picture of a cat) or take actions (e.g., if this is a picture of a cat, save it to my desktop). These types of AI application are often working in the background when we interact with digital systems, for example, to recommend what TV series to watch next or to help us write a text message in record time.

Another key distinction to make when looking at the possibilities and risks of AI systems is between symbolic (knowledge-based) systems and sub-symbolic (data-based) systems.

**Symbolic AI** (also known as Good Old-Fashioned AI) uses **explicit rules and logic** to achieve its goals. These systems work like flowcharts or decision trees, by following predefined steps based on expert knowledge (hence their association with 'expert systems'). An example of a symbolic AI system would be a website that asks you questions to determine what the ideal food for your cat would be.

**Sub-symbolic** AI uses a data-based approach, rather than following rules. These systems identify patterns in large quantities of data and use these patterns to determine its outputs. ChatGPT is an example of sub-symbolic AI.

This differentiation between symbolic and sub-symbolic approaches may seem overly technical, but it reveals a fundamental component in our understanding of the consequences of implementing AI systems: the way in which symbolic systems make decisions can typically be explained in a way that is understandable to humans, because its rules are based on human knowledge and expertise. Sub-symbolic systems cannot be easily explained and (at present) remain a **black box** for most humans – though efforts are being made to make *deep learning systems* more explainable (Samek et al., 2017).

# Concerns about the use of AI in education

Students, teachers, universities and researchers have been investigating how they can harness the power of AI in education for some time.

For **learners**, AI may support their learning through applications such as intelligent or dialogue tutoring systems, AI-assisted simulations, AI-based special arrangements or automated essay writing. For **teachers**, AI may be used for plagiarism detection, automated assessment or the curation of learning materials. Finally, **educational institutions** may use AI systems for student selection for admission, scheduling, security, e-proctoring of digital exams, and the identification of students at risk (Holmes & Tuomi, 2022).

In **language learning and assessment,** AI systems have been deployed for many years for practicing and evaluating writing and speaking skills and, more recently, for automated task generation (ATG), item calibration, adaptive assessment with diagnostic feedback, and in the detection of cheating. The latest generative models also have the potential to deliver innovative approaches, allowing the capture and use of digital data as **evidence *of*** and ***for* learning**, thus furthering the integration of learning and assessment. It is expected that AI will continue to deliver innovations, particularly those related to the **personalisation** of learning and assessment experiences and materials, the provision of adaptive, interactive **feedback**, and the effective use of **learning analytics** for evidence-based learning and assessment decisions.

These current and future applications of AI are very promising, but as with any emerging technology, we also have to consider the risks and challenges. In general terms, these focus on how an AI system's **development** (i.e. functioning and/or technical features), **application** (i.e., how it is used), and level of **human participation** (i.e., the involvement of social actors as developers and/or users) may impact the individuals and societies that engage with or are affected by the system.

Moreover, there are specific concerns related to the application of AI technologies in education and their impact on learning, teaching and assessment. For example, Holmes and colleagues (Holmes & Tuomi, 2022) analysed the responses of 17 leading researchers who were asked about key ethical issues surrounding AI in education. They identified several common themes in their responses, some which are also typically found in the literature on the general ethics of AI, including concerns about **data collection, management, ownership and control, privacy, bias and representation**, and **the transparency and intelligibility** of decisions made by AI systems. They also identified topics that are particular to the education dimension, mostly concerning the **quality of the educational content and experience** provided by AI systems. Indeed, the authors note that there are some aspects of the ethics of education that intertwine with ethical AI, and note four key concerns that should be addressed when discussing the ethics of AI in education in particular: the purpose of learning, the choice of pedagogy, the role of the technology with respect to teachers, and access to education (Holmes & Tuomi, 2022).

**Table 1.** Overview of concerns about the use of AI in education

| Development | Application | Human participation |
|---|---|---|
| Relevant to the functioning and/or technical features of the AI system | Relevant to how the AI system is used | Relevant to the social actors involved in the AI system (as developers and/or users) |
| • Unpredictable errors, i.e., AI systems that fail in ways we cannot predict from the start when they have to deal with data that is different from the data used for training the model<br>• Susceptibility to bias, i.e., the tendency of AI systems to replicate or even accentuate human biases or prejudices<br>• Vulnerability to hacking<br>• Lack of transparency in decision making | • Use of content without consent<br>• Unexplainable models used to generate outputs, i.e., AI systems with inner workings that are very complex and difficult to explain<br>• AI-generated content polluting the Internet<br>• Lack of understanding of the real world | • Worsening digital poverty<br>• Outpacing national regulation adaptation<br>• Reducing the diversity of opinions and further marginalising already marginalised voices<br>• Generating deeper deepfakes |
| UNESCO (2023); Mitchell and colleagues (Jenkins et al., 2020; Mitchell, 2020) | | |

Acknowledging the concerns raised by multiple stakeholders (governments, researchers, thinkers, practitioners, learners and others in the AI space), several attempts have been made to summarise and codify the **ethical principles** that should govern AI applications.

In the UK, the Institute of Ethics in AI brought together insights from a Global Summit on the Ethics of AI in Education through a series of international roundtables over a two-year period (2018-19). An Interim Report (2020) set out a blueprint for considering these issues and the final Framework was published in 2021 (www.buckingham.ac.uk/research-the-institute-for-ethical-ai-in-education/).

This pioneering body of work widened perspectives and sought to foster better knowledge and understandings amongst policy-makers and practitioners with reference to the emerging concerns. The clear and concise documents offered a guide for users, focusing on nine objectives with related criteria and checklists. These were aimed at non-specialists with different roles, covering the procurement, implementation and evaluation of educational solutions that incorporate AI.

Another widescale example comes from the Berkman Klein Center for Internet and Society at Harvard University, whose researchers developed the **Principled AI approach**. Fjeld and colleagues (Fjeld et al., 2020) surveyed key ethical AI policy documents from governments, intergovernmental organisations, civil society, the private sector and multi-stakeholder initiatives, and distilled their commonalities into a set of eight principles that address most of the concerns we have discussed so far. However, as highlighted by Holmes and colleagues (Holmes & Tuomi, 2022) more specificity is needed when applying such principles in education.

In response to this need, Nguyen and colleagues surveyed educational policy documents to obtain seven common principles for **ethical AI in education** (Nguyen et al., 2023).

These three approaches have some common features, as well as some different uses of terminology, as summarised in Table 2.

**Table 2**. Overview of three approaches to ethics in AI

| **Framework for Ethical AI in education** (The Institute for Ethical AI in Education, 2020) | **Principled AI** (Fjeld et al., 2020) | **Principles for AI in education** (Nguyen et al., 2023) |
|---|---|---|
| Equity | Fairness and non-discrimination | Inclusiveness |
| Privacy | Privacy | Privacy |
| | Accountability | Sustainability and proportionality |
| Transparency and accountability | Transparency and explainability | Transparency and accountability |
| | Safety and security | Security and safety |
| Ethical design | Professional responsibility | Governance and stewardship |
| Autonomy | Human control of technology | Human-centred AI in Education |
| Achieving educational goals | Promotion of human values | |
| Forms of assessment | | |
| Administration and workload | | |
| Informed participation | | |

# Ethical principles in language learning and assessment

Set against this background and influenced by the frameworks summarised in Table 2, Pastorino-Campos and Galaczi (2025) carried out a systematic review on behalf of Cambridge English. Their aim was to distil out a spectrum of common principles and guidelines focusing specifically on uses of AI in language learning and assessment that could be adopted by Cambridge and tie in with existing ethical practices more widely .

Beyond AI, Cambridge's approach to ethical practice is grounded in the University of Cambridge's educational mission and in Cambridge English's own Principles of Good Practice for English language assessment (forthcoming, 2025). These Principles were developed in light of ALTE's Code of Practice (1994) and have a strong focus on the traditional concern for *fairness* (Saville, 2005; 2013).

As the uses of language assessments have greatly expanded in recent years, a wider range of issues related to equity, diversity and inclusion have emerged. This has resulted in an increased focus on issues of *social justice* and extending the concept of fairness. Our discussion of ethical AI in language assessment and its wider social impact links into this line of thinking. See also ILTA's Code of Ethics, first published in 2002 and now being fully revised in 2025 (see www.iltaonline.com/page/CodeofEthics).

This review has resulted in six inter-related principles related to the following:

1. Validity and reliability
2. Fairness, equity and justice
3. Consent, privacy and data security
4. Transparency and explainability
5. Accountability
6. Sustainability

These principles and their implementation in practice are exemplified in the next section.

# Implementation

## 1 Validity and reliability

### The principle

Language learning and assessment systems using AI must be valid and reliable for their intended purposes and uses. A valid language assessment is one that defines and accurately measures the intended knowledge, ability or skill of language learners (the construct). A reliable language assessment is one that is consistent and dependable in its measurement and can be used with confidence in making decisions about the learners who take it. This principle applies to all kinds of assessment, including high-stakes tests and school-based assessments that have a formative function or are integrated into learning programmes.

### The context

**High-stakes language proficiency tests**, such as those used for admissions or immigration purposes, have the potential to impact a learner's life in significant ways. Therefore, when a new or emerging technology comes in and changes the educational landscape, high-stakes test providers need to ensure that in using it they do not compromise the validity and reliability of their assessments. In other words, they need to demonstrate that the AI-enhanced system still measures what they intend to measure for a specific purpose, and that the results are an accurate portrayal of a learner's ability and can be used reliably for decision-making.

In the case of **integrated learning and assessment systems**, outcomes are not usually high-stakes in the same way. They are typically learning-oriented, providing formative or diagnostic feedback to support the learners themselves and their teachers in taking classroom-based decisions. However, such decisions can have important consequences for the learners. If automated systems using AI are deployed in these cases, validity and reliability will still need to be addressed, albeit using different types of evidence.

### The practice

Ensuring the **validity of test scores** for a specified purpose and use is a central concern for test providers and collecting evidence to support this is part of their day-to-day work. When gathering **validity evidence** for assessments that use AI, a key concern is whether the assessment retains its focus on the intended language skills, or if limitations imposed by the technology may divert attention from the actual behaviours and knowledge to be measured. For example, would you design a reading test to take advantage of genAI that only creates reading texts in an informal register, but not other kinds of formal texts? Clearly not if the specifications of the test require both types of text to be included to correctly represent the **construct of reading**. In other words, assessment providers should not reduce the construct coverage of their assessments to accommodate the technology but should strive to use it in creative ways to maintain or even enhance validity.

Another crucial aspect is whether the AI provides **reliable results** that are accurate and dependable enough for high-stakes purposes. For this reason, test providers need to collect robust evidence to support the assertion that AI-derived scores meet the same standards as highly skilled and experienced human examiners. It is crucial for test providers to reassure test takers and score users that the introduction of AI has not fundamentally changed the meaning of the scores in any way.

# 2 Fairness, equity and justice

### The principle

Learning and assessment systems using AI must not perpetuate biases or discrimination. They must be designed to ensure fairness, irrespective of cultural, linguistic, or socio-economic backgrounds and contribute to just outcomes in society (Saville, 2010).

The basic idea of social justice is that all members of society should have equal access to fundamental rights, opportunities and conditions. This usually means equal access to schooling, higher education, jobs, housing, health, safety and democratic participation. Access to learning and assessment systems that depend on digital technologies involving AI should also be justly distributed. Potential barriers to access should be anticipated (e.g., availability of the Internet) and any unintended effects and consequences that might lead to unfair or unjust outcomes should be investigated through routine validation procedures that can also lead to remedial action.

### The context

Large Language Models (LLMs) that underpin genAI are said to imitate human patterns of behaviour, leading to the assumption that LLM outputs are 'cognitively' and 'attitudinally' similar to those of humans. But who are the humans that shape the way that these machines 'think' through the data they provide, and do they inadvertently introduce unfair biases and discriminations by the choices they make?

Atari and colleagues (2023) used the World Values Survey to compare a commonly used generative AI application (GPT) to the responses of humans from different countries and cultures in questions related to justice, moral principles, social tolerance and other value-based topics. In general terms, the researchers found that GPT's answers seem to be WEIRD (Western, Educated, Industrialized, Rich and Democratic) and that the AI's view of what constitutes an 'average human' is also WEIRD-ly skewed, as the majority of humans do not necessarily conform to WEIRD values.

This is an example of **algorithmic bias** that can be introduced into AI systems when they are trained with very large datasets that have not been curated to mitigate bias against groups that have been historically disadvantaged or marginalised. Indeed, some researchers caution that the use of 'large, uncurated, Internet-based datasets encode the dominant/hegemonic view, which further harms people at the margins' (Bender et al., 2021, p. 613), and place particular emphasis on the curation and documentation of datasets used for AI training and evaluation to minimise the risks of algorithmic bias creeping in.

### The practice

AI systems to be used in learning and high-stakes testing should be trained on **representative and high-quality** data that has been carefully **curated to mitigate bias** against any cultural, linguistic or socio-economic group. This may mean using smaller language models that are representative of the population of interest and are adequate for the intended uses.

For high-stakes English proficiency tests using AI for scoring or content creation purposes, this may mean the collection and curation of learner datasets that are representative of the test-taker population, e.g., in terms of the scores and key demographic variables such as L1, country/region of origin, gender. Furthermore, these datasets should be evaluated before being used for AI training purposes, to minimise the potential effects of any historical or societal bias being reproduced in the AI outputs. In addition, there must be continuous **monitoring for biases** and **mechanisms to mitigate** them in operational processes.

The approach to social justice also requires us to consider how access to AI and its benefits are distributed in society (Bender et al., 2021). All learners should have equal opportunity to use AI technologies in their learning and assessment journey, and the design and development of AI applications should be done by diverse teams, in consultation with a variety of stakeholders. In other words, the application of this principle calls for efforts from every stakeholder to advance **equality and inclusiveness in the design and impact** of the AI systems that are deployed.

# 3 Consent, privacy and data security

### The principle

In designing and deploying AI systems, assessment providers must inform users about data collection, storage, and usage and must protect individuals' privacy rights when handling sensitive data. Explicit consent must be obtained when collecting such data, including for research and validation purposes or for developing learning materials. They also have a responsibly to keep all data secure and protect it against illegal attempts to access it.

### The context

Concerns about how personal and/or sensitive information is obtained, stored and used are at the forefront of stakeholders' minds, and have been for a long time before AI technologies became as pervasive as they are now. In this digital age, legislation such as the European Union's General Data Protection Regulation (GDPR) has attempted to address these concerns, bringing to the attention of internet users the importance of knowing about and agreeing to the data they provide (mostly through cookie policies pop-ups, etc.).

### The practice

Those who design and deploy AI systems must ensure that users have clear and sufficient **information** to understand what types of data will be collected from them and how it will be processed and used. Users can then choose to give **informed consent** to test providers before the data is gathered. **Storing data securely**, and **maximising the robustness** of AI technologies against attacks, such as hacking, are also key considerations. Only through the development of AI systems with minimal vulnerabilities and robust and transparent data processes is it possible to deploy AI applications for language learning and assessment that users can trust.

# 4 Transparency and explainability

### The principle

AI systems should be designed, developed and deployed in **transparent** ways that facilitate oversight and governance. The outputs (what the system is doing) and the mechanisms used to reach the outputs (why/how the system is doing it) should be **explained** in ways that are understandable by stakeholders with various levels of expertise.

### The context

A critical question about the use of an AI system for learning and assessment is whether the decisions it makes can be explained in ways that makes sense to those who use it. If the AI gives a candidate a score of 6.0 for a speaking test, we should be able to explain how it reached that decision. Testing organisations go through a process with human examiners to understand how they apply the marking criteria and in order to build trust in their marking outcomes, e.g., to explain and justify why a 6.0 was awarded. Can we do the same with AI? It seems the answer is 'sometimes,' at least for now. It depends on the system itself, the humans interpreting the processes and other factors that affect human-computer interactions.

Test-takers need to know whether their results are derived from human or machine marking, or a combination of both. Until recently, they could assume that test scores were given by a human examiner in most cases. This assumption was the basis for understanding the scores and the decision-making processes behind them. It is crucial, therefore, that when an AI system is introduced into an assessment process, **all stakeholders are made aware of this and receive sufficient information** to aid them in their understanding of the test and its impact on them.

### The practice

Stakeholders should be given clear explanations of the AI's functioning in accessible language, empowering them to understand and trust the process. For assessment, this is especially important if decisions about test-takers are made using AI as part of the process.

While complete and unrestricted transparency and explainability may seem a desirable outcome, there is some nuance surrounding the application of this principle in language testing, particularly as it intersects with other priorities of equal importance (e.g., test security and confidentiality). For example, consider the principle of transparency. While it is important that test-takers are aware of the involvement of AI systems in the determination of their scores, complete knowledge regarding the features of the algorithms may encourage sophisticated forms of malpractice that compromise the integrity and fairness of the test. An appropriate balance needs to be found between these competing priorities.

Similarly, there could be a tension between the principle of explainability and the need for better precision and accuracy in the results. As we mentioned, it is desirable to have explanations about AI decision-making processes that are akin to a human's, but this is not currently possible when using 'opaque' AI models, such as deep neural networks. On the other hand, these techniques can be used to refine the output of AI systems, bringing them closer to human performance and increasing their accuracy. Should they *not* be used unless they can be fully explained, or does the imperative for more accuracy supersede the need for explainability?

In these cases, a **consensus** needs to be reached regarding the degree of transparency and explainability that is be acceptable for test providers and test users. This requires open and transparent discussions about the risks and benefits of each approach, and engagement with the complex issues to understand better the current and potential capabilities of AI systems.

# 5 Accountability

### The principle

Establishing accountability is crucial: assessment providers must clearly identify where the responsibility lies in their management and oversight processes for making decisions regarding the development, deployment and maintenance of AI systems. This clarity is needed to ensure accountability in taking corrective actions in case of errors or ethical breaches.

### The context

The term 'artificial intelligence' makes us think of an entity that shares some characteristics of human decision-making processes and, by extension, may also have the **agency and responsibilities** that humans have.

This tendency of humans to 'attribute understanding and agency to machines with even the faintest hint of humanlike language or behavior' (Mitchell & Krakauer, 2023, p. 2) is sometimes called the ELIZA effect, named after the chatbot created by Joseph Weizenbaum that was able to trick people into believing it was able to understand them like a therapist would. However, a big outstanding question is whether AI models are capable of understanding their outputs or producing outputs with actual meaning by themselves, in a way that resembles human intelligence in any 'meaningful' way (Mitchell, 2023; Mitchell & Krakauer, 2023). Indeed, some researchers dub LLMs as '**stochastic parrots**,' noting that any given LLM 'is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning' (Bender et al., 2021, p. 617).

The debate over AI's ability to understand and have agency is fundamental in determining who will be held responsible for both decisions that are (at least partially) made by machines, as well as the potentially massive impact of AI on human society and the natural world (Fjeld et al., 2020). This is why designers, developers and users of AI systems need to have informed and in-depth discussions regarding roles and responsibilities when using AI for language learning and assessment.

### The practice

In order to implement the principle of accountability when introducing AI systems for language learning and assessment, it is necessary to have a clear **understanding of who is involved** at every stage, as well as **awareness of potential intended and unintended consequences** – both positive and negative – of the uses of the technology. This requires engagement with the individuals who are directly involved in or are affected by the use of AI systems, as well as others in society at large. Once identified, all staff and stakeholders should be given opportunities to gain a basic understanding of the types of AI that are being deployed so that attribution of responsibility can be made fairly and logically, and potential harms to individuals, society or the environment can be adequately addressed.

While questions regarding AI's agency and understanding remain unresolved, current thinking emphasises that the ethical responsibility and legal liability for AI applications remains with human agents. It is therefore the responsibility of AI developers and users to assign responsibilities and liabilities in a transparent way, to ensure that AI applications remain human-centred and focused on enhancing and empowering individuals and society.

# 6 Sustainability

### The principle

In light of the climate crisis and the sustainability of educational practices using AI, ethical considerations need to be addressed related to choices we make about the types of AI that are developed and deployed.

### The context

Data centres operating AI systems such as Google's Gemini or OpenAI's GPT-4o crunch vast amounts of data and consume vast amounts of energy, giving off heat, generating CO2 and depleting the world's natural resources. This is not sustainable if climate targets are to be met, as Bashir et al have pointed out: 'the growth of Gen-AI is driving increased electricity demand, which runs counter to necessary efficiency gains to achieve net-zero greenhouse gas emissions' (MIT's Climate and Sustainability Consortium, 2024). Furthermore, it is not clear whether renewable energy sources can keep pace with the current 'arms race' for ever more powerful AI systems. Bashir and his co-authors conclude that a focus is required that goes beyond efficiency improvements in ways that 'support social and environmental sustainability goals alongside economic opportunity' (2024).

The challenge is to align a model of technological development with these goals, and in so doing, ensure that the emerging applications of AI contribute positively to society without exacerbating the environmental crisis.

### The practice

Active coordination is required between diverse stakeholder groups to steer AI developments towards responsible and sustainable growth, and to ensure that ethical principles are adequately addressed along the way.

High-level intervention by governments on a global scale is needed to re-focus the deployment of genAI in ways that are both sustainable and bring the most benefit to society. This will likely include greater regulatory oversight involving international conventions, tighter regulations, and legislation, as envisaged in the Bletchley Declaration.

Educators can play their part by seeking to influence policy-makers and technology providers by encouraging them to focus attention on energy-efficient AI systems that deliver future benefits in terms of performance at a lower cost to the environment. Pastorino-Campos and Galaczi (following Nguyen et al,. 2023) refer to this as a principle of Proportionality, encouraging 'the discerning and commensurate use of AI systems' (2025).

Within the field of language education, the issues of sustainability and proportionality can be considered when making choices about the different types of AI to be developed or used (as outlined above). Decisions on when and how to deploy AI can be balanced against the ethical dimensions of climate impact, e.g. is the use of this AI system necessary or are there ecologically friendly and more sustainable options available?

In the field of language assessment, the Principles of Good Practice that guide practitioners in the development and delivery of their assessment systems will need to be updated to take considerations of sustainability and proportionality into account.

# Conclusion

The six principles set out above endorse many of the issues being debated in the wider global context. In a follow up to the Bletchley Park Declaration, a third high-level summit of global leaders and experts from over 100 countries was convened in Paris on 11 February 2025 to establish an international framework for cooperation in AI development and governance. The delegates considered the First International AI Safety Report coordinated by AI 'godfather' Yoshua Bengio (January 2025) and addressed several key themes, including the following:

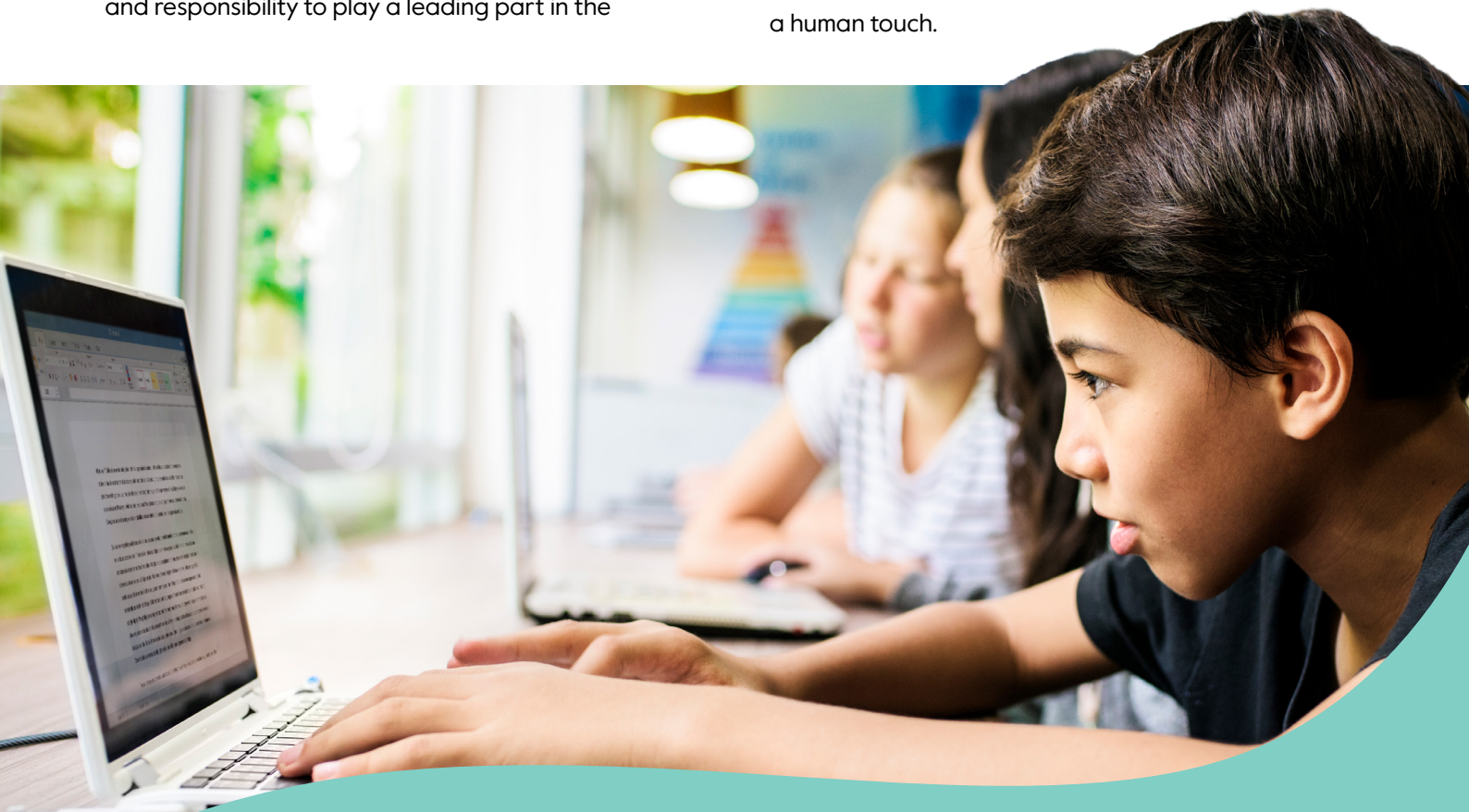
- Promoting AI accessibility to reduce digital divides.
- Ensuring AI is open, inclusive, transparent, ethical, safe, secure and trustworthy, taking into account international frameworks for all.
- Making AI sustainable for people and the planet.

(Bengio et al., 2025)

These points are central to the concerns as set out in this paper. As an interdisciplinary community of AI designers, developers and users working in the context of language education, we have the power and responsibility to play a leading part in the wider debate. Our aim is to create technological applications that are *ethical by design* and that deliver positive impacts on the learning, teaching and assessment of English. These initial reflections on the ethical principles attempt to create a framework to ensure that the AI-enhanced materials we provide remain **trustworthy**, **safe** and **(human) learner-centred**.

Thinking about these topics inevitably brings up interesting and difficult questions: some in the realm of the practical and tangible (How can we use AI to enhance human experience? Are AI-based decisions as accurate as human decisions, or even more accurate?); and some more philosophical and abstract (How does the intelligence we are building and observing compare to human intelligence? How much agency should an AI have?).

Ultimately, we believe that human and artificial intelligence are fundamentally different. Language acquisition and use of language for communication are central to being human and are primarily a human endeavour, and as such, we must retain a human touch.



# References


ALTE. (1994). *Code of Practice.* https://alte.org/Materials

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). *Which humans?* . PsyArXiv. https://doi.org/10.31234/osf.io/5b26t

Bashir, N., Donti, P., Cuff, J., Sroka, S., Ilic, M., Sze, V., Delimitrou, C., & Olivetti, E. (2024). The climate and sustainability implications of generative AI. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.9070dfe7

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bengio, Y. et al. (2025). *International AI Safety Report* (DSIT 2025/001, 2025). https://www.gov.uk/government/publications/international-ai-safety-report-2025

Cambridge English. (Forthcoming, 2025). *Principles of Good Practice.*

Coleman, T. (2021). *Has Covid-19 highlighted a digital divide in UK education?* https://www.cambridgeassessment.org.uk/blogs/has-covid-19-highlighted-a-digital-divide-in-uk-education/

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *SSRN Electronic Paper Collection*. https://doi.org/10.2139/ssrn.3518482

Galaczi, E., & Luckin, R. (2024). *Generative AI and language education: Opportunities, challenges and the need for critical perspectives*. https://www.cambridge.org/gb/cambridgeenglish/research-insights#our-research-papers

High-Level Expert Group on Artificial Intelligence. (2018). *A definition of AI: Main capabilities and scientific disciplines*. European Commission. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, *57*(4), 542–570. https://doi.org/10.1111/ejed.12533

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*, 504–526. https://doi.org/10.1007/s40593-021-00239-1

ILTA. (2002/2018). *Code of Ethics.* https://www.iltaonline.com/page/CodeofEthics

The Institute for Ethical AI in Education. (2021). *The Ethical Framework for AI in Education.* https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf

Jenkins, O. C., Lopresti, D., & Mitchell, M. (2020). *Next wave artificial intelligence: Robust, explainable, adaptable, ethical, and accountable*. https://cra.org/ccc/wp-content/uploads/sites/2/2020/11/Next-Wave-Artificial-Intelligence_-Robust-Explainable-Adaptable-Ethical-and-Accountable.pdf

Joint Committee on Testing Practices. (1988/2002). *Code of Fair Testing Practices in Education.* https://www.apa.org/science/programs/testing/fair-testing.pdf

Mitchell, M. (2020). *Artificial intelligence: A guide for thinking humans*. Pelican.

Mitchell, M. (2023). AI's challenge of understanding the world. *Science*, *382*(6671), eadm8175. https://doi.org/10.1126/science.adm8175

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, *28*(4), 4,221–4,241. https://doi.org/10.1007/s10639-022-11316-w

Pastorino-Campos, C. & Galaczi, E. (2025). Ethical AI for language assessment: Considerations and principles. *Annual Review of Applied Linguistics.* doi:10.1017/S0267190525100081

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*. https://doi.org/10.48550/arXiv.1708.08296

Saville, N. (2005). Setting and monitoring professional standards: A QMS approach. *Research Notes, 22*, 2–5. https://www.cambridgeenglish.org/Images/23141-research-notes-22.pdf

Saville, N. (2010). Developing a model for investigating the impact of language assessment. *Research Notes*, *42*, 2–8. https://www.cambridgeenglish.org/Images/23160-research-notes-42.pdf

Saville, N. (2013). Using standards and guidelines. *The Companion to Language Assessment Volume II: Approaches and Development. Part 7: Assessment Development.* https://doi.org/10.1002/9781118411360.wbcla105

Saville, N. & Taylor, L. (2024). Series Editors' note. In *Language Assessment Literacy and Competence Volume 1: Research and Reflections From the Field* (pp. ix-vii). Cambridge University Press & Assessment. https://www.cambridgeenglish.org/Images/717470-silt-volume-55.pdf

UNESCO. (2023). *Guidance for generative AI in education and research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

World Economic Forum. (2022). *Artificial Intelligence for Children Toolkit*. https://www3.weforum.org/docs/WEF_Artificial_Intelligence_for_Children_2022.pdf

# Author biographies

**Carla Pastorino-Campos** is a Principal Research Manager at Cambridge University Press & Assessment. Her research interests include the cognitive aspects of language learning and assessment, the application of technology in education, and the intersection between technology, ethics and social justice. She conducts research on Equity, Diversity and Inclusion (EDI) and the validity of emerging technologies in language assessment, such as automated marking systems. She obtained her PhD in Theoretical and Applied Linguistics from the University of Cambridge, where she specialised on the psychological aspects of language learning and processing.

**Nick Saville,** PhD, FAcSS is Director of Thought Leadership at Cambridge University Press & Assessment (English) and Secretary-General of the Association of Language Testers in Europe (ALTE). He holds an MA TEFL, MA (Cantab) and a PhD in language assessment specialising in test impact supervised by Prof Cyril Weir. With 45 years' experience in language education, he has been an advisor for the Council of Europe, including for the CEFR and its Companion Volume, and was co-editor of the Studies in Language Testing series (jointly published by Cambridge English and Cambridge University Press) until 2025. His professional interests include English linguistics, plurilingualism, Learning Oriented Assessment (LOA), EdTech combined with educational uses of AI, assessment literacy and ethical frameworks in language assessment.

Find out more at
**cambridge.org**

We believe that English can unlock a lifetime of experiences and, together with teachers and our partners, we help people to learn and confidently prove their skills to the world.

*Where your world grows*

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

All details are correct at the time of going to print in September 2025.