Annual Review of Applied Linguistics (2025), **45**, 294–314
doi:10.1017/S0267190525100081

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Ethical AI for language assessment: Principles, considerations, and emerging tensions

Evelina Galaczi and Carla Pastorino-Campos

Cambridge University Press and Assessment, Cambridge, UK
Corresponding author: Carla Pastorino Campos; Email: carla.pastorino@cambridge.org

Abstract

Many language assessments – particularly those considered high-stakes – have the potential to significantly impact a person's educational, employment and social opportunities, and should therefore be subject to ethical and regulatory considerations regarding their use of artificial intelligence (AI) in test design, development, delivery, and scoring. It is timely and crucial that the community of language assessment practitioners develop a comprehensive set of principles that can ensure ethical practices in their domain of practice as part of a commitment to relational accountability. In this chapter, we contextualize the debate on ethical AI in L2 assessment within global policy documents, and identify a comprehensive set of principles and considerations which pave the way for a shared discourse to underpin an ethical approach to the use of AI in language assessment. Critically, we advocate for an "ethical-by-design" approach in language assessment that promotes core ethical values, balances inherent tensions, mitigates associated risks, and promotes ethical practices.

Keywords: ethical use of AI; language assessment; human-centred AI; fairness; validity

The increased accessibility and rapid integration of artificial intelligence (AI) technology has compelled every sector of society to reflect on its potential benefits and risks. This has led to a proliferation of publications attempting to record the concerns that different sectors have with respect to AI and provide guidelines to minimize existing or potential negative consequences attached to its use. Documents like IBM's *Everyday Ethics for Artificial Intelligence* (2022), UNESCO's *Recommendation on the Ethics of Artificial Intelligence* (2021), and the more recent Australian *Voluntary AI Safety Standard* (Australian Government, Department of Industry, Science and Resources, 2024) and European Union (EU) AI Act (AI Act, 2024) exemplify these efforts across a variety of sectors which might involve high levels of risk, such as healthcare, transport, finance or education. These policy documents attempt to summarize and codify key

© The Author(s), 2025. Published by Cambridge University Press.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

concerns surrounding the use of AI in society while providing ethical and regulatory frameworks to address them as part of a commitment to relational accountability.

The ethical aspects of AI have attracted attention from researchers, ethicists, policymakers, and other interested stakeholders. A systematic review of ethical AI frameworks conducted in 2019 identified over 80 sets of principles (Jobin et al., 2019), and the *AI Ethics Guidelines Global Inventory* (compiled by AlgorithmWatch, n.d.), contains 167 guidelines submitted as of April 2024. It is notable, however, that ethical AI frameworks specifically tailored for AI in education are only recently starting to receive attention (Adams et al., 2021; Holmes & Porayska-Pomsta, 2022; Holmes & Tuomi, 2022; Holstein & Doroudi, 2021; Nguyen et al., 2023), and that frameworks specifically targeted to the uses of AI in assessment, including language assessment, are even more scarce.

The lack of frameworks for ethical AI within language assessment has become more of a concern now that, with the emergence of regulatory frameworks such as the EU AI Act, certain language assessments would be classified as high-risk, as they may determine important decisions about education, training or employment, which might affect an individual's livelihood. AI systems for language testing and for activities such as evaluation of learning outcomes, marking of extended test-taker responses or detecting malpractice, will, therefore, constitute high-risk systems and be subject to specific obligations and responsibilities. The EU AI Act outlines such obligations only within its jurisdiction – the European Union – but it is likely that other countries and regions may soon follow.

The emergence and increased pervasiveness of AI technologies in education has long been perceived as a major change-maker in the education and assessment sectors (Aiken & Epstein, 2000; Holmes & Tuomi, 2022). Its potential for both positive and negative developments (Holmes, 2023) imposes many *ethical* and moral obligations that stand alongside the requirements of *legal* regulatory frameworks. It is therefore timely and crucial that a comprehensive set of principles is developed for use in the domain of language assessment.

Building on the existing corpus of ethical frameworks for general and education-specific applications of AI in the manner depicted in Figure 1, the aim of this chapter is to provide a comprehensive and systematic framework of principles that may support the ethical design, development, use, and evaluation of AI systems within the language assessment domain.

The specificity of this framework stems from the fact that each domain or context will inevitably require principles that address its own use-cases or logic, and that cannot be fully covered even by closely related disciplines (as is the case for education and assessment). In fact, "principles should be understood in their cultural, linguistic, geographic and organizational context, and some themes will be more relevant to a particular context and audience than others" (Fjeld et al., 2020, p. 5).

By assembling the set of principles and considerations presented in this study, we hope to provide a robust starting point for what is certain to be an ever-evolving lens through which to view theory and practice in the use of AI in language assessment. Two questions guided our exploration:

What are widely espoused ethical principles and guidelines on the use of AI in education policies and frameworks?

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

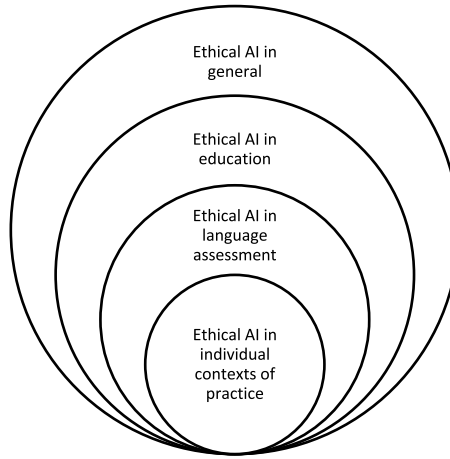


Figure 1. Levels of context-specificity in ethical frameworks.

How might these principles and guidelines apply to the language assessment context?

Method

Defining key terms

Before delving into the analysis of ethical AI guidelines, it is necessary to establish a common understanding of two key concepts. The first is the term Artificial Intelligence itself, which we define based on UNICEF's conceptualization: "machine-based systems that can, given a set of human-defined objectives, make predictions, recommendations, or decisions that influence real or virtual environments. (…) Often, they appear to operate autonomously, and can adapt their behaviour by learning about the context" (2021, p. 21).

A second key term is Generative AI (GenAI). While the availability of GenAI applications, such as ChatGPT, has made this type of AI synonymous with the technology itself, it is important to emphasize that it is not the only AI technology (or even the most widely used in education and assessment). GenAI is a kind of AI technology that "automatically generates content in response to prompts written in natural-language conversational interfaces" (UNESCO, 2023, p. 8). This is in contrast to other AI technologies, which, instead of creating novel content, may be used for prediction, classification, or other automated statistical-, probability-, or rule-based operations.

Data sources and analysis

Best practice in the development of policies or guidelines suggests that a variety of sources of evidence should be included, not only because a plurality of voices enriches the scope of the policy but also to ensure that the most relevant and effective measures are considered. For this reason, this analysis builds on existing ethical frameworks

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

developed for general uses of AI (Fjeld et al., 2020) and AI for educational purposes (Nguyen et al., 2023; The Institute for Ethical AI in Education, 2021) – that is, the two outermost layers in Figure 1, extending them to address issues that pertain to the language assessment domain.

These foundational frameworks used policy analysis (Fjeld et al., 2020; Nguyen et al., 2023) and expert consultation techniques (The Institute for Ethical AI in Education, 2021) to survey the ethical AI policies and viewpoints from key actors in the field, including international and/or intergovernmental organizations (e.g., Organization for Economic Co-operation and Development [OECD]), governments (e.g., Mexico), multistakeholder initiatives (e.g., Institute of Electrical and Electronics Engineers [IEEE]), organizations in the private sector (e.g., Google) and civil society (e.g., Access Now), and individual expert opinion (e.g., policy-makers, educators). Their analyses resulted in consensus-based sets of ethical principles that respond to common concerns regarding the use of AI in general and for educational purposes in particular (summarized in Table 1). These three frameworks are not exhaustive representations of the relevant literature as compared to, for example, the extensive scoping reviews offered by Jobin et al. (2019) and Whittlestone et al. (2019). They were, however, selected as foundational for our purposes due to their comprehensive and robust approach to identifying concepts and principles of relevance for the general and educational domains.

In order to build a comprehensive set of principles regarding the ethical use of AI in language assessment, we designed a thematic analysis procedure based on the thematic analysis process described by Braun and Clarke (2006, 2023), which included several iterations of inductive and deductive (theoretical) thematic analysis, as shown in Figure 2.

In the first stage of the analysis, we familiarized ourselves with the three core frameworks (see Table 1) and employed an inductive method to derive an initial set of codes. This initial set of codes consisted of the principles and sub-principles explicitly stated in the frameworks, compiled and refined by separating principles that appeared merged (e.g., Transparency and Explainability) and by merging principles from different frameworks that overlapped (e.g., the principle of privacy appears in all three frameworks). Two researchers independently revised the initial list of codes and then jointly and iteratively continued improving it until agreement on a coding scheme was achieved. The initial set included 52 codes related to the ethics of AI in the general and educational domains.

The initial coding scheme was then applied to several additional ethical AI policy documents, selected as they allowed us to expand this framework to GenAI and the adjacent domain of language assessment. These documents represent the perspectives of two types of actor that play key roles in the development, use, and regulation of AI in L2 assessment: international or professional organizations (UNESCO, Association of Test Publishers) and English language test providers (Cambridge University Press & Assessment, Educational Testing Service [ETS], Duolingo). The selected documents for coding were:

Guidance for generative AI in education and research (UNESCO, 2023)

Artificial Intelligence Principles (Association of Test Publishers, 2022)

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

Table 1. Summary of principles in the core sources used

	Principled AI (Fjeld et al., 2020)	Ethical principles for AI in education (Nguyen et al., 2023)	The Ethical Framework for AI in education (The Institute for Ethical AI in Education, 2021)
Focus	AI for general purposes, with a focus on human rights	AI in education	AI in education
Method	Policy analysis	Policy analysis	Interviews and roundtables
Sample	Thirty-six ethics of AI policy documents from an array of actors representing: civil society (e.g., UNI Global Union), government (e.g., Government of Japan), inter-governmental organization (e.g., G20), multistakeholder (e.g., Future of Life Institute) and private sector (e.g., Telefónica).	Ethics of AI policies from: UNESCO, OECD, European Commission and European Parliament.	policy-makers, academics, philosophers and ethicists, industry experts and educators.
Principles	Fairness Non-discrimination Privacy Accountability Transparency Explainability Safety Security Professional responsibility Human control of technology Promotion of human values	Inclusiveness Privacy Sustainability Proportionality Transparency Accountability Security Safety Governance Stewardship Human-centred AI in education	Equity Privacy Transparency Accountability Ethical design Autonomy Achieving educational goals Forms of assessment Administration Workload Informed participation

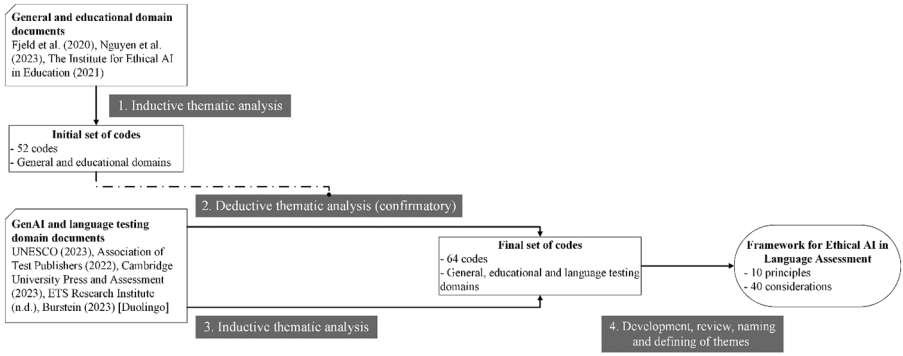


Figure 2. Stages in the analysis of policy documents.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

English language education in the era of generative AI: our perspective (Cambridge University Press & Assessment, 2023)

Responsible Use of AI in Assessment (ETS Research Institute, n.d.)

The Duolingo English Test Responsible AI Standards (Burstein, 2023)

The application of the initial coding scheme to these documents allowed us to confirm that the codes identified in the previous stage were still applicable. We then conducted inductive thematic analysis on the second set of documents to identify codes that may have not previously emerged (e.g., codes specific to L2 assessment). At this stage, working definitions of the codes were constructed based on those found in the documents and, in this process, some codes were merged (e.g., the code "Robustness" became part of the definition of "Security" and was removed from the coding scheme). These emerging codes were added to the coding scheme through an iterative process of corroboration and agreement between the two researchers. The final coding scheme contained 64 codes.

The final set of codes was analyzed and classified into potential themes and subthemes. Themes and subthemes were further reviewed and refined to ensure that the patterns observed were not only accurately identified but also that they sufficiently captured the ethical concerns and considerations reflected in the documents reviewed (Braun & Clarke, 2006). The working definitions constructed in the previous stage were now refined and reformulated by the researchers, using the definitions from previous frameworks to create overarching concepts for each theme and subtheme. The identified themes and subthemes were formulated into the 10 principles and 40 considerations described in the following section.

Principles and considerations

The final proposed framework contained statements related to ethical AI when applied to general, educational, and language assessment domains. They were grouped into the following high-level principles:

- Governance and Stewardship
- Human control of technology
- Accountability
- Human-centricity
- Sustainability and Proportionality
- Privacy
- Fairness and Justice
- Security and Safety
- Transparency and Explainability
- Assessment standards

As a starting point, we have presented these principles and associated considerations in Figure 3, arranged against a general-specific continuum. We believe that this is a useful lens, though we are aware that it is simply one lens through which to view the findings.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

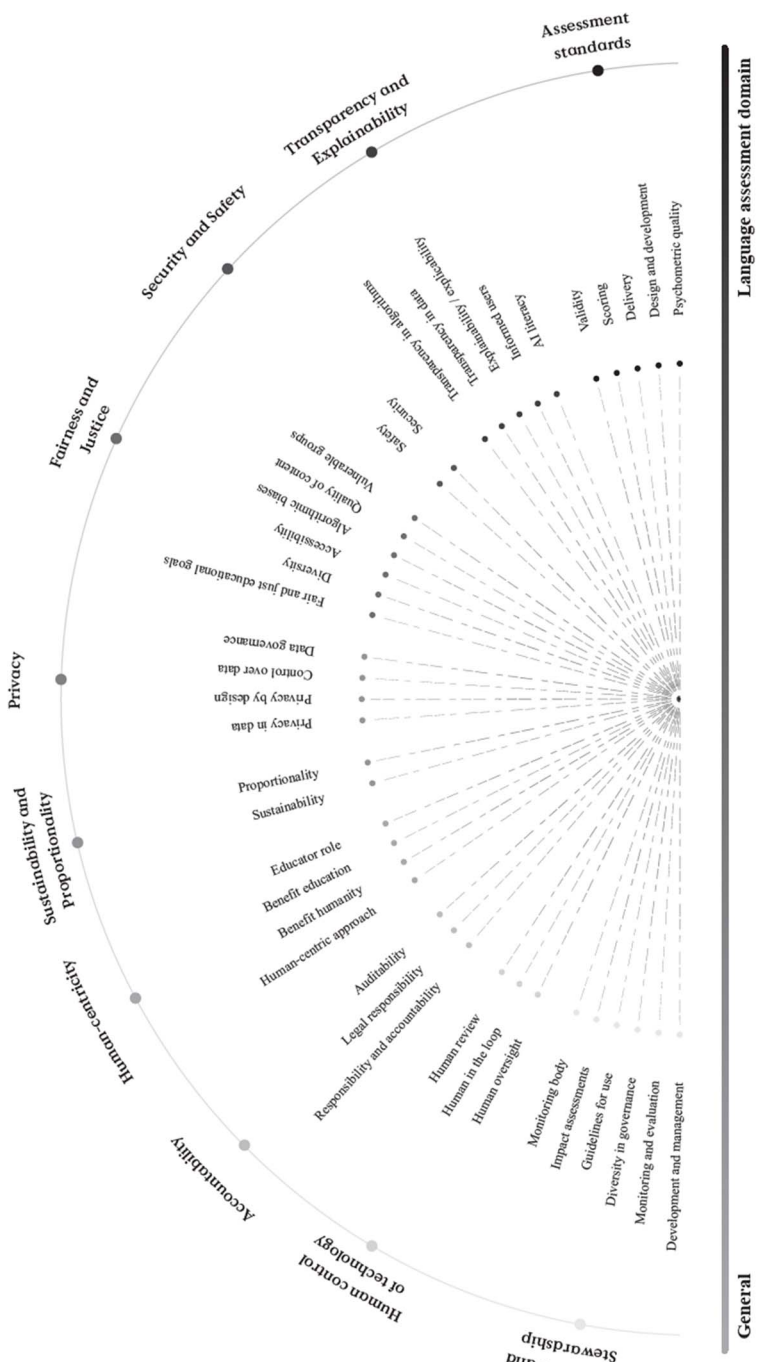


Figure 3. Framework for ethical AI in language assessment.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

Governance and Stewardship

The principle of "Governance and Stewardship" refers to the need to establish and implement relevant and sufficient policies, procedures, and standards for the adequate management of the AI system lifecycle, from design and development to implementation and evaluation. The considerations here aim to provide guidance regarding the processes that must be established to ensure that AI systems are compatible with their intended purposes (Nguyen et al., 2023) and that they uphold all the other identified considerations for the ethical, safe, and responsible use of AI (Leslie et al., 2024). When enacted, these "Governance" considerations turn into "Stewardship" policies, or "governance actions" (Leslie et al., 2024), in that they provide concrete actions to be followed in order to responsibly manage the design, development, implementation, and evaluation of AI systems in educational or assessment contexts.

Six considerations related to "Governance and Stewardship" were found to be common in the documents concerning the ethical use of AI in education and language assessment, as follows:

1.1	Development and management	Policies, procedures, and standards for the development and management of AI systems should be in place and updated regularly to keep up with legislation and changes in the technological landscape.
1.2	Monitoring, auditing and evaluation	A monitoring and evaluation policy should be implemented, where AI systems are audited regarding their functioning and outputs, and the results are used to re-train or tune models where applicable.
1.3	Diversity and inclusiveness in governance	Policies, procedures, and standards related to every part of the AI system lifecycle should consider the perspectives of multiple relevant stakeholders and leverage human expertise whenever possible. Included among these stakeholders should be individuals with knowledge on the ethical implications of using AI in education, who bring a range of perspectives within education and internationally.
1.4	Guidelines for access and use	Providers of education technologies that include AI components should develop and provide guidelines specifying when, how, by whom, and for which purpose the AI systems should be used.
1.5	Impact assessments	An impact assessment policy should be implemented, where the potential negative impacts of AI systems are identified and prevented and/or mitigated.
1.6	Creation of a monitoring or evaluation body	Policies regarding the use of AI in education should recognize the potential need for a specific organization or structure to be created in order to develop procedures, standards, and best practices in this domain.

Human control of technology

The principle of "Human control of technology" highlights concerns related to who is responsible for and in control of any outcomes obtained through AI systems (Fjeld et al., 2020) and emphasizes that AI systems are "not a replacement for trained, qualified, or licensed individuals to arrive at an outcome" (Association of Test Publishers, 2022). The three identified considerations under this principle therefore establish that the locus of control should remain with humans, particularly for those applications that

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

have a high impact on a person's life, such as high-stakes assessment decisions. In other words, "Human control of technology" considerations emphasize that "humans may err, but they alone can shoulder responsibility for mistakes" (Kishimoto et al., 2024, p. 3).

It is acknowledged that complete control over the design, development, implementation, and evaluation of an AI system may not always be possible. Different levels of control may be feasible depending on the features of the AI system, the expertise of the users, ownership over the technologies, and many other variables specific to the different scenarios of use. The purpose of the considerations set out here is to highlight that questions regarding who controls and bears responsibility for the actions of AI systems need to be considered by designers, developers, and users so as to provide safeguards against potential harms derived from the use of AI. Each actor's individual level of responsibility and control may differ based on their capabilities, but the necessity of including humans as overseers, designers, developers, implementers, and evaluators of AI technologies remains.

2.1	Human oversight	The outcomes derived from AI systems, as well as any intended or unintended consequences resulting from their use, should always be attributable to humans or eV.
2.2	Human in the loop	AI systems should harness the unique expertise of humans and the capabilities of both humans and AI systems to produce accurate and trustworthy outcomes. Humans must be involved in every stage of the AI system lifecycle design, development, implementation, and evaluation).
2.3	Human review of automated decisions	When AI systems are used to make decisions about humans, individuals affected should be able to request that the decisions are reviewed by a human auditor.

Accountability

The considerations under the "Accountability" principle provide guidance on who should be held accountable for outcomes or decisions that are made by AI systems (Fjeld et al., 2020). In general terms, this principle refers to the need to "explicitly address acknowledgment and responsibility for each stakeholder's actions" (Nguyen et al., 2023, p. 4,230) throughout the AI system lifecycle.

3.1	Responsibility and accountability	The responsibilities and accountability of each stakeholder involved in every aspect of the AI system lifecycle should be explicitly designated and documented, including specifications regarding auditability and remedies for negative consequences.
3.2	Liability and legal responsibility	Accountability mechanisms should be in place so that the individuals or legal entities that are determined to be responsible for negative consequences derived from AI-driven decisions or AI-generated content can be held accountable.
3.3	Auditability	AI systems should be transparent enough so that knowledgeable humans can inspect and evaluate their design, development and outputs. This includes processes and procedures for the adequate definition, documentation, and recording of the use of AI systems.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

In the context of AI used for assessment purposes, as in many other high-risk domains, there is a strong preference for maintaining human legal and ethical accountability in cases where AI systems are tasked with making high-stakes decisions (UNESCO, 2021, 2023).

Human-centricity

While the determination of roles and responsibilities of human agents involved in the AI system lifecycle has been a concern of the previous principles, the "Human-centricity" principle focuses on the imperative to design, develop, and use AI systems to "complement and enhance human cognitive, social, and cultural capabilities" (Nguyen et al., 2023, p. 4,234). The concept of human-centricity has been key in the development of many frameworks of ethical AI, as it both highlights the need to place AI systems under the control of humans (covered here by the "Human control of technology" principle) as well as establishing that AI systems should be used to contribute to the needs and values of human societies (Kishimoto et al., 2024) and to empower and enable human users (Capel & Brereton, 2023). The four considerations provide further insight into the use of AI for the benefit of human societies and individuals in general and education and assessment in particular.

4.1	Human-centric approach	AI systems should support, enhance, and/or preserve human values, human flourishing, and human well-being.
4.2	Leveraged to benefit humanity	AI systems should be designed, developed, and used to complement and enhance human capabilities, to benefit humanity as a whole and to minimize harm. This includes preserving and enhancing human agency, autonomy, and dignity, and upholding human rights.
4.3	Leveraged to benefit education	AI systems used for education and/or language assessment purposes should be leveraged to benefit the learning and/or assessment journey of the learner. In this context, AI systems should be designed and developed to support a set of predetermined educational goals and should allow learners to be in control of their educational experience, including having the power to negotiate if, how, and when they receive support from the AI system.
4.4	Educator role	AI systems should not replace an educator's role in the learning process and should support and complement the educator's unique capabilities.

Sustainability and Proportionality

The principle "Sustainability and Proportionality" concerns the impact that every part of the AI system lifecycle may have on nature, the economy, society, and individuals. "Sustainability" encompasses considerations focused on the effects of the design and use of AI systems on nature and society, while "Proportionality" encourages the discerning and commensurate use of AI systems. These principles are closely related to notions of beneficence or non-maleficence ("Human-centred" principle) as well as "Fairness and Justice."

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

The two key considerations within this principle help us reflect on how AI design, development, use, and evaluation may affect or be affected by its ecological footprint and its effects on human employment and well-being.

5.1	Sustainability	AI systems should be designed, developed, and used in a manner that minimizes their impacts on the natural environment, society, and the economy.
5.2	Proportionality	The use of AI systems should be proportional to the needs they address and should not be deployed where other more sustainable or less impactful technologies would suffice. This includes discouraging humans from engaging with AI beyond a point that is beneficial to themselves, the use of AI systems to disproportionally replace educational or social activities in the real world, or user over-reliance on AI-generated content.

Privacy

The "Privacy" principle emerged as one of the fundamental concepts to consider when establishing frameworks for the ethical use of AI and has long been at the centre of the regulations and guidelines that govern our increasingly data-based world. Examples of regulatory frameworks that highlight the importance of privacy and data protection are the European Union's General Data Protection Regulation (GDPR) (General Data Protection Regulation, 2016) and the more recent AI Act (2024), both seminal pieces of legislation that outline quintessential data safeguards and foundational regulatory and guidance documents. The four considerations within "Privacy" highlight key concerns and possible remediations.

6.1	Privacy in data	AI systems should ensure that any data collected, processed, or stored by the system is obtained with informed consent and its confidentiality is protected.
6.2	Privacy by design	AI systems should be designed and developed with integration of data privacy and protection considerations from the start.
6.3	Control over the use of data	AI systems should allow human users to have some degree of control over how their data are used and for what purposes. That includes the right to rectify incorrect or incomplete data, the right to erasure of personal data, and the ability to restrict the use of data in AI applications.
6.4	Data governance	AI systems should have clear frameworks for the ethical collection, processing, and storing of data; and be in compliance with relevant data management legislation.

Fairness and Justice

The principle of "Fairness and Justice" enjoys a long tradition of academic debate in the language assessment literature and remains a strong area of academic interest and policy-making. It is also the most recurrent theme/principle found in the ethical AI documents reviewed by Fjeld et al. (2020), which is testimony to its continued importance.

While intrinsically related, "Fairness" and "Justice" may be defined as having different areas of influence. In the context of education and language assessment, fairness

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

refers to the imperative to provide equal treatment and respect for all individuals and subgroups of individuals, as well as the need to minimize bias through the technical qualities of the test (Davies, 2010; Deygers, 2019; Kunnan, 2013, 2018; McNamara & Ryan, 2011). The concept of (social) justice, on the other hand, is closer to that of social impact, and is concerned with maximizing the positive or neutral consequences of education and assessment practices on society, while minimizing their negative consequences (Deygers, 2019; Kunnan, 2013, 2018). In this sense, fairness may be understood as referring to test-internal systematic impacts on individuals/groups (e.g., bias) and justice as referring to test-external systemic impacts on the whole of society or sectors (e.g., access to technology) (McNamara & Ryan, 2011). In light of these definitions, the six considerations under "Fairness and Justice" address questions of diversity, inclusion, accessibility, and bias.

7.1	Advancement of fair and just educational goals	AI systems should treat each individual in an equitable and impartial manner and should be designed, developed, and implemented for the advancement of just educational and/or societal goals.
7.2	Diversity	AI systems should be designed, developed, used, and evaluated including a variety of individuals and perspectives. This includes the use of diverse datasets for training, curated to minimise biased applications or outcomes, and the inclusion of diverse expertise and viewpoints in AI design and development.
7.3	Accessibility	AI systems should be designed, developed, and deployed taking into account the needs (e.g., infrastructure, equipment, skills, and societal acceptance) of a wide range of individuals, e.g., "different age groups, cultural systems, language groups, persons with disabilities, girls and women, and disadvantaged, marginalised and vulnerable people or people in vulnerable situations" (UNESCO, 2021, p. 20), allowing equitable access and use of AI.
7.4	Mitigation of algorithmic biases	Existing bias in the AI algorithm should be mitigated to prevent discriminatory impacts and should be promptly rectified if identified throughout the AI system's lifecycle. That includes using non-static datasets which closely reflect changing values and user populations.
7.5	Quality of generated content	AI systems used for content generation should be designed, developed and used with robust "guardrails" in place to reduce the possibility of generating offensive,[1] biased, inaccurate, or otherwise harmful outputs.
7.6	Protection for vulnerable groups	Special considerations and safeguards need to be in place when vulnerable populations (e.g., children, people with disabilities, marginalized groups) interact with AI systems.

Security and Safety

The principle "Security and Safety" contains two considerations, which could be considered to be on the more technical side of the ethical debate, as they require that the designers and developers of AI systems introduce technical safeguards against external attacks and system-internal misuse. The "Security and Safety" considerations are formulated to minimize the possibility of external or internal risks affecting those who use AI systems or are subject to AI-based decisions.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

8.1	Safety	AI systems should be designed, developed, and used with safety checks in place that ensure the system works as intended and the chances of unintended harms are minimized.
8.2	Security	AI systems should be robust against malicious attacks that threaten the security of the system or the data it processes. When intended for assessment purposes, AI systems should be designed, developed, and used so that every part of the assessment process is protected against external threats. This includes securing items, test versions, test-taker identification, and remote proctoring systems, in addition to securing any data acquired or derived from these processes.

Transparency and Explainability

The "Transparency and Explainability" principle comprises another set of two closely related yet different concepts of a technical nature. "Transparency" refers to the need to clearly and openly communicate to stakeholders the use and functioning of AI systems throughout their lifecycle, to the extent that it does not compromise the security of the system. "Transparency" is not, therefore, a binary quality: different levels of disclosure may be suitable in different contexts and for different purposes. For example, it would be expected that a higher level of disclosure would be provided to knowledgeable auditors or monitors who are in charge of evaluating the AI system; while lower levels of disclosure may be expected for the general public, as revealing certain details of the algorithms used may expose the system to adversarial attacks. Despite these nuances, and the tensions derived from the level of control different actors have over the use of

9.1	Transparency in algorithms	Clear and understandable information regarding the AI system's development, training, operations, and deployment should be made available to auditors and to the general public, where appropriate. Providers of AI systems should also identify scenarios where the disclosure of information about the algorithm may expose the system to malicious exploitation, compromising its security and unnecessarily exposing users to negative outcomes.
9.2	Transparency in data	Data collection, processing, analysis, and reporting processes should be transparent. This entails requesting users of AI systems to give informed consent and providing all users with clear information about data ownership, accessibility of data, and the purposes for which data will be used.
9.3	Explainability/ explicability	The AI-based outputs (what the system is doing) and the mechanisms used to reach these outputs (why/how the system is doing it) should be translated into clear and comprehensible information that is understood by individuals with varying levels of technical expertise. Any accuracy/explainability trade-offs should be explicitly disclosed.
9.4	Informed users	Users of AI systems should be sufficiently informed about their interactions with AI systems so that they can evaluate the consequences of their use. That includes: information (where applicable) when AI makes a decision about an individual, when interacting with AI, and when content is created through GenAI.
9.5	AI literacy – learners and educators	Learners and educators should be trained, informed, confident, and discerning users of AI systems through appropriate AI literacy strategies.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

data, organizations who develop and/or use AI technologies should establish practices that acknowledge and protect intellectual property rights.

"Explainability" may be conceptualized as the requirement that the technical aspects, operations, and decisions of AI systems are explained in such a way that humans with different levels of technical knowledge and expertise may understand them. Similar to the principle of "Transparency," different AI systems may allow for varying levels of explainability. Large language models (LLMs), like the ones powering some of the more well-known GenAI applications, are known to be complex "black boxes" whose inner functioning is beyond explainability. When the use of such models is deemed to be necessary due to, for example, their ability to better predict outcomes, these compromises should be adequately recorded and communicated.

The five considerations contained in "Transparency and Explainability" address concerns surrounding the need for openness and communication of the use and technical operation of AI systems.

Assessment standards

The principle "Assessment standards" groups five considerations that most directly apply to the domain of language assessment. Best practice frameworks within L2 assessment and educational measurement are foundational to these "Assessment standards" considerations (AERA et al., 2014; ALTE, 2020; ILTA, 2020), as well as conceptualizations of justice, fairness, and validity within language assessment (Deygers, 2019; Kunnan, 2013, 2018; McNamara & Ryan, 2011). We have defined distinct considerations which pertain to the fundamentals of language assessment, such as validity, test design, development, delivery, scoring, and psychometric quality of the test. It is important to note that these considerations are not mutually exclusive and there is conceptual overlap between them; for example, the principle of validity can be seen as an overarching concept subsuming all other considerations. Our aim is not to question this conceptual hierarchy, but to highlight key (inter-related) assessment considerations. The aspects of validity we have chosen to include here are the most specific to language assessment; other important aspects of relevance for assessment (e.g., test security, results reporting, accessibility) are included under the other principles presented above.

10.1	Validity	Any outcomes or predictions made by an AI system in an assessment context should support the intended purposes of the assessment and avoid or at least minimize construct underrepresentation and the measurement of construct-irrelevant factors.
10.2	Design and development	Designers of AI-based assessments should consider the positive, negative, and unintended impact of AI on psychometric features of the test. This includes, for example, the impact of using AI-enhanced items and tasks, adaptive test designs, automated item creation and calibration, and dynamic test construction processes. AI-based assessment design should aim to expand accessibility through universal test design and understand its implications for test-takers. More broadly, AI-enhanced assessment design and development should maximize the integration of learning and assessment.

(*Continued*)

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

(*Continued.*)

10.3	Delivery	AI-enabled test delivery (e.g., web-based, offline, mobile) and associated concerns (e.g., identity verification, aberrant response detection, proctoring) should uphold assessment robustness. Unauthorized access to data and content must be prevented and plans for test disruptions need to be in place. Interoperability (i.e., accurate exchange of data between different systems) needs to be robustly handled within the AI test ecosystem.
10.4	Scoring	The scoring model needs to be informed by appropriate training data, a range of scoring features which represent the test construct and a systematic evaluation framework for automarker accuracy. The degree of scoring model change in operational use needs to be controlled in order to ensure that scoring rules are applied consistently.
		The choice of scoring approaches, for example, fully AI or hybrid AI/human, needs to be guided by clear and robust rules, metrics and regulations.
10.5	Psychometric quality	The outcomes of AI-based tests need to be based on reliable measurements in order to ensure accuracy of predictions and measurement fairness. Ongoing gathering of validation evidence needs to be in place to regularly verify the claims behind the intended uses of the AI assessment.

Ethical AI principles: from theory to practice

We now turn to a brief overview of considerations which need to be front-of-mind for stakeholders when engaging with and/or implementing the principles and considerations we have discussed so far. We do so through hypothetical scenarios typical of language assessment and illustrative guiding questions that highlight areas which require attention from an ethical AI perspective. In the interest of succinctness, we have been selective in the guiding questions we pose below, and mainly present questions that have a higher degree of domain-specificity. This does not mean that broader principles and considerations, for example, ones related to "Governance and Stewardship," "Accountability," or "Human-centricity" are less important. Rather, we assume that such general principles and considerations are foundational to all scenarios below, and we have instead put the spotlight on instances of direct application to language assessment.

Using AI for content creation

The recent widespread use of GenAI has opened new possibilities for assessment providers to create tasks and items at speed and at scale. For this scenario, our focus is on a test within a communicative language assessment paradigm designed to include a broad range of tasks representative of the domain (e.g., short and multi-paragraph reading texts, multi-speaker audio and transcripts, and prompts for extended writing and speaking tasks). The AI-generated materials need to accurately measure different ability levels and cover topics which are appropriate for the international test-taker cohort of this hypothetical test. The materials need to be on a par with the content developed by subject matter experts and reach acceptable quality standards, such as

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

domain representation, suitable topics, and appropriate difficulty levels. (For an example of a detailed overview of the content development considerations by subject matter experts, see Galaczi & Ffrench, 2011).

- How appropriate are the selected topics in terms of domain representation, topic suitability and difficulty levels? [*Assessment standards*]
- How reliable are the estimations of item and task difficulty? [*Assessment standards*]

How transparent is the AI-based content development approach to relevant stakeholders? [*Transparency and Explainability*]

Are there adequate mechanisms, tools and training processes in place to ensure human review and feedback in key stages of the process? [*Human control of technology*]

What copyright oversight mechanisms are embedded in order to ensure that the intellectual property rights of content creators are respected? [*Fairness and Justice*]

Is the AI model adequately trained to avoid or minimize inaccurate/inappropriate content and identifiable biases against certain groups? [*Fairness and Justice*]

Using AI for marking

This scenario focuses on another key aspect of L2 assessment – the marking of extended speaking and writing test-taker responses. Currently, predictive machine learning and GenAI models are used to assess responses of this type, typically focusing on linguistic features such as punctuation in writing, pronunciation and fluency in speaking, as well as accuracy and complexity of grammar, breadth of vocabulary, and discourse management (see Xi, 2023, for an in-depth theoretical discussion of the use of AI in this domain; see also Xu et al., 2024, for an overview of key considerations for the practical application of automarking).

- What scoring features are extracted to inform an automarked score and what is the degree of accuracy? [*Assessment standards*]
- What construct do the automarked features represent, and how relevant and representative are the features of the target ability and use domain? [*Assessment standards*]
- How is the potential for malpractice minimized? [*Security and Safety*]
- How explainable are the AI-based decisions? [*Transparency and Explainability*]
- What types of data has the AI system been trained on to minimize algorithm bias? [*Fairness and Justice]*
- What is the role of examiners in the scoring process? [*Human control of technology*]

Using AI for accessible assessment

The final scenario showcases the use of AI to enhance test accessibility. Test accessibility has started to enjoy more visibility in the last decade, especially related to neuro-diverse test-takers (as compared to test-takers with visible disabilities, who have been the focus

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

of research attention in the past). This new and important area of theoretical and practical interest has brought to the fore the positive/neutral/negative impact of AI on test accessibility, fairness and justice considerations (Kormos & Taylor, 2020).

- What are the test design features aimed at enabling accessibility, and are they backed by theoretical and cognitive evidence? [*Assessment standards*]
- What effect on test-taker performance do relevant task design features have? [*Assessment standards*]
- What data have the underlying AI models been trained on to mitigate unfair treatment of neurodivergent test-takers? [*Transparency and Explainability*]
- What is the level of risk that the AI algorithms used (e.g., for scoring or adaptivity) might disadvantage neurodivergent test-takers? [*Fairness and Justice*]
- What procedures are in place for examiners to monitor and intervene in situations where AI models are falling short of the expected standard? [*Human control of technology*]

Final remarks and future directions

The principles and considerations we have proposed in this chapter are both an end and a beginning. They are the culmination of the research we carried out in order to distil principles for the ethical use of AI in the domain of language assessment. At the same time, we see these principles as the beginning of further debate and insights from practical implementation, as different individuals and organizations apply the suggested principles to their specific context (i.e., move into the most context-specific layer in Figure 1). To do that, this next phase of academic scholarship and practical insights needs to successfully grapple with two complex challenges.

The first is the gap between conceptual principles and practical implementation. In their discussions, Jobin et al. (2019) and Fjeld et al. (2020, p. 66) note this divergence and the "wide and thorny gap" between identifying high-level principles and actually implementing them, arguing that translating principles into practice needs to be an important next step for those involved in and impacted by AI in the global community. Deygers (2019) echoes this concern in the domain of language assessment, noting that codes of practice and standards of quality are currently mostly focused on "technical and operational" considerations (p. 23) and not ethical ones. He argues that ethical obligations need to become part of professional codes of practice and standards of quality in order to move from principle formulation to practical implementation. Our personal experience as researchers working in an examination organization frequently brings us face-to-face with this tension between theory and the very practical business of producing tests at a global scale. Explicitly embedding high-level principles will enable the language assessment professional community to deal systematically with the practicalities of embedding ethical regulations for the benefit of all stakeholders impacted by tests. Appropriate training and professional development on the ethical use of AI play a key role as well, since principles and practices are, after all, dependent on trained personnel who are supported through organizational policies in implementing them. We hope that the principles and considerations we have identified here will be a useful foundation for addressing inherent implementation challenges,

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

guided additionally by the influential professional standards offered by AERA et al. (2014), ALTE (2020), and ILTA (2020).

The second key challenge is situated at the intersection between the powerful affordances of AI and the equally significant risks and concerns it poses to human capabilities and values. Discussions of inherent *tensions* are starting to emerge, as seen for example, in Briggs (2024) who argues that one of the key principles in debates of ethical AI use – "transparency" – may, in fact, be "incompatible with the competitive marketplace that is funding" it (p. 694). Similarly, Whittlestone et al. (2019) delve into these tensions in the broader social context of AI application and provide a number of insightful illustrations of the tensions between the values underlying ethical principles of AI and the necessary trade-offs. They point to, for example, the trade-off between the convenience enabled by AI versus the promotion of human agency. Arguing along similar lines, Nguyen et al. (2023) voice an oft-repeated concern that the impact of the automation offered by AI might "reduce learners' interactions with others and their ability to cultivate individual resourcefulness, metacognition, self-regulation, and independent thought" (p. 4235). Other tensions tap into the creation of short-term individual-level benefits versus longer-term collective values, increased personalization versus individuals' data privacy and autonomy, or issues of access and equity across different groups (UNESCO, 2023; Whittlestone et al., 2019). We hope that the principles and considerations we have proposed here will be a spark for uncovering and exploring ways of resolving or minimizing these challenges and tensions.

The complexity of such endeavors is not to be underestimated, especially since the speed of AI innovation which we are currently witnessing is at odds with the typically conservative pace of reform in the education sector, including in large-scale assessment (UNESCO, 2023).

A related issue can be seen in practices that are becoming widely used but which violate ethical principles, and which would be complicated to reverse once they become embedded. Difficult – and perhaps currently impossible to answer – questions emerge. For example: considering that the latest generation of AI tools were created using internet data without regard for intellectual ownership, would all users of such AI tools, even if following ethical guidance, be complicit in violating transparency principles? Another example can be found in the environmental impact of AI tools (and specifically GenAI) and the emerging evidence that the LLMs which power GenAI systems are not environmentally sustainable due to their massive energy needs (see, for example Caines et al., 2023). Without robust empirical evidence that LLMs perform substantially better than traditional AI models, how can informed decisions be made whether the additional computing and environmental costs are justified or not? As we grapple with these complex challenges, as a first step we need to understand the underlying dilemmas and values in order to reach an informed understanding of the trade-offs involved. Addressing these dilemmas in depth is beyond the scope of our chapter, but we hope that our suggested framework will serve as inspiration for further discussion, research and debate in language assessment, on a par with similar discussions of tensions offered by Jobin et al. (2019) and Whittlestone et al. (2019) within a broader societal lens.

Through writing this chapter and exploring the wealth of emerging literature on the ethical use of AI in education and assessment, including language assessment, we

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

have become all too aware of the partial view we have presented in our discussion and our positionality as researchers from specific professional, disciplinary, and cultural contexts. Whittlestone et al. (2019) offer an insightful discussion of the challenges present in engaging in such principle-building endeavours. The authors note challenges stemming from:

Terminological ambiguity: for example, are "black-box" AI models not transparent because algorithms are proprietary, or because of the exceptionally high level of algorithm complexity for all users and even for AI experts?

Differences in disciplinary terminology, for example, the meaning of the concept of "bias" differs in the statistical and social sense.

Cultural tendencies, for example, the concept of "privacy" would have different connotations in Western ethical traditions where individual privacy is viewed positively, in contrast to Eastern traditions where the privacy of collectives is typically seen as more important than that of individuals.

Complexity of conceptual interpretations, for example, the concept of fairness has layers of meaning embedded in philosophy, mathematics, and social sciences, and all of these interpretative lenses need to be considered for an in-depth treatment of this concept. (See also Ryan, 2024, who offers a discussion along similar lines in his critique of human-centred AI through a philosophy-of-technology lens).

These challenges also hold relevance for the data sources which formed the basis of our exploration. Their selection was based on a systematic rationale (as overviewed in the Method section), and yet all the policy documents we included are inevitably shaped by cultural and individual perspectives. We believe that the confluence we often discovered in concepts and themes across the range of diverse perspectives lends a level of universal robustness to the set of proposed principles. However, we must acknowledge that the current availability of mostly Western-based documents presents a limitation of our investigation, which we hope will be minimized through the future development of a broader set of guidelines and documents from a global perspective.

Notwithstanding the challenges we have discussed, we are cautiously optimistic that the language assessment professional community is well placed to develop an informed awareness of potential benefits and harms associated with AI in this domain. We hope that the ethical principles and considerations we have presented are a starting point for further theoretical and practical endeavours and continuous refinement as they are put into practice alongside other professional guidelines and standards. It is through such a focus on theory, policies, and practice that the language assessment community will be able to adopt an "ethical-by-design" approach to developing AI tools for language assessment which benefits all stakeholders at individual, institutional and systems levels.

Note

1. The question of what constitutes "offensive" content is dependent on multiple societal factors, including but not limited to geopolitical and historical forces shaping our perception of what is socially acceptable or not. Nevertheless, documents like UNESCO's guidance for GenAI particularly emphasise the dangers of large language models producing "discriminatory and other unacceptable language" (2023, p. 16) and the need to implement moderation mechanisms to minimise these outputs.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

References

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2021). Artificial intelligence ethics guidelines for K-12 education: A review of the global landscape. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part II* (24–28). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aiken, R., & Epstein, R. (2000). Ethical guidelines for AI in education: Starting a conversation. *International Journal of Artificial Intelligence in Education*, *11*, 163–176.

AlgorithmWatch. (n.d.). *AI Ethics Guidelines Global Inventory*. https://inventory.algorithmwatch.org

ALTE. (2020). *ALTE Principles of Good Practice*. https://www.alte.org/Materials

Association of Test Publishers. (2022, January). *Artificial Intelligence Principles*. https://www.testpublishers.org/ai-principles

Australian Government, Department of Industry, Science and Resources. (2024). *Voluntary AI Safety Standard*. https://www.industry.gov.au/sites/default/files/2024-09/voluntary-ai-safety-standard.pdf

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2023). Thematic analysis. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Research designs: Quantitative, qualitative, neuropsychological, and biological* (2nd ed., pp. 65–81). American Psychological Association. https://doi.org/10.1037/0000319-004

Briggs, D. C. (2024). Strive for measurement, set new standards, and try not to be evil. *Journal of Educational and Behavioral Statistics*, *49*(5), 694–701. https://doi.org/10.3102/10769986241238479

Burstein, J. (2023). *The Duolingo English Test Responsible AI Standards*. https://scite.ai/reports/duolingo-english-test-responsible-ai-4LJW88MP

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., Yuan, Z., Elliott, M., Moore, R., Bryant, C., Rei, M., Yannakoudakis, H., Mullooly, A., Nicholls, D., & Buttery, P. (2023). On the application of Large Language Models for language teaching and assessment technology *arXiv:2307.08393*. arXiv. https://doi.org/10.48550/arXiv.2307.08393

Cambridge University Press & Assessment. (2023, May). *English language education in the era of generative AI: our perspective*. https://www.cambridgeenglish.org/Images/685411-english-language-education-in-the-era-of-generative-ai-our-perspective.pdf

Capel, T., & Brereton, M. (2023). What is human-centered about human-centered AI? A map of the research landscape. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–23. https://doi.org/10.1145/3544548.3580959

Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171–176. https://doi.org/10.1177/0265532209349466

Deygers, B. (2019). Fairness and social justice in English language assessment. In J. Gao (Ed.), *Second handbook of information technology in primary and secondary education* (pp. 1–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-58542-0_30-1

ETS Research Institute. (n.d.). *Responsible use of AI in assessment*. https://www.ets.org/Rebrand/pdf/ETS_Convening_executive_summary_for_the_AI_Guidelines.pdf

European Union. (2024, June). *EU Artificial Intelligence Act*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

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*. Berkman Klein Center for Internet & Society. https://doi.org/10.2139/ssrn.3518482

Galaczi, E., & French, A. (2011). Context validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (Vol. 30, pp. 112–170). UCLES/Cambridge University Press.

General Data Protection Regulation. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council*. https://gdpr-info.eu/

Holmes, W. (2023). *The unintended consequences of artificial intelligence and education*. Education International.

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press

Holmes, W., & Porayska-Pomsta, K. (2022). *The ethics of artificial intelligence in education: Practices, challenges, and debates* (1st ed.). Routledge. https://doi.org/10.4324/9780429329067

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

Holstein, K., & Doroudi, S. (2021). Equity and artificial intelligence in education: Will 'AIEd' amplify or alleviate inequities in education? *arXiv:2104.12920*. arXiv. https://doi.org/10.48550/arXiv.2104.12920.

IBM. (2022). *Everyday ethics for artificial intelligence*. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

ILTA. (2020). *Guidelines for Practice*. https://www.iltaonline.com/page/ILTAGuidelinesforPractice

The Institute for Ethical AI in Education. (2021). *The ethical framework for AI in education*. https://www.ai-in-education.co.uk/resources/the-institute-for-ethical-ai-in-education-the-ethical-framework-for-ai-in-education

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kishimoto, A., Régis, C., Denis, J.-L., & Axente, M. L. (2024). Introduction. In C. Régis, J.-L. Denis, M. L. Axente, & A. Kishimoto (Eds.), *Human-centered AI: A multidisciplinary perspective for policy-makers, auditors, and users* (1st ed., pp. 1–10). Chapman and Hall/CRC. https://doi.org/10.1201/9781003320791-1

Kormos, J., & Taylor, L. B. (2020). Testing the L2 of learners with specific learning difficulties. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 413–421). Routledge.

Kunnan, A. J. (2013). Fairness and justice in language assessment. In A. J. Kunnan (Eds.), *The companion to language assessment* (pp. 1098–1114). John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla144

Kunnan, A. J. (2018). *Evaluating language assessments* (1st ed.). Routledge. https://doi.org/10.4324/9780203803554

Leslie, D., Rincon, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, S., Burr, C., Aitken, M., Katell, M., & Fischer, C. (2024). *AI ethics and governance in practice: An introduction*. The Alan Turing Institute. https://doi.org/10.2139/ssrn.4731635

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly*, *8*(2), 161–178. https://doi.org/10.1080/15434303.2011.565438

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), 4221–4241. https://doi.org/10.1007/s10639-022-11316-w

Ryan, M. (2024). We're only human after all: A critique of human-centred AI. *AI & SOCIETY*, *40*(3), 1303–1319. https://doi.org/10.1007/s00146-024-01976-2

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

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

UNICEF. (2021). *Policy guidance on AI for children*. https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. Nuffield Foundation.

Xi, X. (2023). Advancing language assessment with AI and ML–Leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, *20*(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488

Xu, J., Schmidt, E., Galaczi, E., & Somers, A. (2024). *Automarking in language assessment: Key considerations for best practice*. Cambridge University Press & Assessment. https://doi.org/10.17863/CAM.117098.

Cite this article: Galaczi, E., & Pastorino-Campos, C. (2025). Ethical AI for language assessment: Principles, considerations, and emerging tensions. *Annual Review of Applied Linguistics*, *45*, 294–314. https://doi.org/10.1017/S0267190525100081

https://doi.org/10.1017/S0267190525100081 Published online by Cambridge University Press