BRIEFING SHEET

www.ukalta.org info@ukalta.org @UKALTA2

No. 2/23, August 2023

ARTIFICIAL INTELLIGENCE (AI) IN LANGUAGE ASSESSMENT

This Briefing Sheet is written by UKALTA, the UK Association for Language Testing and Assessment, to answer the following questions.

- What is AI and why is it relevant to language assessment?
- What types of AI are being used in language assessment?
- What are the potential benefits of AI in language learning and assessment?
- What are the risks and concerns in using AI for assessment purposes, and how can they be mitigated?

AUTHOR

Nick Saville – Cambridge University Press & Assessment, University of Cambridge

NOTE: The content of this briefing sheet reflects the author's perspective as a language assessment professional working in the UK and internationally.

For reasons of brevity and space, referencing is kept to a minimum, but the author is happy to supply a full set of references on request.

WHAT IS AI AND WHY IS IT RELEVANT TO LANGUAGE ASSESSMENT?

The term AI originated in the 1950s, but only came to the attention of the wider public in the 1990s (when IBM's Deep Blue computer defeated world champion, Kasparov, at chess).

The Organisation for Economic Co-operation and Development (OECD, 2019) describes AI positively as *a general-purpose technology* 'that has the potential to improve the welfare and well-being of people' and can 'help respond to key global challenges'.

More specifically, it refers to human-like intelligence exhibited by machines that provide tools and systems managed by humans.

In practice, AI systems interact with the world through *capabilities and behaviours* we think of as human.

- ***Capabilities*** are typically specific and narrowly defined and include understanding natural language, solving problems and learning from experience.
- ***Behaviours*** include abilities to perform well-defined tasks, such as answering questions, providing information and making recommendations - all of which are tasks used for assessment purposes.

In the field of language testing, the technology is not entirely new. Computer-delivered language assessments have existed since the 1990s with AI being used to carry out specific assessment functions, such as the scoring of writing. However, in recent years AI capabilities and behaviours have been deployed more widely in long-standing assessment systems (e.g., ETS, IELTS, Cambridge), as well as in on-line assessments that are "born digital' (i.e., that have never existed in a paper-and-pencil format, such as Pearson and Duolingo tests).

More generally in society, AI-enabled systems are now part of everyday life. They are ubiquitous in many familiar tech brands (Google, Netflix, Siri, etc.) and are used to support our public services, including health care, transport, criminal justice and education. In all cases, the AI is embedded in *decision-making* processes or in *making recommendations* that can affect individuals and society.

OpenAI's ***large language model*** (LLM), ChatGPT (November 2022), has attracted widescale attention because of its potential to disrupt the way things are done in many fields, including education and assessment.

Politicians and policy-makers are now considering the potential benefits of this kind of ***generative AI*** (see below), as well as confronting the significant risks posed by it. The pace of change, partly driven by the commercial competition between the biggest names in the AI world, has become a major cause for concern at an international level. Consideration is being given to high-level regulatory guidelines and guardrails (e.g., a global convention or protocol) to ensure AI safety, with suggestions that AI should be treated in a similar way to industries that are highly regulated, such as pharmaceuticals or aviation (including licensing and inspection procedures).

The challenge faced by non-specialists is to understand the AI concepts and related technical features well enough for their own context and purposes.

WHAT TYPES OF AI ARE BEING USED IN LANGUAGE ASSESSMENT?

In language assessment hitherto, AI capabilities and behaviours have been deployed to improve the efficiency and effectiveness of existing assessment systems, but increasingly they are being deployed to develop innovative approaches that have the potential to change the way assessment will be carried out in future.

AI may offer opportunities to streamline human processes and reduce costs based on automation, as well as broadening access to assessment opportunities.

Central to this are two related AI concepts used in developing the relevant AI systems: ***Machine Learning*** (ML) and ***Natural Language Processing*** (NLP).

AI systems based on ML and NLP are already being used for the following purposes:

- assessing productive language skills;
- providing accurate evaluations of proficiency levels;
- analysing language features for diagnostic purposes;
- providing formative feedback.

It is important to develop a basic understanding of these two concepts in order to evaluate how the systems work that deploy them, and the value they add to language learning and assessment.

ML is a subset of AI that uses statistical techniques to enable machines to improve what they do through experience – hence the concept of the machine doing the learning. ML systems are trained using large amounts of data, e.g., written or spoken language in the case of assessment. Some significant challenges and risks of AI-based systems from the social perspective are related to the collection and uses of that data (see below).

NLP is concerned with interactions between computers and human language, and how computers are programmed to process and analyse large amounts of natural language data, both written and spoken. NLP is central to many applications of ML relevant to language learning and assessment systems, including, part-of-speech tagging, parsing, speech recognition, text-to-speech synthesis, and machine translation.

In building an ML model, ***training data*** is the foundation for the model's learning and decision-making process. The data provides *input* that is fed into the model and the *output* is the prediction based on the data. The goal in training the model is to achieve *a mapping from inputs to outputs*, such that it can *accurately predict* the output for new, unseen inputs.

The quality and quantity of the training data play a crucial role; the more data a model is trained on, the more it can learn patterns and relationships in the data. It is important that the training data is diverse, representative and unbiased to avoid biases in the data and in the predictions.

The choice of algorithm is an important consideration.

An ***algorithm*** is a finite sequence of instructions to solve a problem or perform a computation, and a variety of them are used in AI for specific types of ML. How the algorithms work and how they can be explained to those who use the systems they deliver are important considerations, especially as some algorithms are inherently more difficult to understand and explain than others.

Algorithms that are readily interpretable are referred to as **'White Box'** or Transparent models because the internal workings are explainable in ways humans can understand. Others, such as ***deep learning*** algorithms, cannot be easily understood and interpreted and are known as **'Black Box'** or Opaque AI.

Deep learning algorithms, based on ***artificial neural networks*** are inspired by the structure and function of the human brain; they consist of multiple layers of interconnected "neurons" which process and transform input data. A ***deep neural network*** (DNN) has a large number of hidden layers. Typically, the more *hidden layers* a neural network has, the more complex features it can extract, and the more powerful it becomes.

Until recently, ML was largely limited to *predictive models*, used to observe and classify patterns in content. For example, a traditional machine learning problem was to use texts of written language (such as essays) as input and then to predict the level of proficiency (e.g., the CEFR level). In other words, to address the problem of accurately scoring and classifying the texts according to the features of the writing.

Generative models of AI go beyond this and can *generate* human-like text through seemingly natural interactions with users. They can perform many different kinds of task and can be fine-tuned by users to meet their specific needs. ChatGPT is an example of this kind of *generative model of AI* where GPT stands for ***generative pretrained transformer.***

WHAT ARE THE POTENTIAL BENEFITS OF AI IN LANGUAGE LEARNING AND ASSESSMENT?

Early uses of AI in language assessment included the automation of rating systems for assessing writing, and in recent years this has been extended to other skills, including speaking, and to other aspects of the assessment cycle, including ***automated task generation*** (ATG) and ***online delivery***, e.g., ***at-home test taking*** with ***remote proctoring.***

This trend is set to continue as the latest educational technologies using AI - especially generative models - become more widely available and integrated into existing practices.

Automated systems can not only increase the ***efficiency of assessment systems,*** but can also improve their *reliability* and extend the *coverage of the constructs.* The emerging field of ***computational psychometrics*** extends traditional psychometric and statistical techniques to account for the measurement qualities of AI-based systems. This can advance the field

by incorporating the new data sources and by deploying advanced computational models to provide a more comprehensive understanding of learners' cognition and behaviours.

AI has the potential to be transformative in supporting a wider range of assessment *purposes* and in delivering ***innovative constructs*** and *task types,* e.g., bringing language teaching, learning and assessment together more effectively in an integrated way (***learning-oriented assessment, scenario-based assessmen***t, etc.) or to promote plurilingual concepts of assessment. It might also help to improve accessibility for those with specific assessment needs.

Capturing and using specific types of digital data ***as evidence of and for learning*** will be central if these benefits are to be realised, e.g., for the following purposes:

- *Personalized* learning experiences and materials based on a learner's individual needs and abilities.
- *Intelligent tutoring* to provide adaptive, interactive feedback in real time.
- *Learning analytics* on learner performance and behaviour, providing insights into learning and helping teachers make data-driven decisions.
- *Virtual and augmented reality* to create immersive, interactive learning experiences that can enhance learners' engagement and learning outcomes.

An emerging issue is the interplay between humans and machines, i.e., ***hybridization*****.**

Hybrid approaches include:

- Rating systems combining AI and human scorers.
- AI-based tutoring systems combined with human teachers.
- Assessment scenarios combining face-to-face and on-line tasks.

WHAT ARE THE RISKS AND CONCERNS IN USING AI FOR ASSESSMENT PURPOSES?

As the world begins to consider high-level regulatory guidelines and guardrails at an international level (e.g., a global convention or protocol), the discussion in this briefing sheet is set more narrowly within educational contexts and especially for language assessment that uses digital technology.

In order to build trust in these increasingly complex AI systems, a better understanding of the core concepts is needed across a wider range of societal contexts.

There are already comprehensive guidelines for technology-based assessments, e.g., from the International Test Commission and Association of Test Publishers.

- Guidelines for Technology-based Assessment. (2022) https://www.testpublishers.org/assets/TBA Guidelines final 2-23-2023 v4.pdf

And for language testing there are professional ***Codes of Ethics and Good Practice,*** such as those already developed by ILTA, EALTA and ALTE, which can be changed or updated to handle the emerging risks and pitfalls associated with AI.

Seeking the views and experiences from their members is an important first step. In 2023 ILTA initiated a consultation process to revise its ***Code of Ethics***:

- ILTA https://www.iltaonline.com/page/CodeofEthics

ALTE and EALTA are also revisiting their ***Principles of Good Practice*** and practical ***Guidelines.***

- ALTE https://pt.alte.org/resources/Documents/ALTE Principles of Good Practice Online version Proof 4.pdf
- EALTA https://www.ealta.eu.org/guidelines.htm

More generally, language professionals are encouraged to take part in these consultations and to seek ***interdisciplinary collaboration*** to gain better understanding of the emerging issues.

Known risks in using AI in language assessment are associated with the ***collection and use of personal data***, including concerns for the security and privacy of learners, and the potential for bias in outcomes. Other risks may occur when AI systems are not validated for specific uses, or the outcomes cannot be adequately explained to the users.

Bias is a long-standing concern in assessment and is important in machine-learning as the AI models depend on the data they are trained on; if the training data is biased, the model's output will also be biased. This kind of bias may be caused by *sampling error, incomplete data*, or it may exist *in the data itself*, reflecting biases in society (e.g., representation of gender or ethnic groups). The danger is that the AI perpetuates underlying inequalities and prejudices through the outputs of the system.

One of the main concerns in generative AI is related to the ***origin of the data*** that makes up the large language models. Also, despite the size of the data used to build the models, they are still prone to lapses or factual errors in the output (known as ***hallucinations***).

Other concerns related to deployment of AI in assessment include: the reduction of human interactions with test takers and loss of the 'human touch' in assessment processes; a dependence on technologies that offer no improvements in assessing the construct; and the Black Box problem, referred to above.

Explainable AI (XAI) and related dilemmas are central to this debate.

- How can the need for transparency be balanced against other factors that underpin the integrity of an assessment system?
- Should a system that is accurate but difficult to explain be preferred over a more transparent model that performs less well?
- Is full disclosure of algorithms feasible if they underpin proprietary systems and are considered commercially sensitive?

AI creates new opportunities for ***cheating and malpractice*** (e.g., plagiarism) in educational assessment where computer-delivered examinations are used. In this context, ChatGPT has opened up an on-going debate about the impact of generative AI in education. Should we deal with this by restricting the use of the technology, or should we embrace it and change our constructs of learning and assessment?

On the flip side, assessment providers have taken advantage of AI-based ***systems for detecting cheating*** (e.g., in AI-based proctoring) and are extending this approach in checking for plagiarism or use of 'ghost writers' (including generative AIs) in completing academic assignments and dissertations.

HOW CAN THE RISKS AND CONCERNS BE MITIGATED?

In the case of AI this requires a focus on the following concepts that extends the coverage in existing codes of practice for language assessment:

- ***Fairness:*** ensuring AI systems do not discriminate against certain groups of people based on factors such as race, gender, or age.
- ***Transparency:*** ensuring the decision-making processes of AI systems are transparent and explainable, so their outputs can be understood and evaluated.
- ***Accountability:*** ensuring those responsible for the design and deployment of AI systems are held accountable for their actions and decisions.
- ***Human rights:*** ensuring AI systems respect and protect human rights, such as privacy, autonomy, and non-discrimination.
- ***Safety and security:*** ensuring AI systems are safe and secure, and do not cause harm to individuals or society.
- ***Responsible data practices:*** ensuring data used to train AI systems is collected, handled and used responsibly and ethically.
- ***Value alignment:*** ensuring AI systems are aligned with human values and ethical principles, so their decisions and actions are in line with what is considered morally and ethically correct.

Ethical practices in language assessment enable risks and concerns to be addressed and mitigated.

As noted above, perspectives should be sought from a wide range of stakeholders in the development and deployment of AI systems to ensure they are designed and used in keeping with the concepts listed above.

It can be argued that AI systems should complement the roles that humans play rather than replace them entirely, and that a responsible approach to AI development and deployment should bridge both the technical and social aspects.

Human-in-the-loop AI (HITL) systems allow for the integration of human intelligence, knowledge, and decision-making with AI algorithms. HITL focuses mainly on the technical aspects and ensures stakeholder engagement in designing acceptable solutions.

Society-in-the-loop AI (SITL) is an extension of HITL, taking into account the societal impact to ensure that it aligns with societal values and ethical principles.

The combined approach means that known risks can be addressed and emerging issues considered in a timely and collaborative way, bringing together AI specialists with domain experts, practitioners and decision-making bodies (policy makers, regulators etc.) to ensure that AI systems are:

- ***Interpretable*** by developing understandings of how the technology works.
- ***Explainable*** by providing accessible explanations so users can understand why/how an outcome was reached.
- ***Transparent*** by providing information about the data and algorithms used.
- ***Justifiable*** by providing valid arguments in support of outcomes.
- ***Contestable*** by providing information enabling stakeholders to challenge an outcome.
- ***Sustainable*** by developing awareness of the environmental impact of AI and how it can be managed.

Regulatory and legal processes contribute to this debate.

Under the EU's *General Data Protection Regulation* (GDPR), 'data subjects' have the right to request *human review* when automated decision-making is used, and there is an open debate about whether this regulation also contains a *'right to an explanation'* (c.f. XAI).

The **EU 's AI Act** (2023) is seen as a step towards regulating AI from a ***statutory perspective*** and elsewhere governments are engaged with leading tech companies in establishing ***voluntary safeguards,*** e.g., in 2023 the *World Ethical Data Foundation* released a voluntary framework for developing AI products safely.

Ethical AI is emerging as a field in its own right and ethical frameworks are being developed in various societal contexts, including education.

These provide useful information for leaders and practitioners. The *Institute for Ethical AI in Education* (2021), for example, promotes a learner-centric approach that seeks to empower educational leaders to make ethically informed decisions about AI on behalf of their learners. http://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf

SEE ALSO:

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

OECD. AI Principles Overview (2019).https://oecd.ai/en/ai-principles

Office of Educational Technology (2023). Artificial Intelligence and the Future of Teaching and Learning. https://tech.ed.gov/ai-future-of-teaching-and-learning/

The Alan Turing Institute. Understanding artificial intelligence ethics and safety (2019). https://www.turing.ac.uk/research/research-programmes/public-policy

World Ethical Data Foundation (2023). https://worldethicaldata.org/