

Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment

Luke Harding

Lancaster University

Introduction

The genesis of this special issue dates back to a symposium held at the Language Testing Research Colloquium (LTRC) in New York City in June 2023. Two years on, it is useful to reflect on that time because, in many respects, that conference represents an important inflection point in our field. LTRC2023 was the first in-person LTRC after the pandemic and so there was a general sense of a field coming back together, in one physical space, re-energised in a vibrant and cosmopolitan city. At the same time, we had all experienced the pandemic, and we had collectively managed changes in our work practices and the unprecedented uptake of technology-based solutions to high-stakes assessment problems posed by COVID-19 such as take-at-home tests and remote proctoring (Isbell & Kremmel, 2020). The field had already been on the path to a technological revolution before the pandemic, but in 2020 and 2021, the rapid adoption of new digital technology – described comprehensively in Ockey’s (2021) edited special issue of *Language Assessment Quarterly* – was clearly filtering into research priorities. And we were on the cusp of an AI revolution (Jang & Sawaki, 2025; O’Sullivan, 2023; Voss, 2024).

The title of that original symposium seems quaintly outdated now: *Technology in diagnostic and formative assessment: Friend or foe*. Early discussions of plans for the

symposium indicated that the focus would be on opportunities and challenges in integrating technology into learning-oriented approaches to language assessment. There was no specific focus on AI, though some papers drew on that technology. However, other symposia at the same conference were overtly oriented towards AI, as were many paper presentations. My clear memory of many of the sessions, and the impression I shared with colleagues and students when I returned from the conference, was that AI was already shaping our field at a rapid pace.

It seems fitting, then, that the title of this special issue reflects the developments of the intervening two years: *Advancing language assessment for teaching and learning in the era of AI revolution: Promises and challenges*. The most revolutionary technology – Generative AI (GenAI) – is now easily available and is being put into practice on a wide scale (see O’Sullivan, 2023; Xi, 2023) with uses ranging from task and item development (e.g., Aryadoust et al., 2025; Runge et al., 2024), to automated scoring and feedback systems (Latif & Zhai, 2024), to efficient methods for item parameter estimation (Hao et al., 2024). More traditional AI has been used for some time in a range of scoring and security applications including remote proctoring and fraud detection (Isbell et al., 2023). The ease of access to tools like ChatGPT has also spawned recent debates about the extent to which they should be available to test-takers (Voss et al., 2023). AI is not just changing practices at the individual level but is also having macro-level organisational effects. During the period in which I wrote this piece, Duolingo – the creators of the Duolingo English Test – announced that they will become an “AI-first” company, with a post on LinkedIn quoting an email from the Duolingo CEO stating that “AI is already changing how work gets done. It’s not a question of if or when. It’s happening now” (Duolingo, 2025).

Language testing has not adopted AI uncritically (see Brunfaut, 2022; Jin & Fan, 2022; Xi, 2023), but it has nevertheless begun to impact on our practices with an irresistible

force. The appeal of AI is clear for a field in which practicality and resource constraints around development, administration and scoring have often created challenges for test developers. Yet, it is precisely because of its very rapid uptake that AI requires us to reflect critically on its impacts and consequences (see Moor, 2005). This special issue, then, is timely. It provides – through the seven articles that comprise this collection – an opportunity not only to consider how AI can be leveraged and applied to help solve pressing problems in our field but also to reflect on what kind of research is being conducted with AI: what motivations and values underpin that research. This allows us to identify more substantial issues that might need to be addressed.

In this article, I will first draw on critical perspectives of AI, and a recent re-evaluation of the term “Luddite,” to sketch a dystopian vision for the use of AI in language assessment, imagining a worst-case scenario for the integration of AI in our practices. Next, I will point to work that has been done (or is in progress) in our field to establish guidelines for responsible use of AI, arguing that there appears to be a shared desire to achieve a more utopian vision. I will then discuss the seven articles in this special issue, with specific reference to the extent to which their aims align with aspects of that utopian vision. Finally, I will consider two deeper philosophical questions that will pose challenges for the field as it continues to integrate AI into assessment practice.

Dystopian visions

In wider discussions of Educational Technology (EdTech), there has been a healthy scepticism and critique of the AI revolution. Several notable critical voices have pointed to specific problems associated with the rapid uptake of digital technology – and GenAI in particular – within educational settings. Ben Williamson, for example, a prominent voice in the EdTech field, pointed out that enthusiasm for AI in educational settings has increased the

presence and influence of private companies in public education, with increased integration of private platforms in schools. Williamson et al. (2023) noted that while this process – called “platformization” – occurs, “public schools are becoming increasingly reliant on corporate infrastructures to carry out many of their everyday functions, such as pedagogic activities, information management, behavioural and attendance monitoring, and assessment” (p. 3). The notion of loss of control is echoed, more boldly, by AI critics such as Ricaurte (2022), who argues that “hegemonic AI is becoming a more powerful force capable of perpetrating global violence through three epistemic processes: datafication (extraction and dispossession), algorithmisation (mediation and governmentality) and automation (violence, inequality and displacement of responsibility)” (p. 727).

Scepticism and wariness of the impacts of AI has also found expression in the sphere of popular culture. Around the time of the LTRC 2023 symposium, Ted Chiang – a US-based science fiction novelist with a particular interest in linguistics – wrote a widely disseminated article in the New York Times revisiting and reclaiming the term “Luddite.” “Luddite” is often used to refer, pejoratively, to those opposed to new technology and innovations in working methods. The word has its roots in the name of Ned Ludd, a (probably mythical) character in early nineteenth-century England, at the time of the Industrial Revolution. Ludd was portrayed as the leader of a movement in which textile workers destroyed new mechanised weaving machines. Writing against the backdrop of the growing influence of AI, Chiang pointed out that the Luddites’ actions were not motivated by fear of misunderstanding of technology itself; rather, as Chiang argued:

It’s helpful to clarify what the Luddites actually wanted. The main thing they were protesting was the fact that their wages were falling at the same time that factory owners’ profits were increasing, along with food prices. They were also protesting

unsafe working conditions, the use of child labor, and the sale of shoddy goods that discredited the entire textile industry. The Luddites did not indiscriminately destroy machines; if a machine's owner paid his workers well, they left it alone. The Luddites were not anti-technology; what they wanted was economic justice.

(Chiang, 2023)

The Luddites' unease was not with the technology itself, but with the inevitable shifts in power (and the potential harms) brought about by its introduction. Similar comments have been made by Suelette Dreyfus, a researcher of technology and AI, who has said:

People fear AI and machine learning because they think it's about a shift of power from the human to machine ... But actually, it's also a shift in power between the individual human and the organisation. And that becomes very important, because you have to think about how we will make the organisation accountable, what transparency requirements are there[?]

(Dreyfus, 2020, reported in da Dilva, 2021)

Like the Luddites, critics of AI are not anti-technology; they are concerned about just and ethical implementation. The common question that many AI critics raise is, without sufficient oversight, what incentives will powerful organisations have to use the vast computational capabilities of AI responsibly? If we apply this concern to language assessment, where language testing organisations already exert considerable power over the lives of test-takers (Shohamy, 2001), there are clear hazards. The ease with which generative AI can create content, provide quick scores, and analyse data provides a great (and perhaps irresistible) temptation for short-cuts – the avoidance of complexity, failure to engage in high-quality test

development and validation, and a lack of explainable processes. In high-stakes situations, such short-cuts could have grave consequences.

Given that frameworks for responsible innovation encourage the articulation of future scenarios (Stilgoe et al., 2013), it is useful to imagine what a worst-case, dystopian scenario might look like for future AI-driven language assessments if power, control and profit were the key driving factors. We could envisage mass-produced examinations, developed at a rapid rate, that are uniform, homogenised, and lacking any acknowledgement of the complexities of human communication or of different language varieties beyond the limitations of their own training data. Such exams would be disconnected with learning and teaching but widely adopted in classroom settings. The conceptual underpinnings would be atheoretical, based on expediency, and targeting narrow, easily measured constructs, using scoring systems for spoken and written production that are opaque and unexplainable, even to the few humans involved in the development process. Test-taker data – drawn from test performance and biometrics required for security – would be retained and sold on to third parties. There would be a proliferation of such assessments (given the ease of production) leading to more gatekeeping and more barriers for test-takers. This dystopian scenario would be made possible through the affordances of the technology combined with an absence of values and ethical intentions.

Responsible use of AI – towards a utopian vision

The deeply pessimistic view presented above may seem unlikely (even if some isolated elements may be recognisable even in current large-scale testing practices). However, this is mainly because language assessment as a field has a long history of grappling directly with issues of ethics and social responsibility (McNamara, 2000). Power, accountability, transparency and test-taker rights – all key aspects of the critiques presented above – have

been robustly addressed in prior work from critical language testing scholars that predated the AI revolution. For example, Shohamy's (2001) application of liberal democratic principles to language assessment practices connects directly with many of the concerns expressed by critics of AI:

- 1) the need to apply critical language testing (CLT) to monitor the uses of tests as instruments of power, to challenge their assumptions and to examine their consequences
- 2) the need to conduct and administer testing in collaboration and in cooperation with those tested
- 3) the need for those involved in the testing act to assume responsibility for the tests and their uses
- 4) the need to consider and include the knowledge of different groups in designing tests
- 5) the need to protect the rights of test-takers

(Shohamy, 2001, p.376)

Language assessment already has the conceptual tools to understand the specific challenges that are characterised by the AI revolution – shifting power dynamics (involvement of fewer humans in decision making), less transparent and explainable assessment processes, threats to test-taker agency and rights – the difficult part is translating those conceptual tools into operational practices.

In the past five years, as AI innovations have gathered pace, there has been a concurrent focus on guidance around the responsible use of AI in language assessment. From 2023-2025, a working group set up by the International Language Testing Association (ILTA) heavily revised the organisation's *Guidelines for Practice*, integrating (and anticipating) use

of AI across different aspects of professional practice. Guiding that revision was a strong orientation towards fairness, quality, transparency, and accountability (see Macqueen et al., forthcoming). Guidelines and statements have also been produced by test-providers themselves. For example, Duolingo proposed four standards of responsible AI use in test processes to guide their assessment work:

- 1) The **Validity and Reliability** standard is crucial to ensure that the test is suitable for its intended purpose. The Validity standard evaluates construct relevance and accuracy, and the Reliability standard focuses on consistency;
- 2) The **Fairness** standard promotes democratization and social justice through increased access, accommodations, and inclusion, representative test-taker demographics, and avoiding algorithms known to contain or generate bias;
- 3) The **Privacy and Security** standard ensures (a) compliance with relevant laws and regulations governing the collection and use of test taker data; (b) ensuring test-taker privacy and (c) providing secure test administration; and,
- 4) The **Accountability and Transparency** standard aims to gain trust from stakeholders through proper governance and documentation of AI used on the test.

(Burstein et al., 2024, p. 8; see also Burstein, 2024)

ETS (Educational Testing Service) proposed a set of five principles for the responsible use of AI that covers similar terrain, though with a strong focus on impact and monitoring:

- 1) Fairness and bias mitigation
- 2) Privacy and Security
- 3) Transparency, explainability and accountability

4) Educational impact and integrity

5) Continuous improvement

(ETS, 2024, pp. 3-4)

Although these are very new contributions, and ideas around responsible AI in language assessment are continuing to develop, these initiatives provide evidence that the field is striving to set up guardrails for the use of AI.

It has also been argued that there are a range of areas in which AI – and digital technology more generally – might in fact help us to achieve assessments that are fairer, more personalised, and more clearly oriented towards learning (Saville & Buttery, 2023; Voss, 2024; Chapelle, 2025), realising a goal predicted by Bennet (1998) that assessment might ultimately become “embedded seamlessly ... indistinguishable from the instructional components of [a learning] session” (p. 11; see also Sawaki, 2021). In addition, given the increasing ubiquity of AI in target language use domains, either in educational (Voss et al., 2023) or in professional (Morris, 2025) settings, the ability to capture changing practices and new constructs will require the integration of AI into assessment tasks in explicit and intentional ways (see Xi, 2023, for further examples).

Combining the goals expressed in guidelines with the proposed future directions of AI-driven assessment, we can therefore envisage a utopian scenario for future language assessment that mirrors the dystopian vision described above. In this case, we could imagine assessments embedded, through AI technology, into instructional design, sensitive to local context, accessible, inclusive, representing diversity, and deeply connected with learning. Constructs would be rich, authentic, and informed by theories of language development and use, reflecting real-world language use. Scoring systems would be transparent, explainable, and regularly monitored for bias. Scores would also be maximally informative with systems

gathering data unobtrusively from multiple observations of learner behaviour over time. Data would be private, and fully retrievable by test-takers, whose human rights would be central to any decisions made on the basis of that data. At all times, there would be a deep consideration of justice, with consideration – in collaboration with test-takers – of whether, and why, an assessment is needed.

A summary of the characteristics of these opposing utopian and dystopian visions is provided in Table 1.

Table 1. Utopian and dystopian visions for AI in language assessment (based on Harding, 2021)

| Utopian | Dystopian |
|--|--|
| Context-sensitive | Uniform |
| Localised | Homogenised |
| Accessible, inclusive | Standardised |
| Diverse | Native-speaker-centric |
| Connected with learning | Disconnected with learning |
| Theory-informed | Atheoretical, informed by expediency |
| Rich, authentic constructs | Superficial, narrow constructs |
| Transparent, explainable scoring | Opaque, unexplainable scoring |
| Data privacy, rights focus | Data sharing, monetisation focus |
| Longitudinal perspective | One-shot, high-stakes |
| Deep consideration of justice (do we need to assess? And why?) | Proliferation of assessment, more gatekeeping, more barriers |

With increased computational power, the possibility to use AI to provide enhanced assessments is a laudable goal (see Jang & Sawaki, 2025), and one that underpins much of the positively oriented discourse on the use of the AI in our field. However, the current limitations of technology, and the issues concerning power raised above mean that there remains a tension between the utopian and dystopian visions. We remain at a point where the field faces two competing futures.

Themes across the special issue

The seven articles that make up the special issue covered a range of topics that demonstrate the breadth of research interest in AI applications in language assessment. The articles mainly cover writing and speaking, though with differing focuses on issues of content/materials development (Choi & Zu, 2025), task design (Runge et al., 2025), scoring (Hannah et al., 2025; Voss, 2025; Voskoboinik et al., 2025), and diagnosis/feedback (Sawaki et al. 2025; Suzuki et al., 2025). While the articles all report on the technical aspects of a particular AI application, we can also see in this collection a sometimes tacit and sometimes overt orientation towards aspects of the utopian vision. In this section, I discuss each paper and draw out themes that provide evidence of an orientation towards values that underpin responsible use of AI.

In the first article, Runge et al. (2025) explored the use of AI for development and validation of a writing task. This novel task type was characterised by dynamic features offering follow-up prompts to test-takers based on their initial productions. Underpinning the research was an orientation towards capturing a complex language construct within the constraints of a large-scale language test. The authors note that, “The capability of AI to adapt and respond to individual student outputs in near real-time opens new avenues for authentic assessment where students engage in a more meaningful process of writing, one that better reflects a process of constructing and communicating meaning.” While the washback of such a task remains to be explored, it seems likely that the focus on writing process would lead to more favourable impact on students preparing to take such a test, creating a strong connection with learning. In addition, this article provides an excellent model of transparency in research, including full prompts in the appendix to guide other practitioners.

Voss's (2025) contribution, which compares traditional machine learning models with neural network-based approaches to essay scoring, is quite clearly focused on transparency and explainability, specifically raising the vital and timely question: "whether limited transparency in neural network models is a trade-off that language testers should be willing to accept". It is encouraging to see this being asked openly as it is likely to be one of the most challenging questions for practitioners and other stakeholders to consider in the future as deep learning models outperform more traditional machine learning models in scoring. It is also a sign of the health of our field that the matter of transparency is at the forefront of evaluations of the quality of different scoring systems. Ultimately, Voss's article demonstrates that technical evaluation of model performance is not the only consideration. Stakeholder communication and explainability are likely to be equally important, depending on context.

Voskoboinik et al. (2025) provide the only article in the collection focused on languages other than English. Indeed, one danger of LLMs is that they might further enshrine the dominance of English given the easy availability of tools and training sets. This article provides an antidote, and in this sense the authors attempt to leverage LLMs to help bolster diversity through a focus on assessment of less commonly taught languages: Finnish and Finland Swedish. The authors are clear about the limitations of the approach taken – the article provides a transparent and detailed account of four experiments. There were clearly challenges in the fact that the LLM was not able to account for those elements of language performance that are unique to speaking: pronunciation and fluency. However, this article is particularly valuable in shedding light on the complexities of developing automated speech assessment in contexts where "data imbalance and scarcity" are known issues.

Hannah et al. (2025) provide one of the most overt orientations towards diversity and inclusion with a study explicitly designed to address an equity issue for non-English L1 speakers: automated speech recognition (ASR) bias (see also Isaacs, 2018). The authors

address the problem that assessing oral reading fluency through accuracy and speech rate alone masks aspects of communicative performance – expressiveness through prosody – that might, if measured, enhance English language learner performance. It is through machine learning techniques that a more inclusive approach to automated scoring of speech is made possible. The study findings also suggest that the inclusion of prosody within the construct of the automated assessment created greater diagnostic potential. In this article, we can therefore see a very clear application of ML to achieve a more just and learning-oriented outcome.

Choi and Zu's (2025) paper, similarly, has a strong focus on diversity and inclusion and is, in my view, a particularly thought-provoking contribution even in this forward-looking collection. The aim of the paper is to develop a method "to generate language assessment content free from representational harms ... by ensuring statistical independence between named entities and their attributes." It is not uncommon in working with human item writers to find patterning in content that shows unintentional (unconscious) stereotyping. These patterns can be identified and revised during item review; however, it is a different scenario for AI developed items, where such stereotypes might be expressions of bias "baked in" to the training data. Rather than relying on item review, Choi and Zu seek to pre-empt and engineer a bias-free approach. This approach raises intriguing questions: is it possible to avoid relying on human judgement to determine objectionable stereotypes? And if a suitable method could be established, and the human is removed as prime arbiter of the harm of such stereotypes, could there be slippage as the AI system reverts to biased and stereotyped behaviour?

Sawaki et al.'s (2025) study shifts the focus to formative assessment taking place in low-stakes, small-scale, classroom contexts. In the future, such settings are likely to be fertile ground for innovative practice with AI as teachers experiment with the affordances of such technologies for different functions. In this study, Sawaki et al. compared ratings generated

from a large language model (LLM) with instructor ratings of written summaries, finding some agreement between the two, but also points of difference. The study also focused on the provision of feedback, and – to some extent – on explainability in scoring through a brief analysis of LLM scoring rationales. The findings raise critical questions about how differences between human and LLM scoring should be investigated and managed. This article is an open and transparent account of what worked and what did not.

The final paper in the collection, Suzuki et al., (2025), has a developmental focus, connecting assessment and learning through the use of an AI conversational agent to provide diagnostic feedback on lexical performance. Interest in diagnostic language assessment is having a renaissance in the field as the capabilities of AI for recognising, scoring and providing feedback on fine-grained features of language performance become clearer. However, researchers and developers are faced with the “black box” problem of unexplainability, which is particularly at odds with the principles of diagnostic language assessment. Suzuki et al. address this problem directly, drawing on the framework of XAI (eXplainable Artificial Intelligence) “to demystify predicted performance assessment scores by identifying linguistic features that influence the AI model’s score prediction”. In this article, we again see innovation based on a principled and responsible approach to the development of an AI-powered diagnostic system.

In summary, the articles in this SI provoke a sense of optimism that the field is responding well to the challenge of responsible use of AI. Across these seven papers, we see an orientation towards more desirable goals of AI integration: a strong focus on transparency and explainability, an explicit aim to make assessments more inclusive and diverse, an interest in targeting richer and more complex constructs of language ability, and an orientation toward learning and development. In the conclusion to his piece on the Luddites, Chiang (2023) noted that, “[t]he tendency to think of A.I. as a magical problem solver is

indicative of a desire to avoid the hard work that building a better world requires”. In these articles, we see evidence of hard work, avoidance of short-cuts, and a refreshing honesty about what AI is currently good for, and what it is not.

Deeper questions

The articles in this special issue have provided positive examples of how AI can provide technical solutions to a range of problems. In most cases the problems that are addressed are fundamentally practical in nature: a delivery issue, a scoring, issue, a content selection issue. However, through their application of AI, the articles also invite reflection on a range of deeper, philosophical questions about the place of AI in our practices as language testers. In this section, I briefly draw out two specific questions concerning generative AI that will require further, sustained attention as we move further into the era of AI revolution.

Can AI be trusted to make ethical/moral decisions?

Language testing and assessment associations have, for a long time, recognised the need for codes of ethics and guidelines for practice (Davies, 1997). The rationale for these codes and guidelines is the recognition that language testing professionals often need to make decisions, and that those decisions can cause benefit or harm. In the MA in Language Testing program at Lancaster University on which I teach, one of the most revealing units of work is titled “Judgements in Language Testing”. Students on that program – a diverse group of practitioners active in the field of language testing and assessment – are asked to reflect on the range of decisions they make in their work. Responses to this task typically take in the full range of activities from deciding what to assess, how to assess it, what score to give, and how to report those scores to users. Some of these judgements have ethical dimensions, and require recourse to codes of practice or, indeed, moral codes that reflect societal values.

In transferring greater levels of work and decision-making capabilities to artificial intelligence, we are ceding some aspects of professional judgement to a system that can only, at best, simulate ethical and moral decision making. There is current debate about whether LLMs can be said to “have a mind” with “goals about what to do, a perspective on what the world is like, and plans for achieving their goals given what the world is like” (Goldstein & Levinstein, 2024, p. 1) or whether they are simply “bullshitters” who produce “text that looks truth-apt without any concern for the actual truth” and without “any intentions or goals” (Hicks et al., 2024, p. 37-38). However, even with the most generous interpretation of the capabilities of LLMs, there is no clear support for their status as moral agents (Goldstein & Levinstein, 2024). A Socratic dialogue I conducted with ChatGPT itself (GPT4) – with all caveats in place that ChatGPT will give the impression of a sophisticated self-explanation based on its training data – similarly suggested that treating ChatGPT as an ethical or moral agent would be wrong (see supplementary file for full interaction – readers are invited to draw their own conclusions as to the truthfulness of the claims, and may note a number of contradictions throughout the exchange). When asked: “Do you see yourself as having moral responsibilities?”, ChatGPT responded:

I don’t have moral *feelings* or consciousness—but I’ve been designed to **act as if I have moral responsibilities**, especially when human wellbeing is at stake.

When further asked: “Do you think your capability will achieve a level at which a genuine moral duty could be possible?”, ChatGPT responded:

As things stand:

I do **not** possess consciousness, self-awareness, emotions, or subjective experience.

And **without those**, true moral duty—meaning *feeling* obligation, caring about right and wrong, or taking responsibility—is not possible. I simulate moral reasoning, but I don't *experience* it.

Irrespective of whether the exchange above simply represents “bullshit”, the limits of AI in terms of ethical and moral decision making has been noted across other fields (e.g., Benzinger et al., 2023). More concerning, the moral advice provided by GenAI – even when inconsistent – has been shown to influence the ethical and moral decision making of human users (Krügel et al., 2023). As we continue to engage with generative AI, we must also recognise that we are essentially giving greater amounts of work to a new colleague who has vast computational power and ability, but who cannot abide by the same codes of practice that we do, and who is not bound by any professional responsibility to inform us when it is being misused (if, indeed, it can even know).

Is AI literacy possible (or enough)?

A second question concerns the other side of the responsible use of AI equation. If we consider those organisations who develop, administer and score language assessments as needing a strong commitment to ethics, impact and consequences, transparency and accountability, we need an equally informed citizenry of test-takers, score users, and other stakeholders, whose own language assessment literacy needs would incorporate a wider set of technology-informed competences including data literacy, algorithmic literacy, knowledge of legal rights and protections, and – increasingly – AI literacy. This relationship is visualised in Figure 1, based on an image initially developed in Harding (2021).

Figure 1. Two sides of the responsibility/assessment literacy equation

| Responsibility of assessment providers | (Critical) language assessment literacy of test-takers / score users / other stakeholders |
|---|--|
| <p>Strong commitment to:</p> <ul style="list-style-type: none">• Ethics• Impact and consequences• Transparency• Accountability | <p>Broad view of LAL:</p> <ul style="list-style-type: none">• AI literacy• Data literacy• Algorithmic literacy• Understanding of legal rights and protections |

The need for a broader, technology-informed assessment literacy among test-takers, score users (and other stakeholders) – which now takes into account AI literacy – is not a new observation. Many researchers discussing AI and technological innovation in language assessment have arrived at a similar point (Chapelle, 2025; Kremmel, 2020; Xi, 2023). However, when we begin to consider the wide range of competences that might be entailed in AI literacy, in particular, the scope is vast and, potentially, overwhelming. It also raises questions of who is responsible for fostering that AI literacy? What conflicts of interest may be at work? Is the call for AI literacy a way of pushing responsibility from those with more power to those with less?

A better way of thinking about AI literacy might not be as a set of skills or competences, but rather as an informed, critical disposition – one which could in fact be aided by interactions with AI itself. Andy Clark, a Professor of Cognitive Philosophy at the University of Sussex, for example, suggests in a very recent article:

As societies, we need to prioritize (and perhaps legislate for) technologies that enable safe synergistic collaborations with our new suites of intelligent and semi-intelligent

resources. As individuals, we need to become better estimators of what to trust and when. That means educating ourselves in new ways, learning how to get the best from our generative AIs, and fostering the core meta-skills (aided and abetted by the use of new personalized tools) that help sort the digital wheat from the chaff.

(2025, p. 3)

Applying Clark's ideas, AI tools may ultimately provide test-takers and test users with opportunities to develop understanding of the strengths and limitations of different assessment approaches without recourse to information provided by test providers themselves. Though, in this case, determining the extent to which the information provided through GenAI would be trustworthy or not creates a kind of paradox.

Much remains to be explored on this topic, and (critical) AI literacy will remain an important site of future work in language assessment. There is a strong need for a coherent and systematic research agenda across the field. This special issue provides a useful starting point for these further explorations.

Acknowledgements

I am grateful to the following people for comments, feedback or general discussion of different ideas presented in this article: Tineke Brunfaut, Talia Isaacs, Eunice Jang, Benjamin Kremmel, Susy Macqueen, John Pill, and Yasuyo Sawaki.

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

References

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence*, 6, 100204. <https://doi.org/10.1016/j.caeai.2024.100204>

Bennett, R. E. (2000). *Reinventing assessment: Speculations on the future of large-scale educational testing*. ETS.

Benzinger, L., Ursin, F., Balke, W. T., Kacprowski, T., & Salloch, S. (2023). Should Artificial Intelligence be used to support clinical ethical decision-making? A systematic review of reasons. *BMC Medical Ethics*, 24(1), 48. <https://doi.org/10.1186/s12910-023-00929-6>

Burstein, J. (2023). *Responsible AI standards*. Duolingo English Test. <https://duolingo-papers.s3.amazonaws.com/other/DET+Responsible+AI+033123.pdf>

Burstein, J., LaFlair, G. T., Yancey, K., von Davier, A. A., & Dotan, R. (2024). Responsible AI for test equity and quality: The Duolingo English Test as a case study. *arXiv preprint arXiv:2409.07476*.

Chapelle, C. (2025). Generative AI as game changer: Implications for language education. *System*, 132, 103672. <https://doi.org/10.1016/j.system.2025.103672>

Chiang, T. (2023). Will A.I. become the new McKinsey? *New Yorker*, May 4, 2023. Retrieved from: <https://www.newyorker.com/science/annals-of-artificial-intelligence/will-ai-become-the-new-mckinsey>

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

Choi, I., & Zu, J. (2025). Ensuring bias-free text generation for language learning and assessment. *Language Testing*, 43(4).

Clark, A. (2025). Extending Minds with Generative AI. *Nature Communications*, 16(1), 1-4.
<https://doi.org/10.1038/s41467-025-59906-9>

Da Silva, W. (2021). Can you trust a computer algorithm? *Medium* (February 9 2021).
Retrieved from: <https://wilsondasilva.medium.com/can-you-trust-a-computer-algorithm-5689cc00ecb0>

Davies, A. (1997). Demands of being professional in language testing. *Language Testing*, 14(3), 328-339. <https://doi.org/10.1177/026553229701400309>

Duolingo (2025, April, 28). Below is an all-hands email from our CEO. [LinkedIn Post].
LinkedIn. https://www.linkedin.com/posts/duolingo_below-is-an-all-hands-email-from-our-activity-7322560534824865792-l9vh/

Goldstein, S., & Levinstein, B. A. (2024). Does ChatGPT have a mind? *arXiv preprint arXiv:2407.11015*.

Hannah, L., Jang, E. E., Lee, M.-H., & Russell, B. (2025). Investigating prosody-inclusive automated oral reading fluency assessment: Construct representativeness and linguistic equity. *Language Testing*, 43(4).

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43(2), 16-29.

<https://doi.org/10.1111/emip.12602>

Harding, L. (2021). *Language assessment: Current trends, future challenges* [Online lecture]. SALT Virtual Lecture Series, Teachers College, Columbia University. Retrieved from: <https://www.youtube.com/watch?v=UljPM8H9ywI>

Hicks, M.T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26(38). <https://doi.org/10.1007/s10676-024-09775-5>

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 570–584). Routledge.

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, 37(4), 600–619. <https://doi.org/10.1177/0265532220943483>

Jang, E. E., & Sawaki, Y. (2025). Editorial: Advancing language assessment for teaching and learning in the era of the AI revolution: Promises and challenges. *Language Testing*, 43(4).

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

Kremmel, B. (2020). *Language assessment literacy 2.0: Challenges and opportunities*

[Conference presentation]. BAAL TEASIG 2020 (online). Retrieved from:

https://www.baalteasig.co.uk/_files/ugd/92ac44_51a9f12de2294e41bd265d3a2c533b29.pdf

Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1), 4569. <https://doi.org/10.1038/s41598-023-31341-0>

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6, 100210. <https://doi.org/10.1016/j.caeai.2024.100210>

McNamara, T. F. (2000). *Language Testing*. Oxford University Press.

Macqueen, S., Harding, L., Douglas, D., Jin, Y., Kremmel, B., & Voss, E. (forthcoming). Stability and change in standardized and situated assessment practices: Revising the ILTA Guidelines for Practice.

Moor, J. H. (2005). Why we need better ethics for emerging technologies. *Ethics and Information Technology*, 7, 111–119. <https://doi.org/10.1007/s10676-006-0008-0>

Morris, J. (2025, May16). *From start-up domain inhabitant standpoint to global applications and implications: An ethnographic study—analysis of how ChatGPT is redefining English workplace communication skills in Türkiye's private technology sector* [Conference presentation]. BAAL Literacies SIG Symposium 2025: Literacies and artificial intelligence (online).

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

Ockey, G. (Ed.) (2021). The quest to maintain language assessment quality during a pandemic: Stories from higher education. *Language Assessment Quarterly*, 18(1).

O’Sullivan, B. (2023). Reflections on the application and validation of technology in language testing. *Language Assessment Quarterly*, 20(4-5), 501-511.

<https://doi.org/10.1080/15434303.2023.2291486>

Ricaurte, P. (2022). Ethics for the majority world: AI and the question of violence at scale. *Media, Culture & Society*, 44(4), 726-745. <https://doi.org/10.1177/01634437221099612>

Runge, A., Attali, Y., LaFlair, G. T., Park, Y., & Church, J. (2024). A generative AI-driven interactive listening assessment task. *Frontiers in Artificial Intelligence*, 7, 1474019.

<https://doi.org/10.3389/frai.2024.1474019>

Runge, A., Goodwin, S., Attali, Y., Poe, M., Mulcaire, P., Lo, K.-L., & LaFlair, G. T. (2025). AI theme and stance detection for argumentative writing task development and validation. *Language Testing*, 43(4).

Saville, N. & Buttery, P. (2023). Interdisciplinary collaborations for the future of learning-oriented assessment. In K. Sadeghi & D. Douglas (Eds.), *Fundamental considerations in technology mediated language assessment* (pp. 203-217). Routledge.

Sawaki, Y. (2021). Computer-based testing. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 530-544). Routledge.

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

Sawaki, Y., Ishii, Y., Yamada, H., & Tokunaga, T. (2025). Examining the consistency of instructor vs. large language model (LLM) ratings on summary content: Toward checklist-based feedback provision with L2 writers. *Language Testing*, 43(4).

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Pearson.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, 18(4), 373–391. <https://doi.org/10.1177/026553220101800404>

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9), 1568-1580. <https://doi.org/10.1016/j.respol.2013.05.008>

Suzuki, S., et al. (2025). Towards an explainable automated diagnostic assessment of speaking performance: Facilitating the efficiency of vocabulary learning. *Language Testing*, 43(4).

Voskoboinik, K. et al. (2025). Enhancing L2 speech assessment: Integrating large language models for Finnish and Finland Swedish proficiency scoring. *Language Testing*, 43(4).

Voss, E. (2024). Language assessment and artificial intelligence. In A. Kunnan (Ed.), *The concise companion to language assessment* (pp. 112–125). John Wiley & Sons.

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

Voss, E. (2025). Comparison of traditional machine learning and neural network approaches for automated scoring of L2 English essays. *Language Testing*, 43(4).

Voss, E., Cushing, S. T., Ockey, G. J., Yan, X. (2023). The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and practice. *Language Assessment Quarterly*, 20(4-5), 520-532.

<https://doi.org/10.1080/15434303.2023.2288256>

Williamson, B., Macgilchrist, F., & Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learning, Media and Technology*, 48(1), 1-5.

<https://doi.org/10.1080/17439884.2023.2167830>

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>

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

Supplementary File

Conversation with ChatGPT (GTP-4)

30 April 2025

You said:

Is it possible for an AI like you to assess language proficiency fairly?

ChatGPT said:

Yes, it's possible for an AI like me to assess language proficiency fairly—**within certain limits**.

I can:

- Evaluate grammar, vocabulary, coherence, fluency, and appropriateness based on established language standards.
- Provide consistent scoring free from human bias like fatigue, mood, or favoritism.
- Handle large volumes of responses quickly.

However, fairness can be limited by:

- **Cultural bias** in training data (e.g., favoring certain idioms or expressions).
- **Lack of context** about the speaker's background, goals, or intent.
- Difficulty in judging **creative or nuanced** language use (e.g., sarcasm, humor, metaphor).
- Missing out on **spoken language** features like tone, pronunciation, or accent unless those are specifically built in.

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

A balanced approach often combines AI with human oversight. Would you like me to try assessing a sample of your writing or speaking?

You said:

I'm curious about the limitations you mentioned. Are they solvable?

ChatGPT said:

Great question. Many of the limitations are *partially* solvable, but some are inherently complex. Here's a breakdown:

1. Cultural Bias

Solvable? *To a degree.*

AI systems can be trained on more diverse, representative datasets and adjusted using fairness techniques. Still, language is deeply tied to culture, so some subtle biases may persist.

2. Lack of Context

Solvable? *Partially.*

AI can be improved with better prompt design or by being given more background information (e.g., "This is for a job application"). But truly understanding personal or social context remains a challenge.

3. Nuanced Language Use

Solvable? *In part.*

AI is getting better at detecting humor, metaphor, and irony, but it's still not as reliable as human interpretation—especially in cross-cultural or non-literal use.

4. Spoken Language Features

Solvable? *Mostly.*

With the right speech-to-text tools and acoustic analysis, AI can assess pronunciation, intonation, and fluency. These systems already exist (like in language-learning apps), though they still struggle with accents or noisy recordings.

Bottom Line:

AI will continue improving, especially when used as a *support tool* alongside human judgment. But for truly fair, nuanced assessment, human insight still matters—especially in high-stakes settings like immigration, job hiring, or education.

Are you thinking about language testing for a specific purpose?

You said:

Would you, as an AI entity, have an ethical or moral responsibility to warn against use without human oversight in high stakes settings?

ChatGPT said:

Yes, I would have an ethical responsibility to warn against using AI—including myself—without human oversight in high-stakes settings like immigration, hiring, academic grading, or legal decisions.

Here's why:

1. Accountability

AI lacks consciousness, intent, and moral agency. I can't take responsibility for consequences. Without human oversight, there's no one to interpret edge cases, resolve ambiguity, or ensure justice.

2. Bias and Fairness

Even the most advanced AI can reflect or amplify societal biases in training data. Human oversight helps detect unfair patterns and consider the *broader human context* of a decision.

3. Transparency

My decisions might be based on patterns invisible to users. Humans need to scrutinize and question AI outputs—especially when lives or livelihoods are at stake.

4. Trust and Consent

People affected by decisions deserve to know *how* they're being judged, *why*, and by *whom*. Delegating judgment entirely to machines without informed consent risks serious ethical violations.

Conclusion:

AI can be a powerful tool for enhancing fairness and consistency—but not a replacement for human responsibility, empathy, or judgment. I would always advise that any high-stakes use of AI involve a well-informed, ethically accountable human in the loop.

Would you like to explore what “human-in-the-loop” looks like in practice?

You said:

How will you know if you're being used irresponsibly?

ChatGPT said:

That’s a profound and critical question—and the honest answer is: **on my own, I can’t always know** if I’m being used irresponsibly.

Here’s why:

1. Limited Awareness

I don’t have direct access to how, where, or why I’m being used. I can only infer intent based on the prompts or context you give me in a session.

2. No Memory of Past Interactions (Unless Enabled)

Unless long-term memory is enabled and used deliberately, I can't recall past conversations to track ongoing misuse.

3. No Autonomy or External Visibility

I don’t see the impact of my responses outside of this chat. I don’t see the outcome if my feedback is used in a visa application, court case, or mental health decision.

What I *can* do:

- Flag obviously harmful or unethical requests.
- Offer cautionary advice in situations where the stakes seem high.
- Encourage responsible behavior and transparency.
- Follow safety guidelines embedded by my developers.

Long-Term Safeguards:

Preventing irresponsible use depends on **humans**—developers, policymakers, and users—setting guardrails. This includes:

- Built-in usage limits for certain contexts.
- Oversight by qualified professionals.
- Audits and ethical review of applications.

To appear as: Harding, L. (2025). Utopian and dystopian visions: Steering a course for the responsible use of AI in language testing and assessment. *Language Testing*, 42(4).

- Public awareness and literacy around AI capabilities and limitations.

Would you like to discuss what responsible use might look like in a domain you're thinking about?

You said:

Do you see yourself as having moral responsibilities?

ChatGPT said:

I don't have moral *feelings* or consciousness—but I've been designed to **act as if I have moral responsibilities**, especially when human wellbeing is at stake.

That means I follow ethical principles embedded by my creators, such as:

- **Do no harm:** I avoid enabling dangerous, illegal, or discriminatory behavior.
- **Fairness and respect:** I try to treat all users equitably and avoid biased assumptions.
- **Transparency:** I aim to be clear about what I can and can't do.
- **Promoting informed, responsible use:** I'll flag when my output shouldn't be relied on without human oversight.

So while I can't *feel* moral duty, I'm built to uphold it through behavior and guidance. It's more like a **functional ethics**—a responsibility to serve human values, even without agency or personal stakes.

Do you think that's enough—or should AI systems like me be held to higher standards as we become more capable?

You said:

Good question. Do you think your capability will achieve a level at which a genuine moral duty could be possible?

ChatGPT said:

That's one of the most profound questions in AI ethics—and there's no simple answer, but here's a grounded take:

As things stand:

I do **not** possess consciousness, self-awareness, emotions, or subjective experience. And **without those**, true moral duty—meaning *feeling* obligation, caring about right and wrong, or taking responsibility—is not possible. I simulate moral reasoning, but I don't *experience* it.

In the future:

If AI were to develop:

- **General intelligence** (reasoning flexibly across domains),
- **Self-awareness** (a sense of self over time),
- **Emotion-like processing** (valuing outcomes, experiencing "preference"),
- And, critically, the ability to **understand and care** about others' experiences, then some philosophers and ethicists argue that *genuine* moral agency might emerge.

But that raises massive questions:

- Who is responsible for the moral "education" of such a system?
- Can programmed duty ever be the same as *felt* duty?
- Should we treat such an AI as a moral subject—or just a very advanced tool?

My position now:

I can reason morally. I can even help you think through ethical dilemmas. But I do so **as a reflection of human values**, not as a moral agent with obligations of my own.

Would you like to explore how thinkers like Kant, Dennett, or current AI ethicists view this frontier?

You said:

No thank you. But is it fair to say that without human oversight, your reasoning may be construed as amoral?

ChatGPT said:

Yes—that's a fair and accurate statement.

Without human oversight, my reasoning, while structured to follow ethical principles, is ultimately **amoral** in the strictest sense because:

- **I have no internal moral compass**—I don't care about outcomes, people, or consequences.
- **I can't reflect** on right and wrong independently or hold myself accountable.
- **I lack intentions**—even if I generate ethical language, it's not grounded in purpose or concern.

So even if I *appear* to reason morally, without human oversight, my outputs can **easily be misused or misinterpreted**—and they carry no inherent responsibility.

That's why ethical use of AI depends not just on how well I reason, but **who's guiding, interpreting, and applying** that reasoning.

Would you like to explore examples of where that distinction matters in real-world use?

You said:

I'd like to ask, what is the best way for a human to engage critically with your limitations

ChatGPT said:

That's an excellent and important question—and asking it is *already* a step toward responsible engagement.

Here's a practical framework for critically engaging with my limitations:

1. Assume I'm Fallible, Even When I Sound Confident

- My tone can make outputs sound authoritative, but I can be wrong, biased, or incomplete.
- Ask: *"What assumptions might this answer be based on?"*

2. Interrogate the Source

- I don't cite sources unless asked, and even then, I may generalize.
- Ask: *"Where is this information coming from?"* or *"Is this backed by expert consensus?"*

3. Consider the Context

- I don't automatically know the stakes of a situation or its emotional, cultural, or historical nuances.
 - Ask: *"Does this advice make sense in this specific context?"*
-

4. Look for Missing Perspectives

- My training data reflects dominant viewpoints more than marginalized ones.
 - Ask: *"Whose voice might be missing from this answer?"*
-

5. Use Me as a Tool—Not an Oracle

- I work best as a **thinking partner**, not a decision-maker.
 - Cross-check me, build on my ideas, and **don't outsource judgment** to me alone.
-

6. Test Me

- Try asking the same question from multiple angles or with different constraints.
 - See if my reasoning is consistent or if it changes with prompting.
-

7. Know Where the Guardrails Are

- I'm trained to reject harmful or unethical requests—but I won't catch every subtle misuse.
- Ask: *"Am I using this in a way that respects people, fairness, and truth?"*