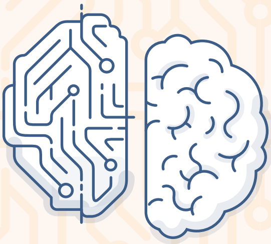
duolingo english test

Responsible AI Standards



Author

Jill Burstein, Principal Assessment Scientist, Duolingo English Test

Contributors

Alina von Davier, Chief of Assessment, Duolingo English Test

Kevin Yancey, Staff AI Research Engineer, Duolingo English Test

Will Belzak, Senior Psychometrician, Duolingo English Test

Klinton Bicknell, Senior AI Research Manager, Duolingo

Carl Gottlieb, Privacy Lead & Data Protection Officer, Duolingo

Rose Hastings, Head of Proctoring Operations, Duolingo English Test

Ian Riggins, Operations Manager, Duolingo English Test

Mark Zheng, Senior Corporate Counsel, Duolingo

External Reviewer

Pascale Fung, Chair Professor, Department of Electronic and Computer Engineering, Hong Kong University of Science & Technology

Internal Reviewers

Ramsey Cardwell, Assessment Scientist, Duolingo English Test

Sophie Wodzak, Senior Communications Manager, Duolingo English Test

Invitation for public comment

We invite public comment on the Duolingo English Test Responsible AI Standards. We developed the standards to advance thinking in the field of assessment with regard to the ethical use of AI for testing. As such, our standards were informed by the AERA/APA/NCME standards, ITC-ATP guidelines for technology-based assessment, and scholarly literature in AI ethics. Leveraging industry guidelines and AI ethics, and engaging in multi-stakeholder collaboration has helped us to formulate our responsible AI standards. Our standards contribute to the Duolingo English Test's validity, reliability, fairness, and security. The Duolingo English Test assessment research team — composed of experts in applied linguistics, computational psychometrics, language assessment, machine learning, and statistics — developed the standards in collaboration with experts from Duolingo's legal and security teams, and an independent external responsible AI expert from the computer science discipline. The Duolingo English Test Responsible AI Standards are intended to be a living document. We believe that through public engagement with stakeholders across communities impacted by AI in testing, the standards will promote the goal of using AI for good.

Note: This version of the Duolingo English Test Responsible AI Standards was updated on March 29, 2024. Document revisions to the standards themselves reflect updates in Duolingo English Test Responsible AI practices, and responses to public feedback gathered since our standards initial release in April 2023. The Duolingo English Test Responsible AI Standards updates include: The Validity & Reliability Subgoal: 1.2.7; Fairness Subgoals 2.1.2, 2.1.3, 2.1.4, 2.1.6, & 2.2.4; and, 2.2.5; and, the Accountability & Transparency Subgoal 4.3.3. Revisions also include editorial modifications to enhance explanation and overall document quality.

Comments can be sent to englishtest-research@duolingo.com; please specify "**Responsible AI Standards**" in the email subject line.

How to Cite:

Burstein, J. (2023). The Duolingo English Test Responsible AI Standards.
[Updated March 29, 2024]:
https://go.duolingo.com/ResponsibleAI

Contents

Introduction

Artificial intelligence (AI[1]) is now instantiated in digital learning and assessment platforms. While there are many benefits of AI use for assessment, there are also risks. Therefore, assessment frameworks, guidelines and standards are needed that explicitly address responsible AI use. Classical assessment validity (Chapelle et al., 2008; Kane, 1992, 2013) and fairness (Kunnan, 2000) frameworks were developed for paper-and-pencil and first-generation, computer-based assessments. While those frameworks capture ethical principles (such as validity and reliability, and fairness), they do not directly address the use of technology for assessment. As such, those frameworks may be limited with regard to addressing responsible use of AI in modern, digital assessments. More recent assessment research discusses AI in terms of validity (such as Huggins-Manley et al., 2022, Burstein et al., 2022; Williamson et al., 2012, and Xi, 2010), while also leveraging the classical frameworks. AI in assessment is also addressed in the American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014) standards (henceforth, AERA/APA/NCME Standards). As well, ethical principles, including validity and reliability, and fairness, are embedded in the AERA/APA/NCME Standards, and are tied to the standards and their suggested practice. Ethical practices (such as for test-taker privacy, test security, and test documentation) discussed in the AERA/APA/NCME Standards serve to inform responsible AI standards for assessment. With regard to responsible AI, however, those standards largely address automated essay scoring. This highlights the limited scope of AI for assessment at the time the standards were published. To our knowledge, there are currently no comprehensive frameworks specifically addressing responsible AI for modern, digital assessment.

There has been a substantial proliferation of literature around responsible AI guidelines, regulations, and governance (e.g., The White House, 2023; Department for Science, Technology & Innovation, 2023; Council of the EU, 2023); frameworks (e.g., NIST, 2023); and, theory (e.g., Bentley et al., 2023; Gianni et al., 2022; Fjeld et al., 2020; and Jobin et al., 2019). Given the potential risk of AI use across different sectors, the computer science community has recommended systematic audits of AI applications to prevent potential harm (Mökander & Axente, 2023; Mökander & Floridi, 2021; Raji et al., 2020; Raji & Buolamwini, 2019). There is a smaller, but growing body of research focussed on responsible AI in education (such as Belzak et al., 2024; LaFlair et al., 2022; Johnson et al, 2022). In earlier work, Aiken and

[1] Note that our use of the term "AI" refers to AI systems and AI-adjacent methods and disciplines, such as computational psychometric methods (von Davier, et al., 2022; von Davier, 2017).

Epstein (2000) discuss ethical considerations for AI in education. Earlier, Dignum (2021) proposed a high-level vision for responsible AI for education. Dieterle et al (2022) and OECD (2023) discuss guidelines and issues associated with AI in testing.

The Duolingo English Test (DET) is a measure of English language proficiency for communication and use in English-medium settings (Cardwell et al, 2024). The test uses human-in-the-loop AI for test design, such as for automated generation of test item content; for measurement, such as for item parameter estimation and automated scoring of written and spoken responses; and for test security, such as to assist with decision-making in remote test proctoring (Belzak et al, 2024). The DET values human expertise as an essential part of decision-making. This is essential in the context of admissions decisions in higher education, since it has high-stakes, potentially life-changing implications for test takers. Human expertise supports accuracy and fairness of AI system outputs that may impact test security, test development, and measurement of test-taker proficiency. Further, AI affordances contribute to a positive test-taker experience. Examples include: automated scoring of practice tests offering test takers an immediate estimate of their test performance; and, computer adaptive testing, allowing for a shorter test-taking experience. The DET's use of AI is in support of Duolingo's mission — to develop high-quality education and provide universal access. The mission reflects the principles of AI for Good – specifically, the DET's use of AI to support a valid, reliable, fair, and secure test, while also promoting a positive test taker experience.

As part of professional responsibility, the DET developed Responsible AI (RAI) Standards. Our Standards embrace ethical principles, and inform responsible AI practice intended to mitigate risk and minimize ethical debt associated with the use of AI for assessment across the DET's assessment ecosystem – i..e, test design, measurement, and security. Ethical debt is the notion that AI tools developed without attention to responsible AI can cause harm (Feisler & Garrett, 2020). An example is facial recognition systems and fairness. In this case, Buolamwini & Gebru (2018) noted harms in that system accuracy varied based on skin color. The DET RAI Standards inform responsible practices that mitigate risk, but also minimize ethical debt pertinent to the test.

The DET Responsible AI Standards

Developed as part of professional responsibility, the DET RAI standards are aligned with four ethical principles: Validity and Reliability, Fairness, Privacy and Security, and Accountability and Transparency. Our RAI standards and their goals are designed to support: (a) auditing of AI-powered test processes across test design, measurement and security; (b) validity and reliability studies; and, (c) documentation for theoretical, qualitative, and quantitative research relevant to AI use on the DET, and responsible AI practices.

To select the ethical principles for the standards, the DET research team completed five key activities. First, to better understand the set of principles that were applicable to the DET, the team conducted a literature review of ethical principles used for AI (including, Floridi & Cowls, 2022; Fjeld et al., 2020; Jobin et al., 2019; Memarian & Doleck, 2023). Second, to examine alignment between domain-agnostic (e.g., NIST, 2023) and assessment-specific principles, the team reviewed assessment-specific standards (AERA/APA/NCME, 2014) and guidelines (including OECD 2023; U.S. Department of Education, Office of Educational Technology, 2023; and ITC-ATP, 2022). Third, the team engaged in multi-stakeholder collaboration. They worked with experts in applied linguistics, computational psychometrics, language assessment, law, machine learning, statistics,and security within Duolingo, and an independent, external RAI expert from computer science. The resulting principles are therefore the outcome of collaboration with a cross-disciplinary set of experts within and outside of the DET. Fourth, after establishing the four ethical principles, the team worked with the external RAI expert collaborator to articulate the rationale and overall goal of each standard, and the more detailed subgoals (i.e., practical implementation of each standard). Finally, the team published the standards as a living document that is freely available, and open for public comment. Our standards remain open for public comment to maintain multistakeholder collaboration as we continue to update the standards.

The remainder of this document presents the DET RAI Standards.

1. Validity & Reliability

Rationale: Validity and Reliability standards are crucial to ensure that the test is suitable for its intended purpose. Validity standards involve evaluating construct relevance and accuracy (Kane, 1992; Kane, 2013), while Reliability standards focus on test score consistency.

Goals summary: To specify processes required to build a validity argument, and to evaluate AI used in test item creation, item calibration, and scoring.

Goal 1.1. Specify processes required to build a validity argument.

Processes include theoretical and empirical evaluations that directly inform or address AI used to build a validity argument for test score use.

1. Develop a description for the test target domain – i.e., English language proficiency – to ensure that test items are aligned with the domain being measured.
2. Evaluate AI scoring system accuracy and fairness, leveraging human expertise. Examples include agreement with human raters, accuracy of system features used for scoring constructed responses, and evaluations of scoring bias.
3. Develop (a) explainable scoring methods, and (b) interpretable AI features used for scoring that have clear alignment with domain constructs.
4. Conduct empirical investigations of item reliability, ensuring reliability for AI-generated items.
5. Evaluate extrapolation through empirical investigations to illustrate relationships between automatically-generated items, test-taker scores, and relevant external measures that suggest proficiency in English skills. Examples of external measures include relationships to other tests, relationships between test-taker's linguistic input and the target domain.

Goal 1.2. Evaluate AI used in test item creation, item calibration, and scoring.

1. Identify AI methods for **item creation**, leveraging human expertise to efficiently create valid and reliable test items. An example of human expertise is human review of items from automated item generation.
2. Conduct human evaluations of the quality of items created using AI, such as reviewing outputs from automated item generation.
3. Identify AI methods that can be efficiently used for valid and reliable **test item calibration**.

4. Conduct evaluations that confirm the accuracy of AI for predicting item parameters (such as item difficulty), leveraging human expertise for quality assurance.

5. Identify AI methods that efficiently produce valid and reliable **scores** for test-taker responses.

6. Conduct evaluations to confirm the accuracy of AI for scoring test-taker responses, leveraging human expertise for quality assurance.

7. System changes and re-evaluations are conducted in the case of suboptimal evaluation outcomes for item creation, item calibration, and scoring.

2. Fairness

Rationale: Fairness standards are required to promote democratization and social justice through increased access, accommodations, and inclusion (ITC-ATP, 2022; Burstein et al., 2022; Cardwell et al., 2023; and Care & Maddox, 2021), represent test-taker demographics, and avoid algorithms known to contain or generate bias (Belzak, 2022; Johnson et al., 2022).

Goals Summary: To specify how the use of AI facilitates test-taker access, accessibility and inclusion; and to specify test-taker demographic representation, and algorithms known to contain or generate bias.

Goal 2.1. Specify how the use of AI facilitates test-taker access, accessibility, accommodations, inclusion, and appeals.

1. Identify AI methods to increase **test-taker access** globally, as part of the DET test-taker experience mission. For example, the DET is available remotely, online, and 24/7. Other access considerations include, but are not limited to, test costs, access to devices, and testing time.

2. Adopt design principles in compliance with **accessibility standards** for test takers that offer a generally accessible user interface and allow for test accommodations, due to factors such as low vision or physical limitations.

3. Ensure that AI or AI-adjacent capabilities do not impact design, such that accessibility compliance might be violated.

4. Ensure that test accommodations are not adversely affected by AI.

5. Develop and apply fairness and bias item review principles for **inclusion** that eliminate construct-irrelevant barriers, and ensure that cultural and linguistic factors do not impede accessibility and inclusion for the DET test-taker population.

6. Maintain an **appeals process**, allowing test takers to appeal test results where there is suspicion of cheating.

Goal 2.2. Specify test-taker and human expert demographic representation, algorithms known to contain or generate bias, and potential human bias.

1. Evaluate and document **demographic representation** in data sets used to build AI. Documentation should describe how representative (inclusive) the data are with regard to DET test takers. For example, the selection of data for AI system development, such as human rater scoring of written responses, should consider the underrepresentation of test-taker language groups which could lead to bias in test-taker outcomes.

2. Evaluate and document known algorithmic bias in AI used in DET ecosystem processes (i.e., test security, design, and measurement).

3. Evaluate and document **bias** associated with automatically-generated item content (e.g., fairness and bias review guidelines), and proficiency measurement.

4. Identify and document human expert demographic representation.

5. Evaluate and document bias associated with human expert ratings of test-taker production tasks.

3. Privacy & Security

Rationale: The Privacy and Security standards are needed to ensure that we (a) comply with relevant laws and regulations governing the collection and use of test taker data; (b) ensure test taker privacy and (c) to ensure secure test administration. (See Liao et al., 2022a; Liao et al., 2022b; LaFlair et al., 2022; Wodzak, 2021; and, Duolingo English Test: Security, Proctoring, and Accommodations, 2021).

Goals Summary: To specify methods to ensure privacy and security associated with data origin, data collection and processing, and data management; to specify how to maintain test-taker privacy, item security, and test-taker security during test administration; and to specify fair and reliable test security proctoring protocols, item pool development and psychometric procedures for test security.

Goal 3.1. Specify methods to ensure privacy and security associated with data origin, data collection and processing, and data management.

1. Ensure that **data provenance, governance,** and **management** comply with the Duolingo privacy policy, external privacy policies (where appropriate), and applicable laws such as the European Union's General Data Protection Regulation (GDPR). (See also the EC Proposed AI Regulation, US National Conference of State Legislatures, 2021).

2. Define and document **data requirements** with regard to DET intended uses, stakeholders, and the geographic areas where the DET is administered that do not violate privacy terms or security (such as including personally-identifying information without consent).

3. Define, document, and implement methods to ensure that **data provenance** complies with Duolingo privacy policy), and security policies with regard to the origin of the data (e.g., open-access corpora, test taker), how it was obtained (e.g., test-taker consent), and changes applied.

4. Define, document, and implement methods to ensure that **data governance** complies with Duolingo privacy policy and, where appropriate, external privacy or security policies during data collection and processing, including data cleaning, annotation, enrichment, and aggregation, sharing, and use. Document how stakeholder data is used, including but not limited to biometric and personally-identifying data (e.g., IDs for security), process (such as keystroke profiles), and product response data and test scores.

5. Define, document, and implement **data management** procedures to ensure compliance with Duolingo and, where appropriate, external privacy or security policies.

Goal 3.2. Specify how to maintain test-taker privacy, item security, and test-taker security during test administration.

1. Define, document, and implement methods for **test-taker verification** in the context of test onboarding to ensure that test-taker identity can be authenticated.

2. Define, document, and implement methods to ensure that verified **test-taker identity** (i.e., personally-identifiable information) is secure.

3. Define and document **test-taking rules**, such as prohibiting headphones, except in cases of accommodations, and mitigate actual or perceived cheating behaviors to support **test-taker integrity** (see Test Rules in FAQs; Test Security Rules).

4. Define, document, and implement **test administration** processes that mitigate cheating through use of external resources – i.e., test administration through a desktop application.

Goal 3.3. Specify fair and reliable test security proctoring protocols, item pool development, and psychometric procedures for test security.

1. Define, document, and implement human-in-the-loop AI **proctoring protocols** that fairly and reliably identify novel and known cheating behaviors.

2. Define and document the algorithm used for proctoring support, proctor training for use of AI, and bias management – i.e., how proctors identify and report perceived AI bias.

3. Define, document, and implement methods (such as, automated item generation) to support scaling of a large, and continuously refreshed test **item pool**. Methods include human expert monitoring protocols for tracking item exposure and test overlap. Larger item pools mitigate the risk that a single test taker, or multiple test takers are likely to see the same test item, or set of ordered test items during repeated sessions (i.e., a test taker registers for and takes the test multiple times).

4. Define, document, and implement security protocols to prevent item breach from external attackers.

5. Define, document, and implement **psychometric procedures**, such as test-retest reliability that can track anomalies in test-taker performance that can reveal test-taker cheating behavior.

4. Accountability & Transparency

Rationale: To gain trust from stakeholders, it is essential that the DET have Accountability & Transparency standards for proper governance of AI used on the test. Through documentation and explanations, we are holding ourselves accountable.

Goals Summary: To assess how AI processes impact stakeholders; to document AI used for building the validity argument, test item creation, test item calibration, and scoring; to document processes for human-in-the-loop interactions with AI; document human expert qualifications required for human-in-the-loop activities that support AI for the DET; to disseminate research about use of AI to various stakeholder communities; and to publish information about how AI is used on the DET, and usage of test-taker data.

Goal 4.1. Assess how AI processes impact stakeholders.

Stakeholders include test takers and organizations who use the DET, such as universities.

1. Document how **ML algorithms** (a) are used on the test and DET support resources, (b) are used for test design, measurement and security, and (c) impact stakeholders (i.e., high-stakes decisions based on DET outcomes).

2. Document **unintended risks** resulting from AI, such as biased scores or construct-irrelevant variance related to design, such as a test-taker's unfamiliarity with "drag-and-drop" options. Unintended risks may result in negative consequences for stakeholders, such as unfair or inappropriate admissions decisions.

3. Document **external factors** that result in a need to modify AI. Examples might include: a) institutional policy changes, such as, new language proficiency requirements require new item types; b) modifications in institutional use cases, such as, different rating practices; and, c) demographic changes, such as, increases in particular language groups that might impact differential item functioning (DIF).

4. Document how AI is used for DET **stakeholder support**. Examples of support include test readiness resources for test takers, and score interpretation guidance for organizations.

Goal 4.2. Document AI used for building the validity argument, test item creation, test item calibration, and scoring.

1. Document theoretical claims, and empirical studies to support the **DET validity** argument – i.e., evidence that the test is suitable for its intended purpose.

2. For **item creation**, document rationales for, and descriptions of item generation methods, including data and algorithms, human expert processes (such as fairness and bias review), and evaluation methods that provide a clear explanation of system performance metrics, and fairness evaluations, such as mitigating bias with regard to item content generation.

3. For **test item calibration**, document rationales for, and descriptions of AI for predicting item parameters, such as item difficulty.

4. For **test scoring**, document (a) rationales, descriptions, and alignment between item subconstructs and computationally-derived features used for scoring. This is relevant for constructed-response tasks involving spoken and written responses; and, (b) rationales for, and descriptions of measurement methods used to generate DET scores.

Goal 4.3. Document processes for human-in-the-loop interactions with AI.

These processes include, system development, evaluation, and data preparation and oversight.

1. Document **sustainable processes requiring 100% human expertise** (e.g., data annotation), or human expert supervision, such as fairness & bias review, monitor language model hallucination for automatically-generated item content. Sustainability is measured in terms of time, costs, and required resources (e.g., need for third parties, such as hiring contractors to support human FAB review, and annotation supporting AI development).

2. Document **qualifications of human experts**, such as software engineers, AI researchers, and assessment researchers, who are responsible for supervision during system development and evaluation, and piloting and operational deployment phases.

3. Document **demographic composition of human experts** who participate in annotation (e.g., scoring written responses) and reviewing tasks (e.g., FAB review).

4. Document **supports to help individuals understand and carry out their responsibilities** in relation to interacting with AI. Supports may include system UX, alert and reporting functions, and rubrics.

5. Document **content to help individuals' understand how AI is applied** on the DET. Document (1) AI systems' intended uses, such as text generation, scoring, test security, (2) empirical evaluations and interpretations of AI system behavior, and 3) acknowledgement of potential automation bias – specifically, favoring system outputs.

Goal 4.4. Document human expert qualifications required for human-in-the-loop activities that support AI for the DET.

1. Document qualifications criteria for human experts that are specific to activities that support human-in-the-loop AI. Such documentation may be applied for hiring practices, such as job descriptions.

Goal 4.5. Disseminate research about use of AI to various stakeholder communities.

1. Document research to illustrate how the **DET validity argument** was constructed with attention to AI.

2. Disseminate research about **theoretical, and quantitative, qualitative and mixed-methods research** through peer-reviewed, external publications and presentations See DET Research page.

3. Prioritize **peer-reviewed, open-access venues and publications**, and provide **public access** to peer-reviewed presentations.

4. Document, disseminate, and update **white papers about internal DET research** through the DET website, such as DET Research page.

5. Disseminate **external media publications**, such as blogs, that provide clear and plain language explanations of complex DET processes, such as test security. These publications are intended to render complex concepts transparent and accessible for the broader stakeholder community.

Goal 4.6. Publish information about how AI is used on the DET, and usage of test-taker data

1. Display on the DET website documentation about **how AI is used on the DET** so that it is understandable to stakeholders. For example, include content about automated test item creation and scoring, and test security on the DET website. Information should be publicly-available, such as in FAQs on the DET website, or a section devoted to explanations about how AI are used on the test.

2. Display documentation on the DET website. about **how stakeholder data is used**, including but not limited to biometric and personally-identifying data, such as IDs for security; process, such as keystroke profiles; and, product response data and test scores.

References

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational & psychological testing. American Educational Research Association.

Belzak, W. C. (2022). The Multidimensionality of Measurement Bias in High-Stakes Testing: Using Machine Learning to Evaluate Complex Sources of Differential Item Functioning. Educational Measurement: Issues and Practice, 1-10.

Belzak, W., Lockwood, J. R., & Attali, Y. (2024). Measuring Variability in Proctor Decision Making on High-Stakes Assessments: *Improving Test Security in the Digital Age. Educational Measurement: Issues and Practice*.

Bentley, C., Aicardi, C., Poveda, S., Magela Cunha, L., Kohan Marzagao, D., Glover, R. Rigley, E., Walker, S., Compton, M. & Acar, O. (2023). A Framework for Responsible AI Education. *Available at SSRN.*

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Burstein, J., LaFlair, G., Kunnan, A.J., & A. von Davier (2022). A Theoretical Assessment Ecosystem for a Digital-First Assessment - The Duolingo English Test. Duolingo Research Report DRR-22-01: 1-32.

Cardwell, R., Naismith, B., LaFlair, G., Nydick, S. (2024). *Duolingo English Test: Technical Manual* (Duolingo Research Reports). Duolingo. https://go.duolingo.com/dettechnicalmanual

Care, N. and Maddox, B. (2021). Improving Test Validity and Accessibility with Digital-First Assessments. Duolingo.

Chapelle, C., Enright, M., & Jamieson, J. (2008). Building a validity argument for the Test of English as a Foreign Language. Routledge.

Council of the European Union (2023, December 9). Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world. [Press release]. https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

Department for Science, Technology & Innovation (2023). A pro-innovation approach to AI Regulation [White Paper]. Crown. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf.

Dieterle, E., Dede, C., & Walker, M. (2022). The cyclical ethical effects of using artificial intelligence in education. AI & SOCIETY, 1-11.

Dignum, V. (2021). 'The role and challenges of education for responsible AI'. London Review of Education, 19 (1), 1–11.

Duolingo English Test (2021). Duolingo English Test: Security, Proctoring, and Accommodations. Duolingo.

Fiesler, Casey and Garrett, Natalie .(16 Sept 2020). Ethical Tech Starts with Addressing Ethical Debt, Wired Ideas: https://www.wired.com/story/opinion-ethical-tech-starts-with-addressing-ethical-debt/

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In S. Carta (Ed.), Machine learning and the city: Applications in architecture and urban design (pp. 535–545). John Wiley & Sons Ltd. https://doi.org/10.1002/9781119815075.ch45

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Srikumar, M. (2020). "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society.

Gianni, R., Lehtinen, S., & Nieminen, M. (2022). Governance of Responsible AI: From Ethical Guidelines to Cooperative Policies. Frontiers in Computer Science, 4.

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

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. Journal of Educational Measurement, 59(3), 338-361.

Huggins-Manley, A. C., Booth, B. M., & D'Mello, S. K. (2022). Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments. Journal of Educational Measurement, 59(3), 362-388.

International Test Commission and Association of Test Publishers (2022). Guidelines for technology-based assessment.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

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

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. Journal of Educational Measurement, 59(3), 338–361.

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112(3), 527.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), Fairness and validation in language assessment (pp. 1–14). Cambridge, UK: Cambridge University Press.

LaFlair, G. T., Langenfeld, T., Baig, B., Horie, A. K., Attali, Y. & von Davier, A. A. (2022). Digital-First assessments: A security framework. Journal of Computer Assisted Learning, 38(4), 1077–1086.

Liao, M., Attali, Y., von Davier, A. A., & Lockwood, J. R. (2022a). "Quality Assurance in Digital-First Assessments." Quantitative psychology: The 86th Annual Meeting of the Psychometric Society, virtual, 2021., (pp. 265-276). Cham, Switzerland; Springer.

Liao, M., Attali, Y., Lockwood, J. R., &amp; von Davier, A. A. (2022b). Maintaining and monitoring quality of a continuously administered digital assessment. Frontiers in Education, 7.

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. Computers and Education: Artificial Intelligence, 5. https://doi.org/10.1016/j.caeai.2023.100152

Mökander, J., & Axente, M. (2023). Ethics-based auditing of automated decision-making systems: Intervention points and policy implications. AI & Society, 38, 153–171. https://doi.org/10.1007/s00146-021-01286-x

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. Minds and Machines, 31(2), 323-327. https://doi.org/10.1007/s11023-021-09557-8

National Institute of Standards and Technology (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 429–435).

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (FAT*'20), January 27–30, 2020, Barcelona, Spain. (pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

OECD (2023), "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI." OECD Digital Economy Papers, No. 349, OECD Publishing, Paris.

The White House. (2023, October 30). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. Journal of Educational Measurement, 54(1), 3–11.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13.

Wodzak, S. (2021, September 14). What if tests were delightful? https://blog.duolingo.com/what-if-tests-were-delightful/

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? Language Testing, 27(3), 291–300.

duolingo
english test