KEY RESEARCH THEMES IN ETHICS AND ARTIFICIAL INTELLIGENCE IN EDUCATIONAL ASSESSMENTS

Tristan Lim[*,a]

[a] Nanyang Polytechnic, School of Business Management, Singapore

*tristan_lim@nyp.edu.sg

Abstract

This systematic literature mapping study aim to provide practical insights on the ethics of artificial intelligence (AI) in assessment. It is important to study the divide between what may be ethically permissible and not permissible, especially in fundamental societal institutions like education, when teaching practitioners or researchers apply AI in academic processes such as assessments. This study applied a systematic literature mapping methodology to scour extant research, so as to holistically structure the landscape into explicit topical research clusters. Through topic modelling and network analyses, research mapped key ethical principles to research archetypical domains, and reviewed the influence of these ethical principles in each thematic domain. Results of this study identified five key research archetypical themes, with presence across the system layers of cognitive, information and physical domains of an AI-based assessment pipeline, namely: (i) AI system design and check for assessment purposes; (ii) AI-based assessment construction and rollout; (iii) data stewardship and surveillance; (iv) administration of assessments using AI systems; and (v) AI-facilitated assessment grading and evaluation. Ten AI ethics principles, namely, (i) fairness, (ii) privacy, (iii) explainability, (iv) accountability, (v) accuracy, (vi) inclusivity, (vii) trust, (viii) human centricity, (ix) auditability and (x) cheating, epitomize the key ethics considerations across each of the five research themes; each manifesting varying levels of importance. The findings of this research can provide researchers and practitioners the insights into the application methods of AI in assessments and their intertwined ethical challenges, and in particular, the generalizable key research themes structured across the assessment pipeline, for follow up studies.

Keywords: *artificial intelligence in education (AIED), assessment, ethics, systematic literature mapping*

Introduction

Artificial intelligence in education (AIED) is the machine mimicry of human-like consciousness and behavior to achieve educational goals, through the use of technology that allows digital systems to perform tasks commonly associated with intelligent beings.

Of the three pillars of education, assessment exists as an important component, alongside pedagogy and curriculum (Hill and Barber, 2014). Within the AIED domain, Chaudhry and Kazim (2022) scoured the landscape and concluded that assessment is one of the four key sub-domains in AIED, alongside learning personalization, automated learning systems, and intelligent learning environments. In an educational context, assessment refers to *'any appraisal (or judgment or evaluation) ... of work or performance*' (Sadler, 1989). The infusion of artificial intelligence (AI) in assessments has grown significantly in recent years. Research on assessments related to digital education in the higher education landscape showed that AI and adaptive learning technologies have tripled between 2011 to 2021 and is likely to surpass immersive learning technologies as a prime research area in the near future (Lim, Gottipati and Cheong, 2022). Among stakeholders, there is a consensus positive view that "*AI would provide a fairer, richer assessment system that would evaluate students across a longer period of time and from an evidence-based, value-added perspective*" (Luckin, 2017).

Infusion of AI in assessments also brings along its own set of concerns. AI implementation comes with technical and operational issues relating to system implementation. Arguably, these challenges have relatively lesser grey areas to contend with, than the complication of navigating the parameters and boundaries of ethics. Evaluators, as practitioners of assessments, will need to acknowledge, respect, and uphold ethical principles that may plague the implementation of an AI-based assessment.

The research objective of this study is to examine the landscape of AI-related ethical issues for educational assessments, through the lens of a systematic literature mapping approach. A systematic literature mapping study is a study concerned with the mapping and structuring of a topical research area, the identification of gaps in knowledge, and the examination of possible research topics (Petersen, Vakkalanka and Kuzniarz, 2015). The research novelty and value of this work lies in the notable lack of research providing a holistic inspection and review of the aforementioned landscape.

This study investigates the following research questions:

- *RQ1: What are the main AI use cases and ethical issues relating to assessments?* This question looks at AI applications and ethical principles in different assessment areas, and how dominantly each area is featured.
- *RQ2: What are the key themes of the systematic literature map?* This question looks to identify key themes of the systematic literature map, and draw up a framework to visualize and generalize the key themes for researchers and practitioners.

The significance of this research is, through a systematic meta-analysis of existing literature in the field, (i) understand and consolidate knowledge regarding what was previously explored relating to AI-based assessment methods and their interconnected ethical issues, (ii) provide an integrated inquiry into the association of the ethical problems faced, and (iii) identify potential future research topics in the field.

Results of this study identified five key research archetypical themes, with presence across the system layers of cognitive, information and physical domains of an AI-based assessment pipeline, namely: (i) AI system design and check for assessment purposes; (ii) AI-based assessment construction and rollout; (iii) data stewardship and surveillance; (iv) administration of assessments using AI systems; and (v) AI-facilitated assessment grading and evaluation. Ten AI ethics principles epitomize the key ethics considerations across each of the five research themes; each manifesting varying levels of importance.

The remainder of the paper is organized as follows: (i) the Methodology section discusses the systematic literature mapping approaches undertaken, explains the machine learning methods utilized; (ii) the Findings section presents the tables and graphic visualizations from topic modelling, and network analyses, and provides in-depth analyses of the data.; (iii) the Conclusion section summarizes the key findings, impact of paper, and closes with proposed future work that can be studied by practitioners and researchers.

Methodology

In this study, we apply the systematic literature mapping approach. The study was conducted using the research methodology in Kabudi, Pappas and Olsen (2021). We apply the methodology undertaken as follows, namely: (i) search and selection, (ii) data extraction, and (iii) classification and analysis.

PRISMA approach, or the Preferred Reporting Items for Systematic Reviews and Meta-Analyses approach, was employed as a guideline to conduct the search and selection phase (Moher et al., 2009). In accordance with the recommended methodology as part of the PRISMA-P checklist, details including the eligibility criteria, sources of information, search protocol, research records, data items and synthesis of data are described in the following sub-sections.

Vivo11, EndNote X9 and *Excel* spreadsheets were used for information organization. Further information extraction, data visualization, and machine learning tools and techniques are described in the following sub-sections.

Search and Selection

As AIED researchers stem from a variety of fields publishing across a wide range of publications, literature search was conducted using Scopus, an interdisciplinary rigorously curated database covering the widest range of disciplines (240 disciplines) relative to similar citation databases, with contents including over 87 million publication items, 1.8 billion cited references, 17 million author profiles, 94,000 affiliation sources and 7,000 publishers. On average, each paper indexed on Scopus has 10% to 15% more citations than similar databases (Elsevier, 2022), which implies a more extensive systematic literature mapping analysis. Summary of PRISMA approach is shown in Fig. 1.

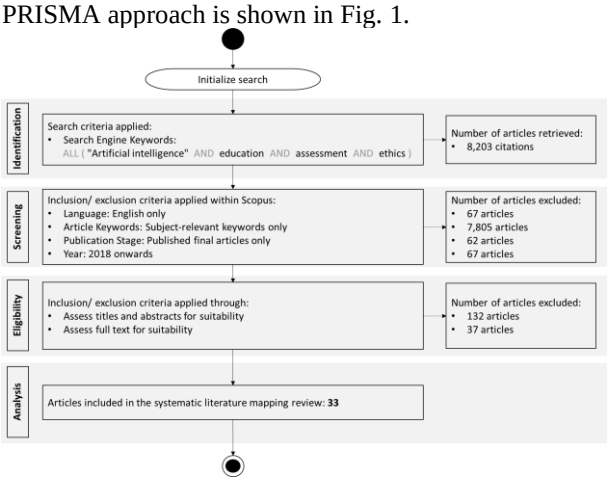

Figure 1: PRISMA - The systematic mapping process

The first stage of PRISMA, or the *identification stage*, identifies the possible papers to be considered using the Scopus search engine. The search entry was as follows: *ALL ("Artificial intelligence" AND education AND assessment AND ethics).* This stage identified a corpus of 8,203 papers.

The second stage of PRISMA, or the *screening stage*, looks at excluding inappropriate and unrelated papers. This stage reduced the corpus count to 202. Search applied the following inclusion criteria:

- *Language:* Only articles written in English language were included. This step omitted 67 articles.
- *Keywording:* Only articles with subject-relevant keywords coded by Scopus for indexing purposes (also known as *Indexed Keywords* by Scopus) were included. This step omitted 7,805 articles.
- *Publication Stage:* Only peer-reviewed final articles published in scientific venues (e.g., books, journals and conferences) were included, for rigority of selection. This step omitted 62 articles.
- *Year of Publication:* Only articles published in 2018 and beyond were included, to ensure recency of literature. Rigorous peer-reviewed articles would

have reviewed key prior related literature within their respective papers. This step omitted 67 articles.

The third stage of PRISMA, or the *eligibility stage*, requires scanning title and abstracts, and full papers to identify relevant eligible articles. This stage yielded a final corpus count of 33 articles. Search applied the following inclusion criteria:

- *Assess Titles and Abstracts for Suitability:* Only relevant titles and abstracts were included. There should be explicit and direct references to the subject matter. This step omitted 132 articles.
- *Assess Full Papers for Suitability*: Only relevant full papers were included. An additional inclusion criterion here was that all articles should have their full text accessible for analysis. This step omitted 37 articles.

Data Extraction

As a citation engine, data in Scopus is highly structured and robustly tagged, delivering metadata for analytical purposes, including (i) author(s), (ii) document title, (iii) affiliation(s), (iv) year, (v) publication, (vi) volume, issue and page source, (vii) citation, (viii) document type, (ix) keywords, and (x) digital object identifier (DOI), among others.

The final pool of 33 primary studies were analyzed to answer the research questions of this study. Information that was extracted from Scopus included: (i) citation information, such as author(s), title, year, publication, and citation count etc., (ii) bibliographical information, such as affiliation(s), and publisher etc., (iii) abstract, (iv) keywords, and (v) references.

Classification and Analysis

Using the data extracted from Scopus, the study utilized *Tableau Desktop Professional version 2021.1.20* to perform exploratory data analyses to address RQ1. Tableau platform allows powerful conversion of complex computations into appealing data visualizations.

With the Scopus extracted data, research utilized a corpus analysis platform *CorTexT* (Breucker et al., 2016) to perform text parsing, and a first pass of topic modelling and network mapping, so as to identify major thematic representations of corpuses comprising of *Author Keywords* and *Indexed Keywords*. This allowed us to perform machine learning for pattern recognition, utilizing unsupervised text mining techniques on these keywords to identify useful patterns.

Using the Python Library *pyLDAvis* (Sievert and Shirley, 2014), topic modelling generated a topic representation of the keyword corpus' textual fields using the *Latent Dirichlet Allocation* method, which allowed a visualization of the most relevant words fitting to the topic. Here, each topic was defined as a keyword probability distribution, and each document was defined as a topic probability distribution. Given the total number of topics defined, the topic model was inferred by probabilistically assigning topics to documents, and positioned in 2D according to a multi-dimensional scaling algorithm for visualization purposes.

While topic modelling provided a sense of the latent themes from the underlying keywords, research further performed network analyses to visualize thematic keyword representations in a clustering format, where each keyword was grouped with distinct members, and linked via proximity measures. The Louvain hierarchical community detection algorithm was used. This algorithm is based on modularity optimization, where the optimal linkage densities are measured, taking into account within-cluster and between-cluster linkages. Louvain algorithm is efficient on large networks (Aynaud, 2020).

The first pass of topic modelling and network analyses above allowed the identification of distinct sub-themes of AI application areas and ethical issues. With the key sub-themes of AI application areas and ethical issues identified as a priori, open and axial coding were conducted for each article to classify the following: (i) application areas where AI is used in assessments (e.g., assessment curation and personalized feedback etc.), and the (ii) type of ethical issues relevant to AI-based assessments as cited in paper (e.g., fairness and explainability etc.). This would allow us to address RQ1.

Using the coded sub-themes of AI application areas and ethical issues, research undertook the second pass of topic modelling and network analyses. The topic modelling and network analyses outputs would be used to guide the identification of the major research themes to address RQ2.

Findings

RQ1: Main AI use cases and ethical issues

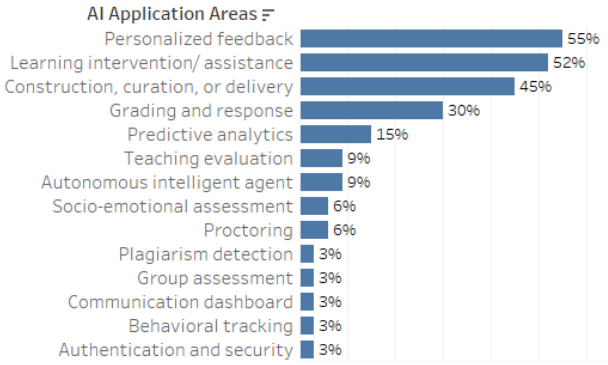

Figure 2: AI application areas and citation proportion

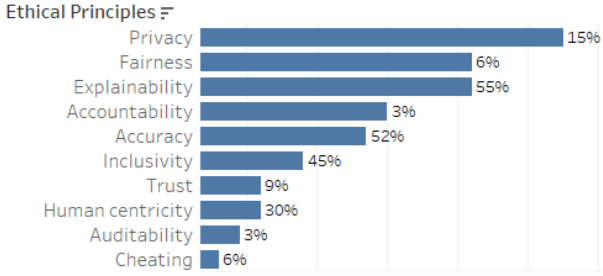

Figure 3: Ethical principles and citation proportion

To address RQ 1, topic modelling was performed, where the optimal number of topics were generated using a model with the highest topic coherence. Further, we

performed network analyses to identify topic clusters. These allowed us to recognize patterns in an unsupervised machine learning approach.

From this first pass of topic modelling, ten latent topics were identified. This aligned well with network analyses, where we observed a more granular fourteen latent topic clusters. The higher granularity of the outputs allowed us to identify distinct sub-themes of AI application areas and ethical issues. Through the review of the first pass of topic modelling, network analyses outputs, and full paper reviews, the study extensively identified fourteen sub-themes of AI application areas and ten sub-themes of ethical issues. We populate them in Figures 2 and 3, respectively.

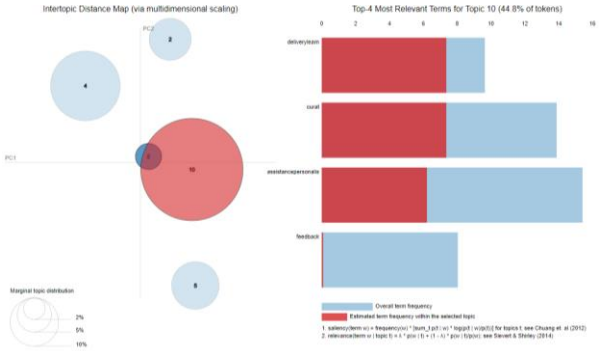

Figure 4: Topic modelling of corpuses involving AI application areas and related ethical principles

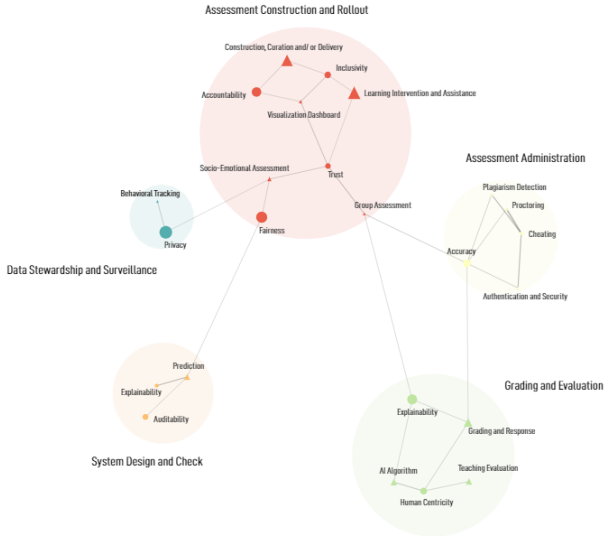

Figure 5: Network analyses of corpuses involving AI application areas and related ethical principles

Table 1: Latent topics and top keywords

Topic No.	Latent Topic	% Tokens	Top Keywords
2	System design and check	*7.5%*	System; Design; Review
9	Data stewardship and surveillance	*2.9%*	Privacy; Sensitive; Data
10	Assessment construction and rollout	*44.8%*	Deliver; Curate; Personalize
5	Assessment administration	*9.6%*	Proctor; Plagiarism; Cheat
4	Grading and evaluation	*20.7%*	Evaluation; Feedback; Response

Next, we utilize the keyword corpuses of fourteen sub-themes of AI application areas and ten sub-themes of ethical issues as an input, to perform the second pass of topic modelling, and network analyses.

Research identified five topical archetypes via topic modelling. For instance, in Fig. 4, we observed the dominant latent topic number 10 linked to AI-based assessment construction and rollout aspects. This aligned well with the network analyses visualization in Fig. 5. In the network analysis diagram, we observed a clear clustering of five topics, with *Assessment Construction and Rollout* similarly dominant in the cluster diagram. The top keywords and latent topics of topic modelling are shown in Table 1.

RQ2: Main AI use cases and ethical issues

Ashok et al. (2022) describes three fundamental domains to conceptually represent the interweaving ethical elements and interrelationships inherent in the design and application of AI in digital technologies. This triadic framework is a modular architecture of an assemblage of technological components that consist:

- *Physical domain (or the referent or object in semiotics):* This includes the device and network layer. Some relevant applications are author systems, intelligent tutoring shells, AI-integrated learning environments, and educational robotics.
- *Cognitive domain (or the symbol or science in semiotics):* This comprises the content layer where data is stored, created, mapped, manipulated, utilized, and shared. Some relevant examples are multimodal structured contents of text, and unstructured contents of images and videos of assessment submissions.
- *Information domain (or the reference or interpretant in semiotics):* This comprises the service layer which encompasses the functionality of the application and its interaction with users, underpinned by AI algorithms. Some relevant examples are use of knowledge representation for instructions, human factor and interface design, and AI-integrated visualization and graphics for feedbacks.

We extend the triadic ontological framework as described by Ashok et al., (2022) to model and visualize the systematic literature map of this paper (Fig. 5). The significance of PCI would enhance understanding of the description of the 5 archetypes below. The five distinct archetypes identified by topic modelling and network analyses in Fig. 4 and Fig. 5 are mapped to the triadic ontological framework in Fig. 6, as follows:

- *AI system design and check for assessment purposes*

  This archetype extends across the physical, cognitive and information domains, and is involved with the design, implementation and maintenance of the AI system for system interactivity, robustness and security. From a predictive analytics point of view, the model constructed should be appropriate – upholding accuracy, inclusivity, accountability, privacy, trust and human centricity.

Here, the overriding ethics considerations are explainability and auditability. The AI system should be created with clear, easy-to-understand and transparent protocols, so that relevant stakeholders and independent third-party auditors can review the processes, perform interventions, mitigate issues, and enable redress in an event of negative outcomes that may arise. In addition, fairness is concerned about the treatment of algorithmic bias to ensure diversity, equity, non-prejudice and non-favouritism towards learners' sensitive attributes, so that needs of minority groups are not disadvantaged or underrepresented.

- *Data stewardship and surveillance*

This archetype extends across the cognitive and information domains, and is involved with the governance and implementation of appropriate data stewardship, and surveillance practices (if any).

Here, the overriding ethics consideration is privacy. One instance is behavioural surveillance, which may be a violation to human rights to privacy especially when data is used beyond academic purposes, for control and surveillance to modify human behaviour. In addition, trust is also an important facet concerned about the preservation of privacy when sensitive data are disclosed.

- *AI-based assessment construction and rollout*

This archetype is predominantly situated in the information domain, and is involved with the construction, curation or delivery of assessment, the communication of evaluation and feedback with stakeholders via AI-integrated communication dashboards, and the carrying out of interventions and assistances to improve assessment and evaluation performance. Assessment and evaluation can be in the form of formative (or summative) individual (or group) cognitive (or socio-emotional) assessment. It can also be a form of teaching evaluation.

Here, the overriding ethics considerations are inclusivity and fairness, so that appropriate and equitable assessments and evaluations are rolled out, embracing diversity, empathy and sensitivity towards the evaluated stakeholders. Furthermore, accountability is an important ethics consideration, as there should exist a responsible discharge of AI ethical principles and compliance with relevant rules and guidelines, when designing and delivering AI-driven assessments. In addition, there should exist trust and confidence on AI systems to achieve assessment and evaluation objectives.

- *Administration of assessments using AI systems*

This archetype is predominantly situated in the information domain, and is involved with the administration of assessment and evaluation, which may comprise authentication and security measures, proctoring and/or plagiarism detection.

Here, the overriding ethics considerations are the overcoming of cheating violations, and the application of accuracy to correctly identify assessment candidates and cheating cases.

- *AI-facilitated assessment grading and evaluation*

This archetype is predominantly in the information domain, and is involved primarily with the interpretation of textual and/or audio-visual responses collected by AI systems, the evaluation of performance, and the provision of feedbacks. These may be performed by autonomous intelligent agents. From an educator's point of view, this phase may involve the evaluation of teaching effectiveness.

Here, the overriding ethics considerations are explainability, so evaluators can understand and adjudge if the grading and/or ranking is accurate and reliable. In addition, there is an element of human centricity. This largely relates to the agency and autonomy of human users, in the presence of AI-generated decisions, and the capacity to intervene for correction and redress.

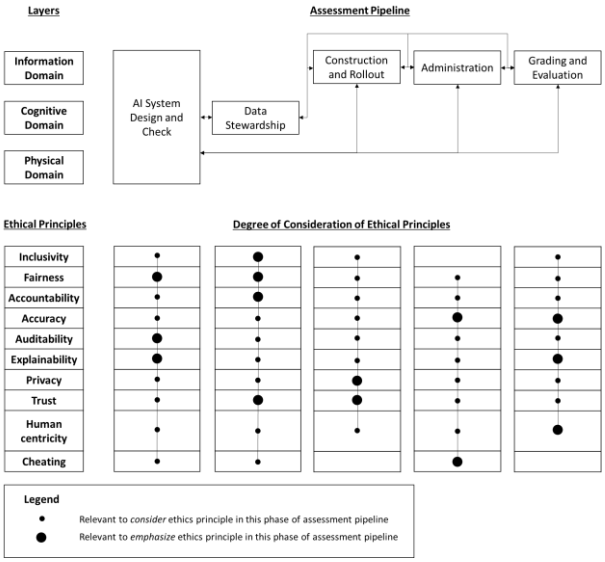

Figure 6: Visualization of the systematic literature map of key research themes

There is an emphasis that the framework does not draw clear delineations when categorizing AI assessment use cases across triadic domains. For instance, the *Grading and Evaluation* research theme is predominantly arising from the cognitive domain. However, coding and rolling out a moral reasoning AI system for AI-generated decisions, evaluations, responses and feedbacks, a sub-item of this research theme, may straddle across all cognitive, information and physical domains. This said, the framework provides a guide to generalize observed phenomena.

Conclusion

As AI becomes more pervasive, it's important to establish ethical safeguards, particularly when there exists the possibility of anthropomorphic influence on AI. Society as a whole, and education institution in particular, should scrutinize the application of AI to mitigate potential violations of ethics, even as we push ahead to reap the benefits of AI.

In this study, we looked at how the design and use of AI in education, and in particular, assessments, can

conform as closely as possible to basic ethical principles. We systematically investigated the key assessment components and ethical principles highlighted in existing literature, mapped them across the end-to-end assessment pipeline while accounting for different assessment types, and constructed a systematic literature mapping framework highlighting key archetypical research themes. The proposed systematic literature mapping framework allows researchers and practitioners to deep dive into key thematic research.

Research identified five key archetypical research themes, namely (i) AI system design and check for assessment purposes, (ii) data stewardship and surveillance, (iii) AI-based assessment construction and rollout, (iv) administration of assessments using AI systems, and (v) AI-facilitated assessment grading and evaluation. Ten literature-derived ethical principles, namely, accuracy, privacy, human centricity, fairness, inclusivity, trust, explainability, cheating, accountability and auditability, were mapped to these research themes.

Future work can extend the use of literature databank beyond Scopus, to include e.g., *Web of Science*, *IEEE Xplore* or *EBSCO Host*, in the systematic literature mapping exercise. While this study is based upon the subject of assessments, the ethical elements of the discourse has relevance beyond assessments, and can be applied to other areas of AIED. Other future works can contribute to the examination on the underpinning theories relating the ontological, semantics, and the epistemological deliberations and practical applications of ethics in this subject matter, across the spheres of philosophy, learning, psychology, sociology and technology. In addition, practical applications of the actionable insights in this paper, in a form of strategic and operational frameworks or case studies, can be another pragmatic endeavor by practitioners and researchers.

Herwix et al. (2022) highlighted the importance of more serious and systematic engagement with the selection, framing and prioritization of ethical issues. There is an emphasis among the state-of-the-art for the need to be more aware, anticipatory, reflecting and informed about the variety of perspectives and contemporary debates concerning AIED ethics. In particular, the relevancy and idiosyncrasy to assessments in our study can help bring forward distinctive actionable applications in this realm.

References

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management*, *62*, 102433.

Aynaud, T. (2020). Python-louvain. Louvain algorithm for community detection. [Online]. Retrieved: https://github.com/taynaud/python-louvain [Assessed 15 Jan 2023].

Breucker, P., Cointet, J., Hannud Abdo, A., Orsal, G., de Quatrebarbes, C., Duong, T., Martinez, C., Ospina Delgado, J. P., Medina Zuluaga, L. D., Gómez Peña, D. F., Sánchez Castaño, T. A., Marques da Costa, J., Laglil, H., Villard, L., & Barbier, M. (2016). CorTexT Manager (version v2). [Online]. Retrieved: https://docs.cortext.net [Accessed 14 Jan 2023].

Chaudhry, M. A., & Kazim, E. (2022). Artificial Intelligence in Education (AIEd): a high-level academic and industry note 2021. *AI and Ethics*, *2*(1), 157-165.

Elsevier (2023). How Scopus works, Scopus. [Online]. Retrieved: https://www.elsevier.com/solutions/scopus/how-scopus-works/content [Assessed 14 Jan 2023].

Herwix, A., Haj-Bolouri, A., Rossi, M., Tremblay, M. C., Purao, S., & Gregor, S. (2022). Ethics in information systems and design science research: Five perspectives. *Communications of the Association for Information Systems, 50*(1), 589-616.

Hill, P., and Barber, M. (2014). *Preparing for a renaissance in assessment.* Pearson, London.

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017.

Lim, T., Gottipati, S, & Cheong, M. (2022). Authentic Assessments for Digital Education: Learning Technologies Shaping Assessment Practices. In *proceedings of 30th International Conference on Computers in Education.* 1, 587-592.

Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour, 1*, 0028. DOI: https://doi.org/10.1038/s41562-016-0028.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. DOI: https://doi.org/10.1371/journal.pmed.1000097.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, *64*, 1-18.

Sadler, D. R. (1989). Formative assessment in the design of instructional systems. *Instructional Science 18*, 119–144.

Sievert, C., & Shirley, KE. (2014). LDAvis: A method for visualizing and interpreting topics, *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70). Baltimore, Maryland, USA.