ISSN: 2278-4012, Volume:12, Issue:2, June 2023

Explainable Artificial Intelligence for Automated Language Assessment Systems

Nguyen Thi Dan Ta[1*]

[1*]Faculty of English, HCMC University of Economics and Finance, Binh Thanh District, Vietnam.

Abstract: Automated language assessment solutions are progressively being integrated into various language tests and instructional structures. The majority of these solutions rest on the deep learning framework, which provides high predictive accuracy, but often regarded as "black boxes," giving very little explanation concerning how a certain score has been obtained or a certain judgment has been made. Consequently, this may create problems for students, teachers, and developers of assessments, especially when a result obtained is of considerable impact on the learner's placement, certification, or academic progress. Explainable AI (XAI) presents a new approach aimed at solving the problem. XAI provides a set of techniques which allow to make the process of obtaining the score understandable for humans without affecting the prediction accuracy. This paper deals with the topic of the interrelation of XAI and automated language assessment. The author presents a classification of the available methods, based on the methods used, their location, and the audiences. In addition, the paper provides an overview of the existing techniques used in essay evaluation, short answer grading, and speech assessment. The aspect of validity, trust, and involvement of students in receiving feedback is also considered. The paper proposes his vision for further research in the field of construction of a reliable, transparent, and pedagogically effective automated language assessment system.

Key Words: explainable artificial intelligence, automated language assessment, automated essay scoring, interpretability, language testing, trust in AI, feedback

(Received: 16 March 2023; Revised: 28 April 2023; Accepted: 24 May 2023; Published: 30 June 2023)

Introduction

The advent of artificial intelligence in scoring and feedback systems has revolutionised language assessment. At present, automated systems such as automated essay scoring (AES), automated short answer grading (ASAG) and automated scoring systems are in use for placement testing, formative feedback in classrooms and in cases of high-stakes testing as well (Ke & Ng, 2019; Mohsen, 2022). These systems are based on machine learning and now increasingly deep learning, which is very useful to identify complex language patterns, but at the same time function as a black box (Adadi & Berrada, 2018). The link between the input and output might be linked through thousands of parameters, making it impossible for humans to decipher the actual logic behind the machine's outcome.

This lack of transparency is a more serious issue than it seems. In educational environments and assessments consequences of decisions made by the software have very important implications for the stakeholders, i.e. course placement, advancement, certification etc., thus, the users expect to receive an explanation about the final decision of the software (Adadi & Berrada, 2018; Arrieta et al., 2020). Explainability in general AI research has different dimensions, including transparency, interpretability, and the provision of human-understandable justifications for model outcomes (Arrieta et al., 2020; Miller, 2019).

ISSN: 2278-4012, Volume:12, Issue:2, June 2023

Evidently, education is one of those fields wherein explainability carries more weight due to some formative and evaluative functions that AI systems perform within learning environments (Khosravi et al., 2022). Predictive models in interpreting dropout risk, for instance, have shown that interpretability does not have to come at great cost in predictive accuracy, so educators can benefit from more informed intervention through explainable models (Baranyi et al., 2020). Research in intelligent tutoring systems has, in similar fashion, proposed methodological standards for how to really scrutinize whether an explanation is practically useful to the learner or the instructor receiving the information, apart from being technically correct (Clancey & Hoffman, 2021).

Within the educational context, automated language assessment presents a strikingly different set of demands. Language scoring is not merely a simple classification task; the system must usually balance things such as lexical sophistication, syntactic complexity, coherence, argumentation; and, in the case of speech assessment, also pronunciation and fluency, often at the same time (Ke & Ng, 2019; Bamdev et al., 2023). It is thus precisely the combinatorial nature of such judgments that has moved the field towards deep learning techniques, which tend to improve on traditional accuracy measures with feature-based models but are thus also much harder to explain Ke & Ng, (2019). Some recent work has begun to tackle this directly by showing that AES systems based on deep learning can become explainable without sacrificing the pedagogical value of the underlying representations are built (Kumar & Boulanger, 2020). Such work on validity arguments for AI-based scores underscores the notion that a score can only justifiably be defended if its rationale can be articulated and examined-a requirement that aligns closely with the goals of XAI.

Despite the convergence of interests, therefore, most of the existing literature treats explanations and automated language assessments as separate literatures: one dealing with the technical mechanics of interpretable machine learning and the other with the psychometric and pedagogical properties of scoring systems. Only a little number of works have, however, juxtaposed both lines of inquiry to inquire what form explanation should take in the case of language testing, where explanations must make sense to the system developers and also to language learners and teachers, who have a range of technical and linguistic metafunction backgrounds (Khosravi et al., 2022; McCarthy et al., 2022).

The main contribution of this paper can be considered threefold from several points of view. First of all, it introduces a taxonomy for organizing the approaches to explainability in automated language assessment. After that, it employs this taxonomy to compare the existing techniques in essays scoring, short-answer evaluation, and assessment of spoken performance, focusing on the strengths and weaknesses of the methods currently being used. And finally, it applies all the technical differences to practical implications for assessment design and affects both learners' trust and teaching practices in the classroom.

In this paper, the gap identified in the area has been bridged by analyzing the use of XAI technologies and principles in automated language assessment systems. The paper is structured in the following way. Section II discusses published materials related to the interpretable techniques used in the automated scoring and feedback systems. Section III offers a schema for the taxonomy of the interpretability methods for language assessment. Section IV compares different methods concerning different kinds of assessments. Section V presents issues of trust, educational effect, and other remaining shortcomings of the field. Finally, Section VI describes the implications in terms of the design of assessments and areas for further investigation.

ISSN: 2278-4012, Volume:12, Issue:2, June 2023

Review of Related Literature

In explaining explainable AI, theoretical principles lie largely far from computer science and include studies on how people generate and evaluate explanations based on human reasoning. It has been claims that the social foundation is crucial since any explanation that is technically complete not necessarily corresponds to human perception and may face incomprehensibility even if its could provide correct information. The importance of the findings is obvious for such an automated operation in language assessment since in this case not computer science specialists are the end users of explanations, but either language learners or teachers, while the use of technical terms in explanations won't make them appealing to learners.

In case of automated writing assessment, the feedback systems have for a long time been equipped with some indicators concerning spelling and grammar so that research has been interested in the effect of such feedback on the overall efficacy of revision process. Overall findings are confirming that the purely mechanical nature of feedback hampers the overall effectiveness partly proving the idea that more meaningful and detailed explanations are needed instead of the narrow mechanisms of error flagging. In this context works have shown that providing learners with such types of feedback may entail certain advantages for the students in its very presentation and perception. This body of research shows that just the existence of automated feedback does not mean that it is educationally effective; the form of the response provided by the automated feedback can influence the learner's response to it.

Recent studies report the application of explainability techniques to the automated scoring methods. In the case of grading short answers using automatic grading methods, attention-based justification techniques have already been applied in the study of which part of the input in the transformer model has determined the resulting grade and allowed to provide explanations for grades of particular answers rather than generalized scoring based on some features (Poulton & Eliens, 2021). A more thorough study of the automated essay scoring shows a historical process of how the models changed from being statistically based and easy to understand to deep learning methods which enhanced accuracy of the automated scoring but decreased interpretability of results by causing many controversial issues in this field of study (Ramesh & Sanampudi, 2022). The topic of similarity scoring methods has also demonstrated that although quite interpretable when compared to deep learning methods are insufficient in covering all the acceptable linguistic varieties used by students while completing writing tasks (Ramnarain-Seetohul et al., 2022). Studies about how people learning second languages interact with automatic feedback further highlight the importance of trust in feedback that is perceived to be credible. Indeed, it can be concluded that the higher the trust in feedback, the more possibility for it to be applied, and vice versa, the lower the trust in feedback perceived as arbitrary, the less possibility for it to motivate any action (Ranalli, 2021). This is similar to the issue of explainability since it plays an important role in how the users perceive the use of automatic assessment as a way of learning. Studies in the field of automatic speech assessment also address this question since it's necessary to understand how to interpret the factors leading to a particular decision about a certain level of proficiency, based on speech assessment technologies such as pronunciation, prosody, and fluency measurement methods that would ensure the trustworthiness of automatic speech assessment (Bamdev et al., 2023).

In this case, other aspects of the XAI literature provide different structural models of reasoning in information systems. One way to achieve this goal is to use knowledge graphs allowing for representing the relations between different components of the model (Rajabi & Etminani, 2024). Such application can have a positive impact on the field of language testing since it uses the principle of correlation between linguistic knowledge and connections between grammatical categories. However, some scholars argue that

ISSN: 2278-4012, Volume:12, Issue:2, June 2023

language assessment practitioners should prefer the use of interpretable systems instead of post-hoc explanation of black-box systems that may not be reliable when explaining how the decisions have been made Rudin, (2019). This argument is relevant for language assessment since scoring decisions might have serious consequences.

Lastly, the findings of the research related to the automated feedback provided to students regarding the written assignments show that explanation quality is vital since it determines learning efficiency. Studies conducted in this field demonstrate that if automated feedback is based on certain textual characteristics, the learner's efficiency rises as well (Zhu et al., 2020). This conclusion can be also supported by the idea about explainability present in multiple projects aimed at the creation of trustworthy AI systems that can rationalize their decisions (Gunning & Aha, 2019).

In summary, the body of research shows that explainability in automated language assessment cannot simply be considered a technical solution applied post hoc to the model. It is seen that explanation value is influenced not just by the design but also by the chosen methodology and the audience. Therefore, it is recommended that the explanation be developed in tandem with the scoring system.

A Taxonomy of Explainability in Automated Language Assessment

The Explanation's Locus

The explanation's locus concerns the nature of the explanation's genesis; either it is generated inherently by employing a model that is already interpretable i.e., one that obtains its informativity from being constructed in a way that allows inspection of the contribution of each of the features toward the score, or it depends on an externally applied explanation tool to explain the output of a non-interpretable model. For example, feature-based scoring models heavily depend on statistical scores based on the input features, such as lexical sophistication or syntactic complexity. This makes them inherently interpretable since one can analyze the contribution of each feature toward the resulting score. On the other hand, deep learning models rely on external mechanisms for their interpretation, such as attention visualization or perturbation methods.

The Explanation's Method

The method dimension is related to the specific technique utilized in generating an explanation. There are a number of methods employed in generating the explanations, including attention weighting, feature attribution, exemplar-based comparison, and graph-based reasoning. Each method has its own shortcomings in terms of fidelity and comprehensibility. For instance, the attention-based approaches enable one to achieve a detailed understanding of how the model works but necessitate some technical knowledge for the correct interpretation. On the contrary, the approaches based on knowledge graphs allow for obtaining a simple explanation but require the underlying linguistic knowledge to be known beforehand.

The Explanation's Audience

The audience dimension has the highest importance for language assessments. Explanations need to be delivered either to the developer who conducts validation, to teachers analyzing group scores, or to students receiving feedback. While explanations targeting teachers of students need to use certain pedagogical terms such as sentence or argumentation structures, those directed to students should avoid using the technical vocabulary and rely on pedagogically appropriate terms.

ISSN: 2278-4012, Volume:12, Issue:2, June 2023

Table 1: Taxonomy of Explainability Approaches in Automated Language Assessment

Locus	Representative Method	Typical Assessment Application	Primary Audience
Intrinsic	Explicit feature weighting	Feature-based AES scoring	Developers, teachers
Extrinsic	Attention visualization	Transformer-based ASAG	Developers, advanced teachers
Extrinsic	Feature attribution / perturbation	AES revision feedback	Learners, teachers
Structural	Knowledge-graph reasoning	Grammar- and lexis-focused feedback	Teachers, curriculum designers
Multimodal	Acoustic cue isolation	Automated speech scoring	Examiners, learners

Table 1 gives a brief summary of how the locus, method, and common use of approach determine the suitable audience for which it will be beneficial. The table emphasizes that there is no approach which benefits all stakeholders equally well, which makes it reasonable to regard audience as a design dimension rather than a lagging consideration. The conceptual framework that connects the key aspects of explainable artificial intelligence to the decisions made with regard to assessment design in automated language evaluation systems is presented in figure 1. The explanations in the framework are further classified based on their location, means of working, and target audience. Accordingly, these aspects lead to several important decisions regarding the process of designing the feedback, transparency, interpretability, fairness, and decision support. All the points made above show how design decisions lead to positive results in terms of learning outcomes achieved by students.

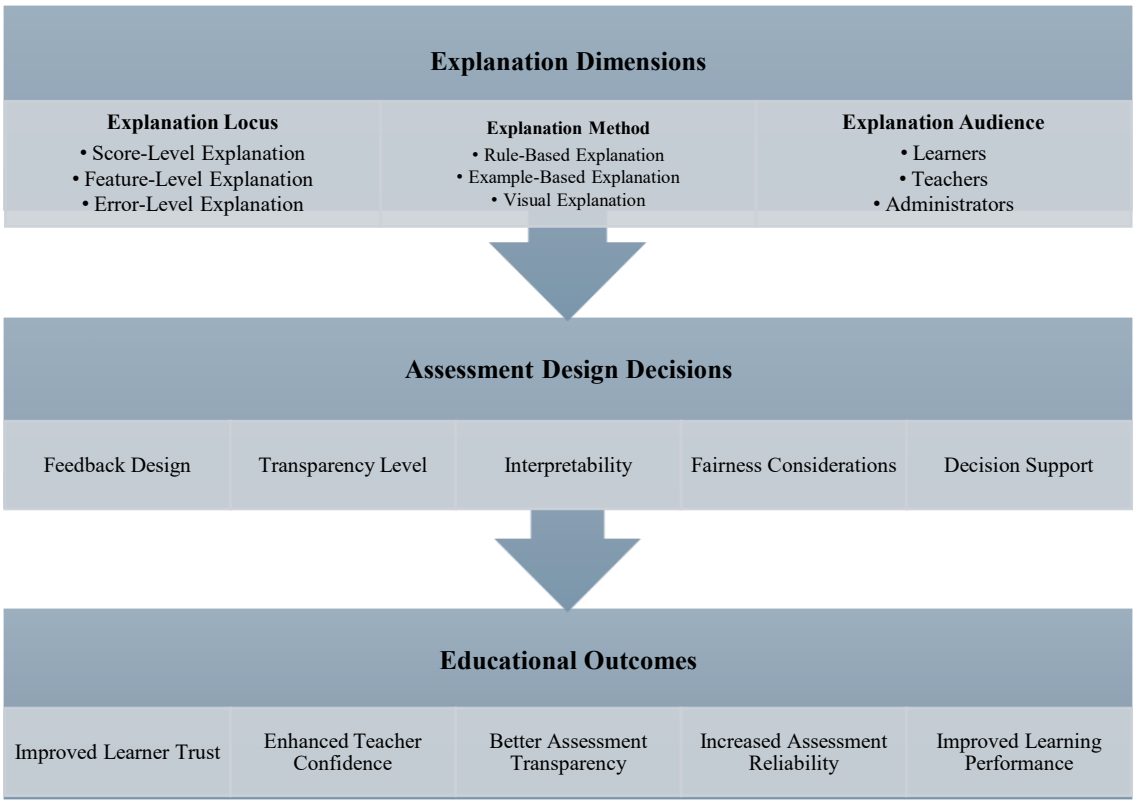

Figure 1: Framework Linking Explanation Dimensions to Assessment Design Decisions

ISSN: 2278-4012, Volume:12, Issue:2, June 2023

Comparative Analysis of Explainability Approaches

Intrinsic and Extrinsic Interpretability

Intrinsic interpretability models provide transparency as a by-product of their inherent scoring logic given that the process of making predictions happens in the model's architecture. However, their downside is that sometimes have difficulties capturing full-set and nuanced language intricacies. Extrinsic interpretability methods applied to deep learning can address this problem successfully but run a greater risk of presenting explanations that would be a mere approximation rather than a true reflection of the model's internal workings.

Fidelity and Accessibility

Another area of disparity in the field of explainability is the oppositional relationship that can sometimes exist between fidelity and accessibility of a specific method. Fidelity refers to the degree of closeness of the explanation, and accessibility denotes how well it can be understood and employed by a non-specialist. While amongst the methods that produce highly faithful and precise outputs, e.g., raw attention scores and feature attributions there are none accessible to wider audiences, the methods that provide accessible results, e.g., knowledge representation methods, may sacrifice some fidelity.

Comparison Across Assessment Modalities

Table 2: Comparative Strengths and Limitations of Explainability Methods

Method	Strength	Limitation
Feature-Based Interpretable Models	Transparent; Easy to Audit	Limited Capacity for Nuanced or Contextual Patterns
Attention-Based Explanation	Fine-Grained; Model-Native	Requires Technical Literacy to Interpret Correctly
Knowledge-Graph Explanation	Intuitive; Semantically Structured	Requires Extensive Manual Knowledge Encoding
Post-Hoc Perturbation Methods	Model-Agnostic; Flexible	Explanation Fidelity is not Guaranteed

Table 3: Explanation Needs by Stakeholder Audience

Audience	Explanation Goal	Preferred Format
Test Developers	Validate Model Reasoning and Diagnose Errors	Feature Attribution, Statistical Diagnostics
Teachers	Interpret Class-Level Patterns; Calibrate Trust	Aggregated Feature Summaries, Exemplars
Learners	Understand and act on Individual Feedback	Concrete Linguistic Examples, Pedagogical Language

According to table 2, the existing methodologies solve only parts of the problem of interpretability, but no method addresses all of the three aspects of interpretability, namely accuracy, fidelity, and accessibility. Table 3 expands this comparison to the stakeholder dimension mentioned in section 3.3, demonstrating that experts, users, and students require different formats of explanation while the underlying model and score remain unchanged. Therefore, both tables demonstrate that a single type of explanation is hardly going to serve the entire audience of the assessment system, and the best strategy

ISSN: 2278-4012, Volume:12, Issue:2, June 2023

would be for the assessment system to generate one explanation that will be presented in various formats depending on the audience.

Discussion

Trust and Learner Engagement

Explainability primarily functions as a means of establishing or maintaining trust between a machine and its human users. In language assessment, trust works at two different levels, including trust in the fairness and accuracy of the score itself, and trust in the value of any accompanying feedback for enhancing one's future performance. When learners perceive automated feedback to be arbitrary, their engagement and willingness to use it are likely to be greatly reduced which undermines the educational value of the system whether the score is accurate or not. An explanation that just repeats back the score but using other words but fails to point to some particular aspects of the learner's response will not possibly help to establish any meaningful trust or support learning.

Pedagogical Implications for Language Teaching

Explanations that identify some concrete linguistic patterns, for example, lack of subordination, limited range of vocabulary, or failure to achieve referential cohesion are going to be perceived as more trustworthy than generic commentary. In order to be able to implement automated language assessment in the classroom, teachers need adequate knowledge of how the system works, so as to determine cases where the score might not be relevant or feedback might be misleading for some particular learners.

Limitations and Future Directions

Several problems still remain unsolved. There is a permanent conflict between predictability and interpretability, and although some methodologies manage to achieve both at the same time, that does not happen automatically for all assessment tasks. Validity is among the main issues, because the explanation can only be useful if it shows what happens in the mind of the model. Most studies so far have dealt only with written assessment ignoring speech assessment, where the relevant factors like prosody or fluency are harder to detect than lexical or grammatical features in the text. At the same time, educational assessment lacks any generalized criteria for assessing the quality of the explanation making it difficult to compare different systems.

Conclusion

The current study aims to analyze the connections between the explainable artificial intelligence and the automated language assessment process, showing that explainability must be viewed as an integral part of the process of design rather than simply an optional characteristic. The elaborated classification, consisting of three main components, represents an effective way to place every variant of explainability into the one analytic framework; at the same time, the report suggests that there is no existing approach that offers a balanced combination of transparency, fidelity, and easy understanding for all stakeholders involved. Additionally, it is demonstrated that quality of explanation plays a role in terms of building relationships based on trust and indicating pedagogical importance of explanation: specific types of feedback based on interpretation of language features are likely to be taken more seriously than regular, non-specific types of feedback. Meanwhile, several obstacles must be tackled, such as the necessity to investigate the core elements of accuracy and reliability of explanation, to expand the use of explainable methods in the field of speech and multimodal testing, and to create commonly agreed standards of

explanation quality assessment. In this regard, future studies analyzing how the development of explainable artificial intelligence will affect the processes of explanation comprehension and implication by students and teachers will be of utmost importance for creating effective automated language assessment systems contributing to reliable, transparent, and pedagogical explanation practice.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.https://doi.org/10.1109/ACCESS.2018.2870052

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bamdev, P., Grover, M. S., Singla, Y. K., Vafaee, P., Hama, M., & Shah, R. R. (2023). Automated speech scoring system under the lens: evaluating and interpreting the linguistic cues for language proficiency. *International Journal of Artificial Intelligence in Education*, *33*(1), 119-154. https://doi.org/10.1007/s40593-022-00291-5

Baranyi, M., Nagy, M., & Molontay, R. (2020, October). Interpretable deep learning for university dropout prediction. In *Proceedings of the 21st annual conference on information technology education* (pp. 13-19). https://doi.org/10.1145/3368308.3415382

Clancey, W. J., & Hoffman, R. R. (2021). Methods and standards for research on explainable artificial intelligence: Lessons from intelligent tutoring systems. *Applied AI letters*, *2*(4), e53. https://doi.org/10.1002/ail2.53

Ferrara, S., & Qunbar, S. (2022). Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, *59*(3), 288-313. https://doi.org/10.1111/jedm.12333

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, *40*(2), 44-58. https://doi.org/10.1609/AIMAG.V40I2.2850

Ke, Z., & Ng, V. (2019, August). Automated Essay Scoring: A Survey of the State of the Art. In *IJCAI* (Vol. 19, pp. 6300-6308). https://doi.org/10.24963/ijcai.2019/879

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and education: artificial intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kumar, V., & Boulanger, D. (2020, October). Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education* (Vol. 5, p. 572367). Frontiers Media SA. https://doi.org/10.3389/feduc.2020.572367

McCarthy, K. S., Roscoe, R. D., Allen, L. K., Likens, A. D., & McNamara, D. S. (2022). Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision? *Assessing Writing*, *52*, 100608. https://doi.org/10.1016/j.asw.2022.100608

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Mohsen, M. A. (2022). Computer-mediated corrective feedback to improve L2 writing skills: A meta-analysis. *Journal of Educational Computing Research*, *60*(5), 1253-1276. https://doi.org/10.1177/07356331211064066

Poulton, A., & Eliens, S. (2021, September). Explaining transformer-based models for automatic short answer grading. In *Proceedings of the 5th international conference on digital technology in education* (pp. 110-116). https://doi.org/10.1145/3488466.3488479

Rajabi, E., & Etminani, K. (2024). Knowledge-graph-based explainable AI: A systematic review. *Journal of information science*, *50*(4), 1019-1029. https://doi.org/10.1177/01655515221112844

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: a systematic literature review. *Artificial intelligence review*, *55*(3), 2495-2527.

https://doi.org/10.1007/s10462-021-10068-2

Ramnarain-Seetohul, V., Bassoo, V., & Rosunally, Y. (2022). Similarity measures in automated essay scoring systems: A ten-year review. *Education and Information Technologies*, *27*(4), 5573-5604. https://doi.org/10.1007/s10639-021-10838-z

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816

Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668