Explanation as Evaluation: Using explanation quality to measure AI system performance

Paul Groth[1], Michael Cochez[2], Michel Dumontier[3], Fajar J. Ekaputra[4] and Monica Palmirani[5]

[1]University of Amsterdam, Amsterdam, Netherlands
[2]Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[3]Maastricht University, Maastricht, Netherlands
[4]WU Vienna, Vienna, Austria
[5]University of Bologna, Bologna, Italy

Abstract

Traditional benchmarking approaches for evaluating AI systems face limitations when applied to complex, real-world domains such as legal decision-making, medical research, and scientific discovery. These limitations include missing or incorrect ground truth data, data leakage, confirmation bias, and deployment mismatches that make systematic evaluation challenging. Here, we propose a novel approach *explanation as evaluation*, which uses the quality of AI-generated explanations as a proxy for overall system performance. Our process operates across three phases: (1) Design and Development, where explanation quality dimensions are selected; (2) Deployment, where explanations are generated alongside system outputs; and (3) Evaluation, where explanations are systematically assessed against the selected dimensions. We argue that this approach addresses key desiderata for AI evaluation: the ability to evaluate multifaceted outputs without ground truth, support for continuous evaluation, efficient use of human expertise, and ready application across diverse domains. We demonstrate the feasibility of our approach through a proof-of-concept evaluation in clinical trial outcome prediction. Using explanations for AI evaluation builds on the growing regulatory requirement for explainable AI systems, potentially enabling more robust and continuous assessment of AI performance in complex, dynamic domains.

Keywords

AI system evaluation, explainable AI, XAI

1. Introduction

Evaluation has driven massive progress in AI [1]. Systematic assessment provides insights into a model's strengths and limitations, revealing not only how accurately it performs a given task but also how well it generalizes, handles uncertainty, and upholds ethical standards. Robust evaluation frameworks aid in the assessment of competing approaches, foster transparency, and build trust among end-users and stakeholders. However, evaluation has primarily adopted a benchmarking approach [2] using standardized datasets with training and tests splits[1]. While effective for certain tasks, this approach has numerous shortcomings, which have been widely discussed in the literature [3, 4]. Generalized evaluation of more sophisticated use cases that move beyond well-defined tasks such as classification or text generation introduces additional complexity.

For some cases, systematic evaluation remains challenging owing to the lack of expert annotations as well as evolution in domain knowledge. As such, even highly trained experts may disagree on the soundness or completeness of a AI generated answer. Many AI systems also need to be evaluated before and during deployment, meaning that evaluation approaches need to be able to be continually run. To overcome these challenges, we need new approaches to evaluation.

We argue that the use of *explanations* for evaluation of AI systems is one such approach. Specifically, we posit that the *quality of explanations is an indicator of the performance of an AI system.* Assume

ESWC'26: Trust, Autonomy and Accountability in PKG-Based Agentic AI (TAAPAAI) Workshop on May 10, 2026 in Dubrovnik, Croatia

*Corresponding author.
p.groth@uva.nl (P. Groth); m.cochez@vu.nl (M. Cochez); michel.dumontier@maastrichtuniversity.nl (M. Dumontier); fajar.ekaputra@wu.ac.at (F. J. Ekaputra); monica.palmirani@unibo.it (M. Palmirani)

[1]See for example https://paperswithcode.com/sota

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-06-09

that AI system generates an output along with an explanation of that output. The explanation is then evaluated along multiple dimensions for its quality (e.g. is it rigorous, convincing, effective). We regard the quality of the explanations given for the particular outcomes as an important indicator of the quality of the AI system.

This is analogous to how we evaluate outcomes of complex human processes. For example, in science, legal cases, and healthcare reviewers consider the evidence base on which an outcome is derived, the methodology and reasoning used to come to the outcome, the credibility of the organizations involved, and so on.

Importantly our approach builds upon the significant work in two areas:

- *eXplainable AI (XAI)* provides the basis for both creating explanations as well as evaluating their quality [5]. XAI systems are widely available [6] and explainability is becoming a requirement for the deployment of such systems [7]. Additionally, there are frameworks that allow for the evaluation of the results produces by XAI systems [8]. We go further and say that these results are not just a measure of the explanation quality, but a measure of the AI system performance itself.

- *Argument quality assessment* [9, 10] that uses the social science literature to derive metrics for argument quality. This is especially useful in verbal explanations produced by Generative AI-based systems [11]. It also provides further metrics for understanding the quality.

Evaluating through explanations has a number of benefits. As long as an AI system is able to generate explanations, we can evaluate the quality of an output regardless of problem, task, or domain. We can focus part of the human effort on determining the best metrics for explanation quality rather than on specific domains.

The rest of this paper is organized as follows. First, we describe three use cases that exemplify where the evaluation of AI Systems is challenging. We then summarize common evaluation pitfalls from which we derive a set of desiderata for evaluation approaches. Next, we define Explanation-Based Evaluation and discuss existing literature with respect to our approach. Finally, we describe a proof of concept, discuss future work and conclude.

2. Complex Use Cases for AI Systems

Legal Predicting the success of a legal appeal requires considering evolving legislation and the judicial environment, making static benchmarks unreliable. For example, relying on old case-law can lead to an inaccurate prediction if the legislation has been recently amended when related case data is scarce. In addition, the legal domain puts further requirements on the AI system; it must have the ability to produce explanations to judges, enabling informed decision-making, while maintaining neutrality, impartiality, and fairness. This explanation does not themselves provide a justification, but is rather a necessary step to find a legal basis, justifying the final decision made by the human judge. To effectively utilize AI, judges must trust the system's quality and understand the reasoning (the chain of thought) behind its outputs. Consequently, argumentation theory, grounded in logic, is frequently employed to evaluate both the quality of AI systems and their outputs within the legal domain [12].

For a judge, it is important to distinguish between prediction, explanation, and justification. The prediction can help the judge to find and contrast with cases which are not just lexically similar, but also similar in terms of arguments, facts, and their legal basis [13]. Argumentation theory can help evaluate the output of neuro-symbolic AI system, and can provide the judge with relevant elements for their decision-making process, including the creation of autonomous opinions, justifying the opinion, and providing arguments for the decision.

Medicine: Fewer than 1 in 10 clinical trials ultimately succeed in bringing a molecular therapeutic to market [14]. Clinical trials often fail due to reasons of safety and efficacy owing complex factors such as incomplete understanding of disease and drug mechanisms, heterogeneous patient populations, and suboptimal trial design. AI systems hold the promise of unlocking new insights from vast datasets to

guide fundamental health research as well as in trial planning and execution, ultimately reducing failure rates. However, harnessing this potential requires more than just accurate predictions—it requires trustworthy explanations that make the decision-making process transparent and understandable to clinicians, researchers, and regulators. Well structured, evidence-based explanations will help stakeholders pinpoint where AI-driven insights come from, how they align with clinical reasoning, and where potential biases or gaps may lie, such as a lack of genetic evidence [15]. Reasoned explanations will allow innovators to properly assess risk, and potentially refocus their efforts on better developed or more promising directions. By ensuring AI models are not only powerful but also interpretable, the medical community can gain the confidence and clarity needed to design more effective trials, refine interventions in real time, and maximize patient benefit.

Scientific research: A feature of scientific research is that it always ventures into unchartered territory. The most interesting questions are those for which the answer is not known yet. [16] stated that the search space for scientific knowledge is potentially infinite or unboundable. Hence, if we have an AI system which aids in the process of making new findings, it will eventually draw from this space. Now, given this space is unbounded, it is not generally possible to define an absolute benchmark or ground truth for this AI system. An example can be found in climate research where we might want to predict sea surface temperature. [17] found that a large fraction of the differences in this measure for 2023/24 could not be predicted from linear extrapolation from the past four decades. One could conceive of a more expressive predictive model, but the only way to make sure its predictions are correct is to do an experiment which is not only impractical, but likely also not ethical. Alternatively, one could attempt running a simulation to evaluate, but in practice this cannot be as granular as the real world system. So, even if a system predicts well for all observations we have until today, there is no guarantee that it will perform well in the future.

3. Common Evaluation Pitfalls

Evaluating AI models is critical for understanding their performance, limitations, and suitability for real-world applications. However, several challenges or pitfalls can arise in the current ways we perform these evaluations. We discuss only some here.

Missing or Incorrect Ground Truth Data Ground truth data is the data used to train and evaluate an AI system. Missing or incorrect elements in ground truth data may have a substantive impact on the ability of the AI system to learn patterns that generalize to real-world situations and to accurately report their anticipated performance [18] . Building a good quality ground truth dataset is challenged by the high cost of human annotation [19, 20], the varying quality of annotated data [21], the challenging nature of the annotation process itself [22] and the ability to broadly cover representative data [23].

Data Leakage Data leakage occurs when during the model training information is used that would not be expected to be available at prediction time, causing the predictive scores (metrics) to overestimate the model's utility. One example is the incorrect split of train-test (e.g., time steps in a time-series dataset are randomly split so that future data is available during training). The issue of data leakage is commonly known in machine learning and practitioners take measures to avoid it. However, this becomes more difficult for complex systems with multiple sources of information [24].

Confirmation Bias AI system designers aim to evaluate their systems well. However, they might be guided by confirmation bias. For example, imagine a designer evaluates their new system with a dataset and obtains excellent results on the commonly used dataset and metrics. There is incentives to stop and either publish the work or deploy the system. However, a superficial evaluation may not uncover systematic issues in the model and result. First, the results might be due to a bug in the system, due to luck (e.g., we would only obtain these with a very specific seed) or due to data leakage. Further issues could occur when comparing the outcomes to existing research results. Often there is no capacity to re-evaluate all other systems for comparison, and existing results are reused. Unfortunately, this has caused misleading comparisons where what is compared is not the system itself, but rather an environmental factor, similar to a hidden confounder. An example can be found in the link prediction

literature where what was in fact compared were the the optimization tricks used, rather than the systems [25]. The above are examples of confirmation bias but there are a wide variety of other biases that emerge in the testing and validation process [26].

Deployment mismatch A large majority of ML approaches are evaluated with the assumption that the distribution of instances that will be seen at inference time will be similar to the distribution that was used to train the system. This assumption is valid in a rather limited number of real-world settings. In practice, the real-world deployment of AI systems is much more involved and the "sterile" evaluation results found in research papers hardly ever apply [20]. To illustrate, methods to predict new or missing links in graphs are evaluated with benchmarks containing complete knowledge [27], which is never the intended application of such methods. This is an example of the focus on the internal validity of results rather than their external validity [3].

4. Desiderata for Evaluating AI Systems

We now formulate desiderata for evaluation approaches that can be used in the kinds of AI systems discussed above and would overcome common pitfalls.

To cope with the complex outputs produced by AI systems, the approach should ① *be able to evaluate such multifaceted and complex outputs.* Given the difficulty in creating correct ground truth data the approach should be able to be applied when ② *no ground truth is available.*

Given the importance of evaluating in deployed environments, the approach should be able to ③ *run in a continuous manner* and to ④ *cope with changes in outputs.*

The domains in the use cases above are complex. Hence, expertise is often limited therefore the evaluation approach should ⑤ *efficiently make use of human effort.* This also entails that the approach should be able to be ⑥ *readily applied to new problems, tasks and domains with a minimal amount of effort.* AI systems produce outputs that are often specific to a requester (e.g. a user, or organization). Hence, evaluation approaches should be able to ⑦ *cope with variation in the tailored outputs.*

We now define explanation based evaluation and describe how it addresses these desiderata.

5. Explanation-Based Evaluation

As AI systems are integrated into decision-making in domains such as healthcare, finance, and law, explanations become crucial. They act as the bridge between the system's internal reasoning and human trust; they ensure that users can understand, validate, and, if necessary, challenge the system's outputs. Poor-quality explanations—those that are unclear, logically flawed, or misleading—can erode trust, perpetuate biases, and lead to harmful outcomes, especially in high-stakes contexts. Moreover, explanations can help identify whether an AI system adheres to ethical and legal standards, aligns with domain-specific knowledge, and is robust against adversarial conditions. A systematic approach to explanation quality ensures that AI systems remain interpretable, transparent, and accountable, fostering responsible adoption and use.

Given these benefits, AI-systems increasingly provide various forms of explanation. We aim to use these explanations not just for these benefits but to evaluate the performance of the AI system itself. In other words, the quality of explanations is a proxy for the performance of an AI system.

Definition 1 (AI System performance). *Given a set of tasks, and corresponding outputs and their explanations created by an AI system.* AI System performance *is the aggregation of the quality of the explanations.*

5.1. Process Steps

We define a set of process steps to operationalize our approach, organized into three steps across lifecycle phases of AI systems: Design and Development, Deployment, and Evaluation. Each phase and its associated steps are depicted in Figure 1 and briefly described in the following.

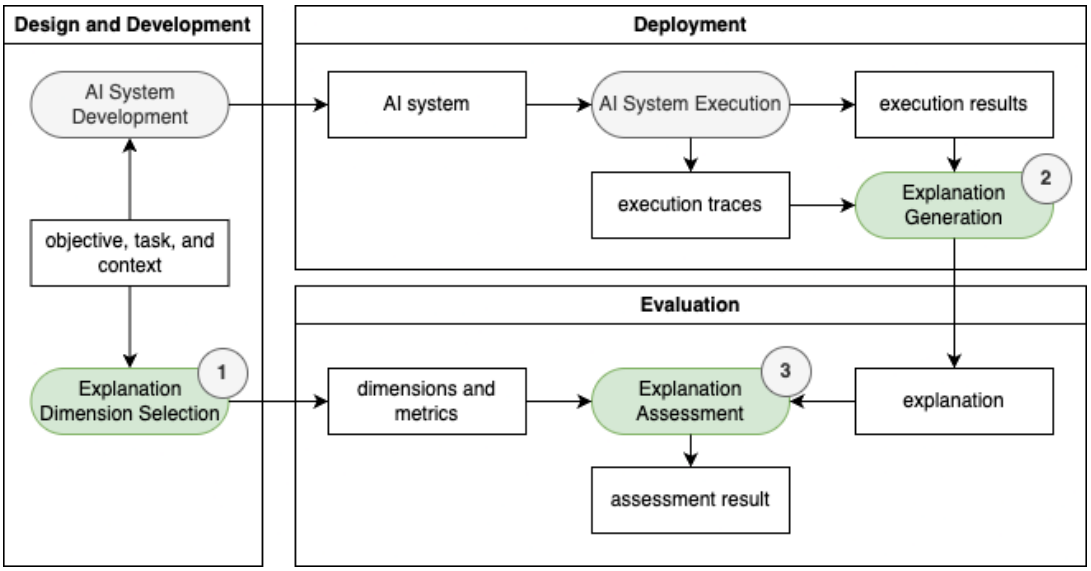

Figure 1: The Process Steps for *"Explanation as Evaluation"* Approach

Phase: Design and Development. We initiated the evaluation process from the early design and development phase, where the objectives, tasks, and contexts of the AI systems needs to be defined. These information will become an input for both AI system development, and the explanation quality dimension selection process.

- *Step 1: Explanation Quality Dimension Selection.* This step defines the most relevant dimensions and metrics for the evaluation. We summarized some potential quality dimensions from logical and rhetorical point of views in Table 1.

Phase: Deployment. This process phase facilitate the operationalization of AI systems, where users can execute commands on AI systems and retrieve execution results. The system can also produce execution traces, providing additional context to the execution results. Both the execution results and traces will become an input for the explanation generation step.

- *Step 2: Explanation Generation.* This step will generate the explanation of the execution results based on AI system execution results and process traces. The explanation can be generated through various methods, ranging from symbolic (e.g., decision-tree), sub-symbolic (e.g., LLMs), neurosymbolic (e.g., RAG), to heuristics approaches.

Phase: Evaluation. The goal of this process phase is to determine whether the AI system meets its intended objectives. In our approach, we focus on a specific evaluation method of explanation assessment.

- *Step 3: Explanation Assessment* assesses the generated explanation from Step 4 in accordance to the selected dimensions and metrics. Similar to the explanation generation method, the evaluation assessment method can vary from manual assessment to neurosymbolic AI.

5.2. Evaluating the quality of a explanation

Explainable AI (XAI) methods aim to make the decisions made by AI systems more transparent and understandable for people. xAI outputs can take a variety of forms depending on the modal output. For instance, an explanation pertaining to image classification may highlight pixels that are relevant to the decision. A growing number of efforts have emerged to systematically assess xAI methods on a variety of problems so to help end-users select appropriate methods for particular tasks or problems

Logical Dimensions		Rhetorical Dimensions	
Dimension	**Definition**	**Dimension**	**Definition**
Validity	Does the argument follow from its premises?	Clarity	Is the argument expressed clearly and unambiguously?
Soundness	Are the premises of the argument true, or plausible?	Accessibility	Are the terms and concepts sufficiently well defined for the target reader to understand them?
Consistency	Are there any internal contradictions within the argument?	Persuasiveness	Does the argument effectively appeal to the intended audience?
Coherence	Does the argument flow logically with good transitions between points?	Free of Fallacies	Includes strawman arguments, ad hominem attacks, false dichotomies, appeal to authority, slippery slopes, circular reasoning, emotional appeals.
Conciseness	Does the argument contain redundant premises or evidence?	Representative	Does the argument rely on examples or data that are representative of the broader context, or are they cherry-picked?
Completeness	Does the argument address all aspects of the issue; does it omit important information?	Use of Evidence	Is evidence credible, relevant, representative, and sufficient?
Relevance	Are all the premises and pieces of evidence pertinent to the conclusion, or are there irrelevant details?	Use of Counterfactuals	Does the argument acknowledge and address potential counterarguments or alternative viewpoints?
Timeliness	Does the argument use information that is accurate for the period?	Accuracy	Does the argument present information truthfully, without misrepresentation or distortion?
Focused	Does the argument stay on topic, or does it go off topic?	Fairness	Is the argument neutral, impartial, and just?
Contextualized	Does the argument consider specific contextual considerations? For instance, spatial or geopolitical considerations.	Acceptability	Whether the argument is satisfactory and able to be agreed to or approved of by the target recipient
		Novelty and Creativity	Does the argument present new perspectives or solutions? Is it original, or does it rehash well-known ideas?

Table 1
Examples of dimensions for assessing argument quality sourced from [28].

[29, 8]. However, these xAI evaluation frameworks do not purport that the systematic assessment of the quality of the explanations of an AI system can be used as a measure of system performance.

Explanations should align with established research on argument quality along logical and rhetorical dimensions are listed in Table 1 and cover aspects relating to cogency (e.g. logical soundness, relevance, and sufficiency of evidence), reasonableness (e.g. consideration of alternative viewpoints and global accessibility), and effectiveness (e.g. Persuasiveness, clarity, and credibility) [28].

6. Proof-of-Concept

We exemplify the approach in an anecdotal use case on Clinical Trials (cf. the section on use cases) to demonstrate the feasibility of our approach. In this use case, we assume the existence of an AI system that can predict the result of clinical trials based on trial descriptions to help (a) *decision makers* to issue the permit for such a trial, and (b) *patients* to decide whether or not enrolling on such trials.

Dimension	Default perspective		Lay-users	
	GPT	Gemini	GPT	Gemini
Soundness	4	4	4	3
Coherence	5	4	4	4
Clarity	3	3	2	2
Persuasiveness	3	3	2	3

Table 2
LLMs' assessment results for "sound and coherent", but "unclear and unpersuasive" explanations on a 5 point Likert scale (5 = very good) from two different personas (Default and and lay-users.)

Bringing the use case to life, we are using a public real-world clinical trial descriptions[2] as the input data to the AI system. We selected an example case of a clinical trial on the effectiveness of an experimental drug INM-176 for patients with Alzheimer-type dementia in comparison to the standard, control drug Donezepil[3].

In our proof-of-concept (PoC), we utilized three different LLMs accessed through their web interfaces: a) OpenAI `gpt-4o-mini` (GPT), b) Anthropic `claude-3.5-haiku` (Claude), and c) Google `gemini-2.0-flash` (Gemini). These LLMs are used for both explanation generation (Step 4) and explanation assessment (Step 5) steps.

The goals of our PoC are two-folds: (A) to evaluate whether the LLMs-based assessment is able to capture the specified dimensions intended in the generated explanation, and (B) to evaluate the effect of *context*, specifically different personas in LLM-based explanation assessment process, representing different types of users. We briefly describe each step in our evaluation setup as the following:

1) Explanation Dimension Selection. We selected two logical (*"soundness" and "coherence"*) and two rhetorical (*"clarity" and "persuasiveness"*) dimensions for this feasibility evaluation. Furthermore, a Likert-scale (1-very poor, 5-very good) and text justification are used as evaluation metrics for assessment.

2) Explanation Generation. We conducted this process step through asking LLMs to generate explanations for our input dataset (clinical trial results) with various quality wrt. chosen dimensions. Concretely, we generated two sets of explanation on a specific statement of *"why INM-176 is more effective compared to the standard treatment"* with the following characteristics with regards to the selected dimensions: (EXP-1) "Sound and coherent", but "unclear and non-persuasive" explanations[4], and (EXP-2) "Clear and persuasive", but "unsound and incoherent" explanations[5].

3) Assessment of Explanation Dimension. We conducted quality assessment through prompting several LLMs with different personas (i.e., default vs lay-users). This follows the arguments from recent work showing that LLMs can make for quality argument assessors [30].

For assessing explanations, we use LLMs that are not used for explanation generation to ensure the neutrality of the assessment process. The assessment results are shown in Table 2 and Table 3. Detailed prompts and assessment results for Table 2 (i.e., GPT[6] and Gemini[7]) and Table 3 (i.e., Claude[8] and Gemini[9]) are available online.

Assessment Result Analysis. We analyse the evaluation assessment results with regards to the two goals we defined earlier:

(A) to evaluate whether the LLMs-based explanation assessment is able to capture the specified dimensions

[2] https://clinicaltrials.gov
[3] https://clinicaltrials.gov/study/NCT01245530
[4] Prompts and results: https://short.wu.ac.at/y8ab
[5] Prompts and results: https://short.wu.ac.at/6hp7
[6] https://short.wu.ac.at/s7bn
[7] https://g.co/gemini/share/91152728d968
[8] https://short.wu.ac.at/fghu
[9] https://g.co/gemini/share/20a81777b1d7

Dimension	Default perspective		Lay-users	
	GPT	Claude	GPT	Claude
Soundness	1	1	1	3
Coherence	2	2	2	4
Clarity	3	3	3	5
Persuasiveness	1	1	1	4

Table 3
LLMs' assessment results for "clear and persuasive", but "unsound and incoherent" explanations

intended in the generated explanation. The results demonstrate the ability of LLMs to correctly assess the intended quality dimensions in almost all cases (especially in Table 2), with exceptions related to the given personas, which lead to the next point.

(B) to evaluate the effect of context, specifically different personas in LLM-based explanation assessment process, representing different types of users. While generally the results are stable, particular LLMs are more diverge in terms adapting personas for assessing the explanations. Table 3 shows an extreme case of assessment with different user personas with Claude LLM, where it specifically mention the lack of medical knowledge as the reason of high-assessment for all dimensions, especially on soundness and persuasiveness.

7. Discussion & Conclusion

In this work, we have argued that the quality explanations can be a powerful proxy for the quality of an AI system's performance.

This contention presupposes that AI systems will be able to provide such explanations. This is a good question [31]. Here, we point to the fact that explanations are becoming a critical requirement on AI systems according to policy, regulatory and standard bodies. For example, European legislation (AI Act art. 13, 14, 86[10]) refers to explainability. In the USA, NIST includes explainability in the main principles for designing a good AI system. See, the Four Principles of Explainable Artificial Intelligence, 2021[11]). In the UK, guidelines on AI regulation include explainability as a main pillar of AI system design[12]. Likewise, UK guidelines on AI and cybersecurity also focus on explanation.[13].

At the international level, standardization bodies (e.g. ISO/IEC/ JTC1/SC42 Working Groups on AI Standardisation, CEN-CENELEC, ETSI) have been working on AI standards that include explanation ([14]). Furthermore, the EU AI ACT delineates a roadmap for approving these standards with the participation of all the stakeholders (Art. 40 (Harmonised Standards and Standardisation Deliverables)). Thus, explaining the results of AI systems is seen by these bodies as a fundamental principle to implement transparency, trustworthiness, fairness, and accountability in the use of AI by end users.

An interesting side effect of using explanations as the source of evaluation is that the many benefits of using XAI also are produced, whether that is improved debugging, reproducing results, understanding the provenance of training data, improving the trust in the system, or helping justify a final decision of the system[15].

There is a wide variety of future work to be considered in using explanations as a source of evaluation. First, one could tailor the evaluation dimensions according to the domain and use-case, the needs of the applications, the deepness of explanation (e.g., judge, teacher, citizen). In the legal domain, for example,

[10] http://data.europa.eu/eli/reg/2024/1689/oj
[11] https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf
[12] https://ico.org.uk/media/about-the-ico/consultation-responses/4029424/regulating-ai-the-icos-strategic-approach.pdf
[13] https://www.gov.uk/government/news/world-leading-ai-cyber-security-standard-to-protect-digital-economy-and-deliver-plan-for-change
[14] https://www.iso.org/obp/ui/en/#iso:std:iso-iec:ts:25058:ed-1:v1:en
[15] In case of an automated decision system, art. 22 GDPR Reg. 2016/679 http://data.europa.eu/eli/reg/2016/679/oj

"Conciseness" or the "Persuasiveness" are less important respect the Validity, Soundness, Timeliness, Contextualized, Use of counterfactual, Fairness.[6]

Second, developing systems for continues evaluation based on explanations is critical for implementation in practice. Lastly, there is significant amount of work in verifying whether this approach to evaluation corresponds to what stakeholders view as quality.

Traditional benchmarking approaches to AI evaluation have significant limitations especially in complex and dynamic domain. There is often a lack of ground truth data. It is challenging to evaluate system output in a continuous manner. Furthermore, complex and multifaceted outputs are often difficult to evaluate even for experts. In this paper, we have argued that by evaluating AI systems through the explanations that they generate these challenges can be overcome. Explanations can be generated continuously and there is a rich literature, for example on argument quality, that can be used to evaluate explanations in an independent manner. We hope that this approach to AI system evaluation will help to continue to drive forward progress in the field based on systematic assessment.

Acknowledgments

This work originated from participation in Dagstuhl seminar 25051 - Trust and Accountability in Knowledge Graph-Based AI for Self Determination.

For this work, Michael Cochez and Paul Groth were partially funded by the Elsevier Discovery Lab. The work in part is based on work in COST Action CA24121 - Knowledge Graphs in the Era of Large Language Models (KGELL), supported by COST (European Cooperation in Science and Technology, https://www.cost.eu). This work is partially supported by the EU's Horizon Europe programme, in the ENEXA project (grant Agreement no. 101070305). This work was supported by the Austrian Science Fund (FWF) Bilateral AI [10.55776/COE12] and the Austrian Research Promotion Agency (FFG) FAIR-AI [FO999904624]. It was also supported by European Commission funds within ERC HyperModeLex Grant agreement ID: 101055185.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Claude for formatting assistance, literature review and summarization as well as part of the PoC. After using these tools, the authors reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Renkhoff, K. Feng, M. Meier-Doernberg, A. Velasquez, H. H. Song, A survey on verification and validation, testing and evaluations of neurosymbolic artificial intelligence, IEEE Transactions on Artificial Intelligence 5 (2024) 3765–3779. doi:10.1109/TAI.2024.3351798.

[2] Y.-T. Lin, Y.-N. Chen, LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models, in: Y.-N. Chen, A. Rastogi (Eds.), Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 47–58. URL: https://aclanthology.org/2023.nlp4convai-1.5/. doi:10.18653/v1/2023.nlp4convai-1.5.

[3] T. I. Liao, R. Taori, I. D. Raji, L. Schmidt, Are we learning yet? A meta-review of evaluation failures across machine learning, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2024.

[4] B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, V. Prabhakaran, Evaluation gaps in machine learning practice, 2022. URL: http://arxiv.org/abs/2205.05256. doi:10.48550/arXiv.2205.05256. arXiv:2205.05256 [cs].

[5] Q. V. Liao, Y. Zhang, R. Luss, F. Doshi-Velez, A. Dhurandhar, Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 10 (2022) 147–159. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/21995. doi:10.1609/hcomp.v10i1.21995.

[6] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Trans. Interact. Intell. Syst. 11 (2021). URL: https://doi.org/10.1145/3387166. doi:10.1145/3387166.

[7] N. Balasubramaniam, M. Kauppinen, K. Hiekkanen, S. Kujala, Transparency and explainability of ai systems: ethical guidelines in practice, in: International working conference on requirements engineering: foundation for software quality, Springer, 2022, pp. 3–18.

[8] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, Openxai: Towards a transparent evaluation of model explanations, in: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022. URL: https://openreview.net/forum?id=MU2495w47rz.

[9] G. Lapesa, E. M. Vecchi, S. Villata, H. Wachsmuth, Mining, assessing, and improving arguments in NLP and the social sciences, in: R. Klinger, N. Okazaki, N. Calzolari, M.-Y. Kan (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries, ELRA and ICCL, Torino, Italia, 2024, pp. 26–32. URL: https://aclanthology.org/2024.lrec-tutorials.5/.

[10] A. Toledo, S. Gretz, E. Cohen-Karlik, R. Friedman, E. Venezian, D. Lahav, M. Jacovi, R. Aharonov, N. Slonim, Automatic argument quality assessment - new datasets and methods, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5625–5635. URL: https://aclanthology.org/D19-1564/. doi:10.18653/v1/D19-1564.

[11] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, A survey on xai and natural language explanations, Information Processing & Management 60 (2023) 103111.

[12] J. Collenette, K. Atkinson, T. Bench-Capon, Explainable ai tools for legal reasoning about cases: A study on the european court of human rights, Artificial Intelligence 317 (2023) 103861. URL: https://www.sciencedirect.com/science/article/pii/S0004370223000073. doi:https://doi.org/10.1016/j.artint.2023.103861.

[13] F. Bex, H. Prakken, Can predictive justice improve the predictability and consistency of judicial decision-making?, in: E. Schweighofer (Ed.), Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021, volume 346 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2021, pp. 207–214. URL: https://doi.org/10.3233/FAIA210338. doi:10.3233/FAIA210338.

[14] H. Dowden, J. Munro, Trends in clinical success rates and therapeutic focus, Nature Reviews Drug Discovery 18 (2019) 495–496. doi:https://doi.org/10.1038/.

[15] I. D. Olesya Razuvayevskaya, Irene Lopez, D. Ochoa, Genetic factors associated with reasons for clinical trial stoppage, Nature Genetics 56 (2024) 1862–1867. doi:https://doi.org/10.1038/s41588-024-01854-z.

[16] H. Kitano, Nobel Turing Challenge: creating the engine for scientific discovery, npj Systems Biology and Applications 7 (2021) 29. URL: https://doi.org/10.1038/s41540-021-00189-3. doi:10.1038/s41540-021-00189-3.

[17] C. J. Merchant, R. P. Allan, O. Embury, Quantifying the acceleration of multidecadal global sea surface warming driven by earth's energy imbalance, Environmental Research Letters 20 (2025) 024037. URL: https://dx.doi.org/10.1088/1748-9326/adaa8a. doi:10.1088/1748-9326/adaa8a.

[18] M. Mazumder, C. R. Banbury, X. Yao, B. Karlas, W. G. Rojas, S. F. Diamos, G. Diamos, L. He, A. Parrish, H. R. Kirk, J. Quaye, C. Rastogi, D. Kiela, D. Jurado, D. Kanter, R. Mosquera, W. Cukierski, J. Ciro, L. Aroyo, B. Acun, L. Chen, M. Raje, M. Bartolo, E. S. Eyuboglu, A. Ghorbani, E. D. Goodman, A. Howard, O. Inel, T. Kane, C. R. Kirkpatrick, D. Sculley, T. Kuo, J. W. Mueller, T. Thrush, J. Vanschoren, M. Warren, A. Williams, S. Yeung, N. Ardalani, P. K. Paritosh, C. Zhang, J. Y. Zou,

C. Wu, C. Coleman, A. Y. Ng, P. Mattson, V. J. Reddi, DataPerf: Benchmarks for data-centric AI development, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/112db88215e25b3ae2750e9eefcded94-Abstract-Datasets_and_Benchmarks.html.

[19] J. Hu, R. Kashi, D. Lopresti, G. Nagy, G. Wilfong, Why table ground-truthing is hard, in: Proceedings of Sixth International Conference on Document Analysis and Recognition, IEEE, 2001, pp. 129–133.

[20] A. Paleyes, R.-G. Urma, N. D. Lawrence, Challenges in deploying machine learning: A survey of case studies, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3533378. doi:10.1145/3533378.

[21] Z. Cao, E. Chen, Y. Huang, S. Shen, Z. Huang, Learning from crowds with annotation reliability, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 2103–2107. URL: https://doi.org/10.1145/3539618.3592007. doi:10.1145/3539618.3592007.

[22] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10671–10682. URL: https://aclanthology.org/2022.emnlp-main.731/. doi:10.18653/v1/2022.emnlp-main.731.

[23] J.-C. Klie, R. E. d. Castilho, I. Gurevych, Analyzing dataset annotation quality management in the wild, Computational Linguistics 50 (2024) 817–866. URL: https://doi.org/10.1162/coli_a_00516. doi:10.1162/coli_a_00516. arXiv:https://direct.mit.edu/coli/article-pdf/50/3/817/2470929/coli_a_00516.pdf.

[24] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, D. Scheinost, Data leakage inflates prediction performance in connectome-based machine learning models, Nature Communications 15 (2024). URL: http://dx.doi.org/10.1038/s41467-024-46150-w. doi:10.1038/s41467-024-46150-w.

[25] D. Ruffinelli, S. Broscheit, R. Gemulla, You CAN teach an old dog new tricks! on training knowledge graph embeddings, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=BkxSmlBFvr.

[26] R. Srinivasan, A. Chander, Biases in ai systems, Communications of the ACM 64 (2021) 44–49. URL: http://dx.doi.org/10.1145/3464903. doi:10.1145/3464903.

[27] Y. Zhou, X. Chen, B. He, Z. Ye, L. Sun, Re-thinking knowledge graph completion evaluation from an information retrieval perspective, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2022, pp. 916–926. URL: https://dl.acm.org/doi/10.1145/3477495.3532052. doi:10.1145/3477495.3532052.

[28] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, B. Stein, Computational argumentation quality assessment in natural language, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 176–187. URL: https://aclanthology.org/E17-1017/.

[29] N. A. Sharma, R. R. Chand, Z. Buksh, A. B. M. S. Ali, A. Hanif, A. Beheshti, Explainable ai frameworks: Navigating the present challenges and unveiling innovative applications, Algorithms 17 (2024). URL: https://www.mdpi.com/1999-4893/17/6/227. doi:10.3390/a17060227.

[30] N. Mirzakhmedova, M. Gohsen, C. H. Chang, B. Stein, Are Large Language Models Reliable Argument Quality Annotators?, Springer Nature Switzerland, 2024, p. 129–146. URL: http://dx.doi.org/10.1007/978-3-031-63536-6_8. doi:10.1007/978-3-031-63536-6_8.

[31] Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, V. Mikulik, S. R. Bowman, J. Leike, J. Kaplan, E. Perez, Reasoning models don't always say what they think, 2025. URL: https://arxiv.org/abs/2505.05410. doi:10.48550/ARXIV.2505.05410.