Faithfulness All the Way Down: A Wittgensteinian Diagnosis of the XAI Evaluation Regress

Nicolò Cristante[1]

[1]Ca' Foscari University of Venice, Dept. of Philosophy and Cultural Heritage, Dorsoduro 3484/D, Venezia, 30123, Italy

Abstract

Faithfulness metrics are aimed at resolving the disagreements among the explanations produced by XAI methods. However, they tend to reproduce those discrepancies at the next level: the metrics themselves diverge, requiring additional criteria for comparison. In this context, *faithfulness* denotes the extent to which an explanation accurately represents the model it purports to account for. Relying on Wittgenstein's rule-following considerations, as formulated in the *Philosophical Investigations*, this paper diagnoses the instability as a regress of interpretations. Each metric operationally reformulates the rule for the application of the predicate *faithful*, treating agreement with a selected behavioural profile as criterial. A meta-metric introduces a further interpretation without grounding the authority of the standard it applies. The difficulty is therefore normative rather than a case of ordinary empirical underdetermination. Even ground-truth benchmarks validate metrics solely against predetermined explanatory targets. Faithfulness should thus be maintained as an evidential constraint, not as an autonomous criterion of correctness. The relevant standards, following Wittgenstein's appeal to practice, must instead be articulated and tested within public forms of justification, challenge, and revision. Contestability names this requirement, while formal argumentation offers a framework for operationalising it.

Keywords

Explainable AI, faithfulness, evaluation, rule-following, Wittgenstein, contestability, argumentation

1. Introduction

Much of XAI treats explanation quality as a relation between an explanatory representation and the model it purports to explain. This relation is typically discussed in terms of faithfulness. Measuring this correspondence would allow competing explanations to be ranked without relying on visual plausibility or the confidence they elicit.

Yet the practice proves less stable than this picture suggests. Post-hoc methods regularly generate incompatible accounts of the same prediction, and the faithfulness metrics introduced to arbitrate among them likewise disagree. Practitioners consequently turn to ad hoc heuristics [1], whereas meta-evaluations report weak or unstable metric performance even under controlled conditions [2, 3].

This instability is commonly regarded as a symptom of methodological immaturity; hence more representative benchmarks, more appropriate perturbations, or cleaner ground truth are expected eventually to produce convergence. This paper investigates whether this phenomenon may instead be structural. My starting point is Wittgenstein's rule-following argument. According to the *Philosophical Investigations*, an interpretation cannot by itself determine how a rule is to be applied, since applying that interpretation may call for a further interpretation (PI §§198–202). Faithfulness evaluation reproduces this structure. The rule governs the application of the predicate *faithful*: a metric specifies what constitutes a correct application by selecting an aspect of model behaviour as criterial. Deletion, insertion, stability, or some analogous relation is therefore not read neutrally from the model as the meaning of *faithful*; it is an operational interpretation of what faithfulness demands. When rival metrics select different profiles, their disagreement concerns less the accuracy of competing measurements than which interpretation of *faithful* should govern.

BEWARE-26 (Bias, Risk, Explainability and the role of Logic and Logic Programming), 8 October 2026, Perugia, Italy, co-located with AIxIA 2026

nicolo.cristante@unive.it (N. Cristante)

0009-0002-0573-5324 (N. Cristante)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-17

CEUR-WS.org/Vol-4332/paper3.pdf

Settling this dispute with a meta-metric requires a further standard whose legitimacy must itself be established: this is the regress of interpretations in the form it assumes within XAI evaluation. This argument should not be understood as evaluative nihilism; its claim is more limited. Faithfulness remains a necessary constraint: an explanation that conflicts with robust evidence about the model is defective. What faithfulness cannot determine by itself is which, among the compatible accounts, should qualify as the correct explanation. I argue that this question falls within a justificatory practice: reasons must be articulated and remain open to challenge; if a challenge succeeds, they must be revised. The paper thereby extends Páez's pragmatic turn [4] but accords it a different status: once faithfulness is not self-grounding, the pragmatic shift is structurally required rather than merely advisable. Accordingly, the object of the argument is to be distinguished from both accounts that situate opacity in the model–world relation or in the phenomenon itself [5] and Wittgensteinian work on AI alignment [6]. It pertains to the grounding of explanation evaluation itself.

2. The Evaluation Practice and Its Anomaly

Consider first the conception of explanation presupposed by a substantial portion of the XAI literature, which I will call "the extraction picture": the model contains determinate explanatory structure, an XAI method reconstructs it, and a metric assesses the success of the reconstruction.

This picture is not imposed externally; it is evident in both the field's definitions of faithfulness and the design of its most influential post-hoc methods. Jacovi and Goldberg define an interpretation as *faithful* when it accurately depicts the model's reasoning process, while noting the absence of a consistent formal definition [7]. Nauta et al. similarly treat correctness as a dimension of explanation quality and survey quantitative tests intended to measure correspondence with the predictive model [8]. The same reconstructive assumption is embedded in LIME and SHAP, although the two implement it differently. LIME constructs an interpretable surrogate intended to locally approximate the model [9], while SHAP represents a prediction through an additive attribution model whose values are meant to account for the model's output [10]. In each case, explanatory content is recast in a simpler representation and assessed in terms of its relation to the original model. Here the distinction Lipton draws between transparency and post-hoc interpretation underscores an important limit: such a representation may track the model's output without disclosing the mechanism underlying its production [11].

The instability emerges first among the explanations themselves. Krishna et al. report conflicts in selected features, rankings, and contribution signs. Among the 14 survey respondents who reported encountering explanation disagreements, 86% described ad hoc heuristics or uncertainty about resolving them [1]. It then manifests among the metrics. Miró-Nicolau et al. find substantial deviations from the expected score even when the benchmark is a transparent decision tree [2]. For chain-of-thought evaluation, Gur-Arieh et al. construct ground-truth-labelled tasks, yet most prominent metrics remain near-chance, with different metrics prevailing at the step and trace levels [3].

The proposed arbiters require arbitration; thus, disagreement begins to assume the form of a regress: each criterion introduced to settle the dispute seems to imply a further criterion.

3. The Regress of Interpretations

Wittgenstein's rule-following considerations bring this difficulty into sharper focus. No interpretation settles its own application. Whatever is done can be brought into accord with a rule under some interpretation; interpretations alone cannot determine meaning (PI §198) [12]. A further formulation leaves the gap open, since it must be applied as well. Kripke's plus/quus example dramatises the problem: a finite history of additions is compatible both with the familiar function and with a deviant one diverging outside the cases considered so far [13]. His discussion of an addition algorithm then transfers the challenge to the terms in which the algorithm is stated. The attempt to support a rule by another rule has not resolved the difficulty. It has merely relocated it.

But this is not Wittgenstein's conclusion. PI §201 regards the repeated demand for one more interpretation as evidence of a misunderstanding: there is a manner of grasping a rule that is not itself an interpretation, and it is displayed in actual cases of obeying and violating the rule. PI §202 identifies rule-following as a practice [12]. This appeal to practice is easily misconstrued, which is why McDowell's anti-sceptical reading matters here. Communal assent does not manufacture correctness, nor can shared behaviour substitute for meaning. A practice is the background within which a performance can count as a correct or incorrect application at all [14]. Publicity is therefore not a vote; it is a condition for the operative distinction between seeming right and being right.[1]

4. The Regress of Faithfulness

The previous section pointed out the mistake that leads to the regress. Faithfulness evaluation reproduces the equivalent structure when a metric is used not just to test an explanation, but to supply the rule for applying the predicate *faithful*. Suppose f is a trained model, e is a post-hoc explanation, and M is a faithfulness metric. When M returns a score, the comparison is restricted to a finite selection of f's behaviour. The metric first creates a test environment through choices such as a neighbourhood, a baseline, or a perturbation scheme. It then observes a finite set of responses. Strictly speaking, $M(e, f)$ compares e with a behavioural profile $B_M(f, e)$ that is shaped by both the metric and the explanation being tested. This dependence on e is not incidental. In deletion and insertion tests, the explanation's ranking determines the perturbation path that becomes visible. As a result, aspects like preservation, locality, relevance, and causal influence are built into the profile. M counts as an interpretation in the sense established in §3: it reformulates the rule for applying *faithful* as success under a selected test, rather than reading that standard neutrally from f, as the extraction picture would require.

The variation among familiar tests extends beyond technical implementation to the questions each test asks. For example, deletion tests look at what happens when highly ranked features are removed, while insertion tests check whether restoring them recovers the output. Stability relates to nearby inputs, surrogate metrics bear on selected predictions, and randomisation tests concern sensitivity to changes in model parameters or training labels. Each approach is defensible, but none of them alone captures faithfulness as such.

Metric plurality is not itself the target of the argument. Taken as stipulative, purpose-relative indicators, different tests may legitimately capture different dimensions of the model–explanation relation without requiring a meta-metric. The regress begins only when a metric is promoted into a context-independent criterion of what counts as *faithful*. The comparisons reviewed in §2 motivate the demand for arbitration among explanations. The present argument concerns the further claim that a metric's verdict can settle explanatory correctness independently of the purpose and target of the test. Moving evaluation up one level makes the Wittgensteinian difficulty fully visible. To rank M_1 and M_2, a meta-metric has to choose which cases to test, define success, set tolerances, and decide which invariances or discriminations matter. This creates another behavioural profile. But why should this new profile be the standard? More data can make results more reliable for a given purpose, but they cannot determine that purpose on their own. In Kripke's idiom, rival metrics extend the predicate *faithful* differently beyond the cases assessed so far. The observed behaviour does not by itself decide which extension is correct. Like the interpretations in PI §201, the meta-metric is still part of the practice, not outside it. So, the regress shifts to a new level.

Nevertheless, the analogy must be delimited. The regress does not concern f's input–output function, which is assumed to be determinate.[2] Ordinary empirical underdetermination presupposes a fixed target

[1] The argument remains neutral between Kripke's sceptical reconstruction and McDowell's anti-sceptical reading. It uses the regress structure, while following McDowell in denying that Wittgenstein's positive appeal to practice reduces correctness to communal ratification.

[2] A more radical route would question whether computational rules themselves are determinately implemented. It is not pursued here: the trained model's input–output function is granted as mathematically determinate (cf. Sprevak [15], who uses the Church–Turing thesis to address an auxiliary problem about generalising local replies across natural language, not Kripke's original rule-following problem).

independent of the available evidence; rival operationalisations may therefore converge as evidence improves. In this case, further facts about f awaiting discovery would leave open the standard for deciding when an application of *faithful* is correct. The rule-following argument says that such a standard is not made self-applying by a further formulation. It becomes operative only in practice. The contrast is therefore between not yet knowing which empirical hypothesis is true and asking what makes this use of the predicate correct. The latter is the rule-following problem proper.

This also shows why the argument does not generalise to every scientific operationalisation. Some measurements can be assessed against a target articulated independently of their indicators. The faithfulness literature surveyed here, however, offers no generally agreed, independent account of "the model's reasoning" against which competing criteria can be appraised [7, 8]. Operationalisation itself supplies the norm that it was meant to measure.

Ground truth might seem like an obvious solution. Transparent models and synthetic tasks can create strong validation settings. Miró-Nicolau et al. use a decision-tree path as the known explanation [2]; Gur-Arieh et al. design tasks where the answer shows that specified intermediate computations have occurred [3]. Do such cases stop the regress? Not in general. These strategies succeed because they fix in advance what counts as the explanatory target: a tree path, an unavoidable bottleneck step, or a controlled response to a hint. They validate a metric against a stipulated, publicly specified criterion. Far from overturning the rule-following diagnosis, they illustrate its positive point: correctness becomes assessable where a criterion has been fixed within a practice of application. What these benchmarks do not provide is a general criterion of explanatory correctness derived from model behaviour alone.

The empirical literature is best read as illustrating limitations of behavioural criteria, not as confirming a grammatical thesis. Failed parameter and data randomisation tests [16], local fragility under minor perturbations [17], and adversarial scaffolding that conceals discriminatory behaviour from LIME and SHAP [18] give practical reasons to scrutinise the authority assigned to behavioural profiles. The prediction is conditional: if behavioural criteria cannot ground their own significance, then a purely behavioural meta-metric will not terminate the regress without importing a purpose, target, or standard from outside the measured behaviour. Given the regress, evaluation requires an account of how the justificatory practice should be organised; practice itself cannot be eliminated.

5. Ending the Regress: From Faithfulness to Contestability

In §3, I argued that the regress is not terminated by a further interpretation, but by the practice in which a rule is applied. In XAI, that practice is asking for, giving, challenging, and revising reasons. Consequently, the evaluation of explanations cannot remain a detached scoring exercise: a request for explanation is a situated normative act. When someone asks about a decision, they ordinarily do so in relation to an implicit or explicit alternative, having a particular purpose in mind. Miller's synthesis of the social-science literature offers converging support. Explanations are contrastive, select only some of the possible causes, and transfer knowledge through interaction [19]. An account constructed without regard to the question, the audience, or the possibility of follow-up may match a selected behavioural profile and still fail as an explanation.

Under these circumstances, the role of faithfulness changes. It remains a necessary constraint: conflict with strong evidence about the model disqualifies an explanation, although compatibility alone falls short. Once the explanatory task is situated, adequacy also becomes question-relative. Feature attributions, counterfactuals, mechanistic descriptions, and procedural justifications respond to different challenges; their adequacy depends on whether the response addresses the relevant fact–foil contrast and the stakeholder's epistemic or practical task. This relocates standards of correctness, rather than eliminating them, within a practice where claims can be examined and objections can influence outcomes.

Contestability gives this practice its positive form. A good explanation exposes its commitments: claims, evidence, and assumptions must be open to scrutiny. The interlocutor must be able to compare alternatives and raise objections, and a successful challenge should influence the assessment and,

where possible, the decision itself. Operationally, an explanation should be able to answer publicly assessable critical questions, such as "Why this feature rather than the corresponding feature in the foil?" This is achieved by disclosing its supporting evidence and assumptions, together with potential defeaters of its claim. A faithfulness score may serve as evidence in this exchange, but it is not the final arbiter. This evidential role does not reintroduce the regress. A score is assessed relative to an explicitly specified explanatory target and remains defeasible within the exchange; it may be weighed against other considerations or set aside after a successful challenge. The problem arises only when the score is treated as an autonomous final criterion. The proposal remains programmatic, but the diagnosis clarifies its direction: metrics should be situated within a justificatory practice, not treated as the sole standard.

Consider a denied credit application. The same SHAP attribution may be relevant to a regulator asking whether a protected proxy influenced the decision and to an applicant asking what could change the outcome. Yet it performs different explanatory functions in the two exchanges. For the regulator, it may form part of the evidence of compliance with an external norm; for the applicant, it may be inadequate unless it identifies an actionable contrast or the conditions under which the decision could be revised. The technical result has not changed. The explanatory task has.

Practice in this context is not about consensus, but about the level at which standards can be articulated, applied, and critiqued. Wittgenstein's argument alone does not determine this institutional form. In high-stakes XAI, contestability also relies on additional normative premises: affected parties have standing to request reasons, evidence must be publicly assessable, successful objections must be able to alter the assessment, and standards must remain accountable to external legal and ethical norms. These premises are reflected in Henin and Le Métayer's distinction between descriptive explanation and normative justification, as well as in their account of justifiability and contestability grounded in such norms [20]. Walton provides the corresponding dialogical structure by treating explanation as a rule-governed exchange tested through critical questions and, when necessary, a shift to examination [21]. Argument mining can help recover claims and relations of support or attack from natural language [22], while formal accounts of reliability and computational trust can make declared reliability requirements checkable [23, 24]. These tools make visible an auditable practice in which standards are declared, applied, challenged, and revised, without extracting norms from the model.

6. Conclusion

The paper began with a familiar anomaly in XAI: the disagreement among explanations is reproduced among metrics, while meta-evaluation relocates rather than resolves the need for justification. I have argued that this instability is characterised by the structure of Wittgenstein's regress of interpretations. A behavioural test does not read a self-applying standard from the model–explanation relation, as the extraction picture supposes; it selects a finite projection of that relation. A further test repeats the same move, producing another projection whose authority must in turn be justified. The question is no longer which measurement is more refined, but what makes an application of *faithful* count as correct. The demand for another interpretation comes to an end only in public practices in which standards are applied, challenged, defended, and revised.

Faithfulness to the model should be retained as a necessary evidential constraint and not elevated into an autonomous criterion of explanatory correctness. Relevance, justificatory adequacy, and contestability must come into play where compatibility with model behaviour no longer decides among competing explanations. Subsequent work should formalise these dimensions within a framework of argumentation and investigate their operational consequences. The diagnosis may apply to the domain of mechanistic interpretability as well: even where internal causal structure is determinate, choosing among causally adequate decompositions repeats the criterial move within the model. Explanation is possible. Faithfulness was simply asked to do too much: it measured a constraint while being asked to ground a norm.

Declaration on Generative AI

During the preparation of this work, the author used GPT-5.6 (OpenAI) in order to: Translation; and Claude (Anthropic) in order to: Grammar and spelling check. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, H. Lakkaraju, The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective, Transactions on Machine Learning Research (2024).

[2] M. Miró-Nicolau, A. Jaume-i-Capó, G. Moyà-Alcover, A Comprehensive Study on Fidelity Metrics for XAI, Information Processing & Management 62(1) (2025) 103900. doi:10.1016/j.ipm.2024.103900.

[3] Y. Gur-Arieh, A. Marasović, M. Geva, Faithfulness Metrics Don't Measure Faithfulness: A Meta-Evaluation with Ground Truth, arXiv:2605.25052 (2026).

[4] A. Páez, The Pragmatic Turn in Explainable Artificial Intelligence (XAI), Minds and Machines 29(3) (2019) 441–459. doi:10.1007/s11023-019-09502-w.

[5] J. C. Bjerring, J. Mainz, L. Munch, Deep Learning Models and the Limits of Explainable Artificial Intelligence, Asian Journal of Philosophy 4, 22 (2025). doi:10.1007/s44204-024-00238-8.

[6] J. A. Pérez-Escobar, D. Sarikaya, Philosophical Investigations into AI Alignment: A Wittgensteinian Framework, Philosophy & Technology 37, 80 (2024). doi:10.1007/s13347-024-00761-9.

[7] A. Jacovi, Y. Goldberg, Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?, in: Proceedings of ACL 2020, 2020, pp. 4198–4205.

[8] M. Nauta, et al., From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI, ACM Computing Surveys 55(13s) (2023) Article 295, 1–42. doi:10.1145/3583558.

[9] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of KDD '16, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[10] S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems 30, 2017.

[11] Z. C. Lipton, The Mythos of Model Interpretability, Communications of the ACM 61(10) (2018) 36–43. doi:10.1145/3233231.

[12] L. Wittgenstein, Philosophical Investigations, 4th ed., trans. G. E. M. Anscombe, P. M. S. Hacker, J. Schulte, Wiley-Blackwell, 2009, §§198–202.

[13] S. A. Kripke, Wittgenstein on Rules and Private Language: An Elementary Exposition, Harvard University Press, 1982.

[14] J. McDowell, Wittgenstein on Following a Rule, Synthese 58 (1984) 325–363.

[15] M. D. Sprevak, Kripke's Paradox and the Church–Turing Thesis, Synthese 160 (2008) 285–295. doi:10.1007/s11229-006-9120-2.

[16] J. Adebayo, et al., Sanity Checks for Saliency Maps, in: Advances in Neural Information Processing Systems 31, 2018.

[17] D. Alvarez-Melis, T. S. Jaakkola, On the Robustness of Interpretability Methods, in: ICML Workshop on Human Interpretability in Machine Learning, 2018.

[18] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods, in: Proceedings of AIES '20, 2020, pp. 180–186. doi:10.1145/3375627.3375830.

[19] T. Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences, Artificial Intelligence 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.

[20] C. Henin, D. Le Métayer, Beyond Explainability: Justifiability and Contestability of Algorithmic Decision Systems, AI & Society (2021). doi:10.1007/s00146-021-01251-8.

[21] D. Walton, A Dialogue System Specification for Explanation, Synthese 182(3) (2011) 349–374. doi:10.1007/s11229-010-9745-z.
[22] J. Lawrence, C. Reed, Argument Mining: A Survey, Computational Linguistics 45(4) (2019) 765–818. doi:10.1162/COLI_a_00364.
[23] F. A. D'Asaro, F. A. Genco, G. Primiero, Checking Trustworthiness of Probabilistic Computations in a Typed Natural Deduction System, Journal of Logic and Computation 35(6) (2025). doi:10.1093/logcom/exaf003.
[24] G. Primiero, On the Foundations of Computing, Oxford University Press, 2020. doi:10.1093/oso/9780198835646.001.0001.