

Leveraging Complementary AI Explanations to Mitigate Misunderstanding in XAI

Yueqing Xuan

School of Computing Technologies
RMIT University, Australia
yueqing.xuan@student.rmit.edu.au

Kacper Sokol

Department of Computer Science
ETH Zurich, Switzerland
kacper.sokol@inf.ethz.ch

Mark Sanderson, Jeffrey Chan

School of Computing Technologies
RMIT University, Australia
{mark.sanderson, jeffrey.chan}@rmit.edu.au

Abstract—Artificial intelligence explanations can make complex predictive models more comprehensible. To be effective, however, they should anticipate and mitigate possible misinterpretations, e.g., arising when users infer incorrect information that is not explicitly conveyed. To this end, we propose *complementary explanations* – a novel method that pairs explanations to compensate for their respective limitations. A complementary explanation adds insights that clarify potential misconceptions stemming from the primary explanation while ensuring their coherency and avoiding redundancy. We introduce a framework for designing and evaluating complementary explanation pairs based on pertinent qualitative properties and quantitative metrics. Our approach allows to construct complementary explanations that minimise the chance of their misinterpretation.

Index Terms—machine learning, artificial intelligence, explainability, intelligibility, comprehension, evaluation, human-centred.

I. INTRODUCTION

Artificial intelligence (AI) explanations assist users in interpreting the general functioning as well as selected details of complex predictive models. When those AI models are opaque, their explanations become a crucial bridge ensuring that users can better understand them. *Effective* explanations should not only maximise comprehension of insights that they convey but also mitigate their misinterpretation. While relevant research has mainly focused on the former task – by improving the intelligibility and informativeness of AI explanations – a critical, yet often overlooked, challenge remains in minimising user misunderstanding, in particular when explainees infer spurious information that an explanation does not provide.

The key cognitive bias contributing to this phenomenon is *the illusion of explanatory depth*, leading users to believe that they understand a system in greater detail than they actually do [1], [2]. This can result in flawed judgment, automation misuse and miscalibrated trust, where users either over-rely on or unjustifiably dismiss AI output [3]–[5]. Prior work has shown that explanations for which users struggle to identify unspecified information – thus realise their limitations – are particularly prone to misinterpretation [6]. To ensure responsible AI adoption, it is critical not only to enhance user comprehension of what an explanation conveys but also to prevent users from making unwarranted generalisations about information that remains unknown.

A common approach to mitigate misinterpretation is to explicitly inform users about the limitations of an explanation

and caution them against drawing conclusions from missing information [7]. However, given the breadth of missing information in an explanation, listing all such omissions is neither practical nor effective as it risks overwhelming users with excessive detail. Another approach is to make explanations interactive, which allows users to independently explore AI models [8]. While this strategy provides access to richer information, it does not guarantee that users will develop the necessary understanding; also, we still lack systematic methodology for constructing interactive explanations while preventing their individual misinterpretation.

In this paper we introduce *complementary explanations* to mitigate user misunderstanding of AI explanations. A complementary explanation provides additional information specifically designed to address common misinterpretations and clarify missing details that users are likely to incorrectly infer from the primary explanation. Intelligible explanations, while effective in conveying explicated information, can paradoxically be misleading if they unintentionally cause users to overgeneralise the information that they provide [6]. To this end, we propose a framework for systematically identifying misleading aspects of an explanation and pairing it with a complementary explanation that addresses these gaps.

Our framework is applied in three steps. First, we identify the explicitly conveyed and unspecified information of an AI explanation. Second, we select its complementary explanation based on four desiderata: (1) degree of *novelty*; (2) difference in *granularity*; (3) extent of *non-redundancy*, i.e., capacity to provide additional information rather than reiterate existing insights; and (4) level of *coherency*, i.e., ability to deliver sufficient shared context to help users integrate multiple explanations into a meaningful bigger picture. Third, we quantify the degree of complementarity between explanations using dedicated metrics. This structured approach enables constructing principled complementary explanation pairs that enhance user comprehension while minimising misinterpretation risk.

II. BACKGROUND AND RELATED WORK

When users interact with AI models through their explanations it is crucial that they understand how to correctly interpret such insights [9]. A common approach is to explicitly communicate the limitations of AI explanations, e.g., indicate

the information that remains unspecified by explanatory artefacts [7]. Consider an AI model assisting doctors in predicting diabetes risk; a local explanation tailored to a particular patient might state the limitation of its information in a form of a disclaimer: “the explanation applies exclusively to this patient and does not imply that *glucose* is the most important factor across all the cases”. Explicitly communicating explanation scope helps to prevent its overgeneralisation.

Simply disclosing the limitations of an explanation may nonetheless be insufficient to curb its misinterpretation. This is because explanations tend to vary in scope (e.g., global, local or sub-space) and information content (e.g., feature influence or counterfactual insight) [10]. It is thus impractical to identify and communicate all the unspecified aspects of explanatory insights. Moreover, this approach will only succeed if users can understand and operationalise such details. Additionally, while the limitation disclosure narrows down the scope of consideration and reduces the required cognitive effort, it does not facilitate richer understanding of AI models and can potentially erode user trust and engagement.

Another strategy is to allow users to interact with explanations (without supervision) – e.g., through a dynamic interface [11] – helping them to explore an AI model and incrementally expand their knowledge. This approach, however, cannot guarantee that users will learn what is necessary to achieve the desired level of comprehension. Furthermore, the sequence in which explanations are presented affects how users process information and could lead to inconsistent interpretations [12]. Thus providing interactive explanatory information without any underlying structure risks confusing or overwhelming users, possibly leading to unexpected misinterpretations.

Complementary explanations address such challenges by combining structured (interactive) exploration with carefully selected explanatory information. They present different yet coherent insights that guide users through the explainability process while minimising cognitive load. Our complementary explanations therefore align with the concept of *explanations as a social practice* by facilitating interaction and co-construction [13]. While current research has explored unifying different explanation types – e.g., combining dataset analysis with global feature importance [14] and contextualising local feature attribution with partial dependence plots [15] – it is mostly limited to preselected pairs of explanations in isolated contexts, lacking systematic guidelines on selecting and integrating various explanations. Our work addresses this gap by demonstrating how to design complementary explanations that minimise misunderstanding and support learning.

III. COMPLEMENTARY EXPLANATION FRAMEWORK

Our complementary explanation framework builds upon existing explanation types, aligning them in a systematic way.

A. Identifying Information

Complementary explanations match different explanation types by identifying their pairs that offer complementary information to compensate for their respective shortcomings

TABLE I: Selected explanation types – split into three groups: global, local and sub-space – and their information scope. These explanations can be generated with well-established, model-agnostic tools such as LIME, SHAP and PDP.

Explanation	Explicated Information	Unspecified Information
partial dependence plot	overall effect of a feature on the model’s output	feature interaction & instance-level insights
(surrogate) decision rules & trees	rules & trees approximating overall model behaviour	surrogate fidelity & feature relationship
feature importance	overall importance of features for predictions	instance-wise importance & feature interaction
data distribution analysis	dataset statistics, outliers & feature distribution	context of how feature distribution affects predictions
counterfactual & decision surface	changes required to alter a specific prediction	how changes affect other predictions
feature attribution	contribution of individual features to a prediction	feature interaction & global feature importance
nearest neighbours	closest training points to a given input	global patterns & model-level insights
influence function	influence of a data point on a prediction	how data & features interactions impact predictions
prototypes & criticisms	representative examples & edge cases or exceptions	broader regional patterns & model generalisability
regional feature importance	feature importance within a specific data sub-space	global feature importance & variation across spaces

and limitations. When linked to a primary explanation, the complementary explanation adds information that the former does not fully convey and is likely to be misinterpreted. For example, user studies have shown that lay people often infer local feature attribution from decision surface visualisation and counterfactuals [6]. By presenting feature importance alongside these explanation types we can prevent such misconceptions. In this context, local feature attribution is complementary to decision surface visualisation and counterfactuals since it minimises the chances of users inferring incorrect insights that the latter do not specify.

Given the variety of available explanation types, we require a systemic guideline to assess their information scope – whether explicated or unspecified – to apply our framework. Specifically, we need to know: (1) what information an explanation *communicates*, and whether the explicated information is intelligible; and (2) what information an explanation *misses*, and what *unspecified* insight is the most likely to be misconstrued by users. Table I provides answers to these questions for a selection of representative AI explanation types.

B. Design Principles

Identifying the information scope of different explanations allows us to determine their complementarity, which we chart across four dimensions: novelty, granularity, non-redundancy and coherency.

1) *Novelty*: The complementary explanation should provide information beyond what is communicated by the corresponding primary explanation. For example, local feature

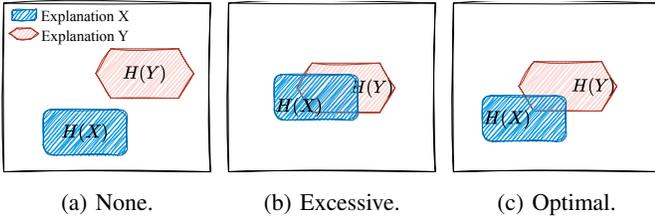


Fig. 1: *Mutual Information* examples for explanations X & Y .

attribution can be complemented by data distribution analysis, which contextualises feature values of an individual within the broader population. Since the former explanation type lacks insights about (a) typicality of feature values, the latter addresses this shortcoming by showing how these attribute values align with or deviate from the modelled population. This extension connects local insights to broader data patterns.

2) *Granularity*: The complementary explanation can present a different level of granularity, providing either a broader or a more detailed view. For example, a high-level explanation of model behaviour may highlight that *location* is the most important feature in predicting house pricing, while a local explanation might reveal that *size* is more influential for a specific home. Providing explanations at multiple granularity levels helps users to engage with AI models on various analytical planes. This combination also caters to different cognitive approaches: inductive reasoning, which generalises from specific instances to broader patterns, and deductive reasoning, which applies general rules to specific outcomes [16].

3) *Non-redundancy*: The complementary explanation should avoid excessive repetition of information communicated by the primary explanation. For example, decision rules show conditions leading to specific outcomes, and counterfactuals describe how to tweak input features to change outcomes. Since both focus on feature values that determine the output of a model, they are largely redundant in conjunction.

4) *Coherency*: An explanation pair should share context to help users integrate these insights into a coherent mental representation of an AI model. Coherency ensures that explanations are perceived as interconnected parts of a whole rather than isolated pieces of information. An example of incoherency would be pairing a counterfactual with global feature importance. While the former focuses on instance-level changes, the latter ranks features by their overall impact. The lack of a direct connection between the individual and global perspectives can prevent users from reconciling these explanations to form a unified understanding of an AI model.

C. Evaluation Metrics

Next, we introduce three metrics to quantify complementarity of explanation types and guide their selection.

Metric 1 (Information Richness). *Information richness* $H(X)$ of explanation X measures the amount of intelligible information that X conveys about the behaviour of an AI model.

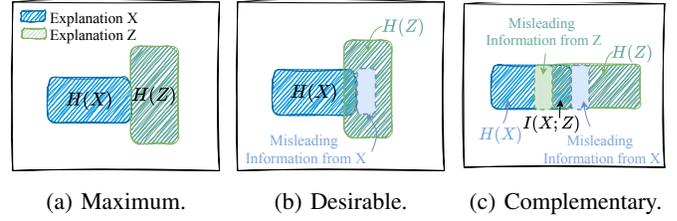


Fig. 2: *Information Gain* examples for explanations X & Z .

The value of *Information Richness* – Metric 1 – depends on factors such as the number of features included in X , the modality and size of X , and its presentation structure. For example, the length of an explanation or the number of new concepts that it introduces as well as the overall size of the feature set can be used to this end [6], [17], [18]. Similarly, the cognitive load of visual AI explanations can be quantified by analysing feature counts and visual trends [19]. Since overly complex explanations can cause cognitive overload [20], $H(X)$ exhibits sub-linear growth as explanation content increases, reflecting the diminishing returns when additional information becomes harder to process.

Metric 2 (Mutual Information). *Mutual information* $I(X; Y) = H(X) \cap H(Y)$ quantifies the amount of information shared between explanations X and Y .

Information Richness characterises a single explanation. When aligning multiple explanations, we use *Mutual Information* – Metric 2 – to measure the degree of redundancy between them. As discussed in Sect. III-B, effective complementary explanations should have small content overlap (non-redundancy) while preserving some shared information to maintain thematic coherency. Figure 1 illustrates this concept. Ideally, $I(X; Y)$ should be low but not zero; when $I(X; Y) \approx 0$ users may struggle to reconcile the explanations and develop cohesive understanding (Fig. 1a). As an example, consider counterfactuals paired with global feature importance.

When $I(X; Y)$ is excessive (Fig. 1b), such as decision rules and counterfactuals, the second explanation adds little new insight but processing it requires extra cognitive effort. The optimal amount of *Mutual Information* (Fig. 1c) delivers novel insights and facilitates coherent comprehension. For example, pairing counterfactuals with local feature attribution offers both diagnostic (what contributed to the decision) and actionable (what to change to alter the outcome) insights.

Metric 3 (Information Gain). *Information gain* $IG(Y, X) = H(Y) - I(X; Y)$ quantifies the amount of new information that explanation Y provides in the context of explanation X .

While *Mutual Information* captures how well a complementary explanation aligns with the primary one, we also need a metric to assess its ability to offer new insights and mitigate any potential misunderstanding caused by the primary explanation. *Information Gain* – Metric 3 – captures insight novelty and granularity difference that the complementary explanation

planation introduces – see Fig. 2. Maximum *Information Gain* is achieved when there is no overlap between explanations, i.e., $I(X;Y) = 0$, and $H(Y)$ is large enough to ensure sufficient *Information Richness* (Fig. 2a). This, however, is suboptimal as some shared information is needed for coherency. It is more desirable for the insights provided by Y to fill the informational gap left by X to mitigate any misunderstanding (Fig. 2b). In this instance, the complementary explanation not only provides novel insights but also rectifies potential misconceptions. Furthermore, maintaining sufficient shared information helps users to merge the explanations into cohesive understanding and reduces cognitive load.

Based on *Mutual Information* and *Information Gain*, we introduce *complementary* explanations – illustrated in Fig. 2c. Each explanation in this pair mitigates potential misconceptions introduced by the other. Their shared information fosters coherent understanding, ensuring that they work together to broaden comprehension. For example, consider local feature attribution and counterfactual explanations in the context of predicting heart attack risk. The former highlights the contribution of the most influential attributes – e.g., cholesterol, blood pressure and exercise frequency – to a model’s output; it shows why a prediction was made but lacks information about how changes to feature values affect this outcome. A user may thus falsely believe that lowering cholesterol alone is sufficient to reduce their heart attack risk since it is the biggest factor. This interpretation, however, overlooks possible feature interactions.

Counterfactual explanations fill this gap by suggesting the smallest actionable feature change, e.g., reducing blood pressure in combination with increasing exercise frequency, which implicitly accounts for attribute interaction. Thus rather than applying a big change to the most important feature (cholesterol), small adjustments to multiple attributes bring the desired outcome. When provided in isolation, counterfactuals may nonetheless mislead a user into believing that blood pressure and exercise frequency are the most significant factors, causing them to overlook the contribution of each attribute independently [6].

Feature attribution addresses the limitation of counterfactuals and clarifies that cholesterol is still the most influential factor. Together, these explanations address each other’s limitations. Their shared information – common explainability scope, context and content (i.e., feature set) – brings them together into a coherent narrative. In summary, complementary explanations are mutually enriching, helping to improve overall user understanding of AI models.

IV. CONCLUSION AND FUTURE WORK

Ensuring that AI explanations do not mislead users about the information they do not communicate is crucial. This paper introduced a framework and guidelines for designing complementary explanations that minimise the chances of any such misunderstanding. It also defined criteria to quantify the complementarity of explanations and identified complementary explanation pairs that address each other’s limitations.

Future work will focus on user studies aimed at empirical validation of our framework. Specifically, we will investigate if complementary explanations reduce misconceptions more effectively than simply disclosing the limitations of each individual explanation. Accounting for user demographics will be crucial when applying our framework as the perceived complementarity of explanation types may vary across stakeholders [21]. We further plan to examine the impact of complementary explanations on user trust and cognitive load as overly complex explanation pairs may adversely affect both.

ACKNOWLEDGEMENTS

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (project number CE200100005), funded by the Australian Government through the Australian Research Council. Additional support was provided by the Hasler Foundation (grant number 23082).

REFERENCES

- [1] M. Chromik *et al.*, “I think I get your point, AI! The illusion of explanatory depth in explainable AI,” in *IUI*, 2021, pp. 307–317.
- [2] R. Byrne, “Good explanations in explainable artificial intelligence (XAI): Evidence from human explanatory reasoning,” in *IJCAI*, 2023.
- [3] R. Parasuraman and V. Riley, “Humans and automation: Use, misuse, disuse, abuse,” *Human Factors*, pp. 230–253, 1997.
- [4] A. Jacovi *et al.*, “Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI,” in *FACCT*, 2021, pp. 624–635.
- [5] D. Ahn *et al.*, “Impact of model interpretability and outcome feedback on trust in AI,” in *CHI*, 2024, pp. 1–25.
- [6] Y. Xuan *et al.*, “Comprehension is a double-edged sword: Overinterpreting unspecified information in intelligible machine learning explanations,” *International Journal of Human–Computer Studies*, p. 103376, 2025.
- [7] N. van Berkel *et al.*, “Effect of information presentation on fairness perceptions of machine learning predictors,” in *CHI*, 2021.
- [8] K. Sokol and P. Flach, “Glass-Box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant,” in *IJCAI*, 2018, pp. 5868–5870.
- [9] K. Sokol and J. Vogt, “What does evaluation of explainable artificial intelligence actually tell us? A case for compositional and contextual validation of XAI building blocks,” in *CHI EA*, 2024, pp. 1–8.
- [10] R. Guidotti *et al.*, “A survey of methods for explaining black box models,” *ACM Computing Surveys*, pp. 1–42, 2018.
- [11] H.-F. Cheng *et al.*, “Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders,” in *CHI*, 2019, pp. 1–12.
- [12] H. Kaur *et al.*, “Sensible AI: Re-imagining interpretability and explainability using sensemaking theory,” in *FACCT*, 2022, pp. 702–714.
- [13] K. Rohlfing *et al.*, “Explanation as a social practice: Toward a conceptual framework for the social design of AI systems,” *IEEE Transactions on Cognitive and Developmental Systems*, 2020.
- [14] A. Bhattacharya *et al.*, “EXMOS: Explanatory model steering through multifaceted explanations and data configurations,” in *CHI*, 2024.
- [15] C. Bove *et al.*, “Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users,” in *IUI*, 2022.
- [16] D. Wang *et al.*, “Designing theory-driven user-centric explainable AI,” in *CHI*, 2019, pp. 1–15.
- [17] M. Narayanan *et al.*, “How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation,” *arXiv preprint arXiv:1802.00682*, 2018.
- [18] I. Lage *et al.*, “Human evaluation of models built for interpretability,” in *CHI*, 2019, pp. 59–67.
- [19] A. Abdul *et al.*, “COGAM: Measuring and moderating cognitive load in machine learning model explanations,” in *CHI*, 2020.
- [20] J. Bo *et al.*, “Incremental XAI: Memorable understanding of AI with incremental explanations,” in *CHI*, 2024.
- [21] U. Ehsan *et al.*, “The who in XAI: How AI background shapes perceptions of AI explanations,” in *CHI*, 2024, pp. 1–32.