Maintaining Coherence in Explainable AI: Strategies for Consistency Across Time and Interaction

Alan Dix[1,2,*], Tommaso Turchi[3], Ben Wilson[2], Alessio Malizia[3,4], Anna Monreale[3] and Matt Roach[2]

[1]*Cardiff Metropolitan University, Wales, UK*
[2]*Computational Foundry, Swansea University, Wales, UK*
[3]*Department of Computer Science, University of Pisa, Pisa, Italy*
[4]*Molde University College, Molde, Norway*

Abstract

Can we create explanations of artificial intelligence and machine learning that have some level of consistency over time as we might expect of a human explanation? This paper explores this issue, and offers several strategies for either maintaining a level of consistency or highlighting when and why past explanations might appear inconsistent with current decisions.

Keywords

human-AI interaction, explainable AI, synergistic human-AI systems, user interface, artificial intelligence, design, adaptive interfaces, user experience

1. Introduction

This paper considers how XAI systems can behave in ways that are coherent over time, mirroring the expectations of consistency for human explanations.

It is widely believed that there are advantages to having AI systems that are comprehensible to human users. This has been part of the literature since the early 1990s [1], in particular highlighting the potential for ethnic, socio-economic and gender bias in black-box ML and the way that explanation as a form of transparency can help expose this. However, over recent years the issue has become a major area of both research and practical development, with numerous algorithms [2, 3, 4, 5], frameworks and surveys [6, 7, 8, 9, 10, 11, 12].

Myers and Chater argue that a human explanation is not just an atomic utterance, but that we expect a level of *coherence* over time [13, 14]; that is future statements and explanations should be consistent with previous ones. Indeed, this is part of the implicit contract between the parties that enables mutual trust, effective communication and collaboration. For example, if Alan explains a food choice by saying "I prefer sausages to poultry", you would expect him to subsequently choose sausages if given a choice. Myers and Chater extend their argument from the realm of human explanation to highlight 'what it would really mean for AI systems to be explainable'. They argue that AI explanations equally should have some level of consistency. Myers and Chater build their position based on extensive theoretical and empirical literature from psychology, sociology and XAI; we do not repeat this here beyond motivating examples. In this paper, we take a next step exploring the different ways that this consistency can occur within XAI settings and some potential algorithmic strategies to ensure this in practice.

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy
*Corresponding author.
alan@hcibook.com (A. Dix); tommaso.turchi@unipi.it (T. Turchi); b.j.m.wilson@swansea.ac.uk (B. Wilson); alessio.malizia@unipi.it (A. Malizia); anna.monreale@unipi.it (A. Monreale); m.j.roach@swansea.ac.uk (M. Roach)
https://alandix.com/ (A. Dix)
0000-0002-5242-7693 (A. Dix); 0000-0001-6826-9688 (T. Turchi); 0009-0004-5663-5854 (B. Wilson); 0000-0002-2601-7009 (A. Malizia); 0000-0001-8541-0284 (A. Monreale); 000-0002-1486-5537 (M. Roach)

In the next section we'll look at the different ways (in)coherence may manifest in AI systems, and then move on to consider ways this can be managed such as explaining incoherence or avoiding it happening. Notions of nearness, closeness or local neighbourhoods are crucial to both.

Note this paper will deal principally with single point explanations: "the system made this decision about input X because ...". Contrastive explanations may also be very powerful, that is answering questions of the form, "why are the decisions about inputs X and Y different (or the same)?" We also principally focus on coherence *between* explanations and decisions for different inputs or models, that is *inter-response consistency*. In addition for complex explanations (particularly LLMs), we can ask whether the parts of the explanation are coherent, that is *intra-response consistency*. There are also important issues regarding *instability* of explanations [15, 16] to the same input (in the case of stochastic algorithms or ongoing learning) and of explanations of the same decision given to different people (where there is personalisation, for example in LLMs). We leave a detailed discussion of these issues to a future paper.

2. Types of Incoherence

Within both human–human there are many different meanings of coherence or consistency, with no single clear definition, In general the term 'coherence' seems to be used more for internal consistency with an argument *intra-response consistency* and 'consistency' more to do with the relationship between multiple utterances or between utterance and action *inter-response consistency*. Here we are looking predominantly at the latter, especially in relation to AI explanations, that is the extent to which the decisions/outputs and explanation given by an AI at different times appear to agree or make sense relative to one another. However, there are several ways in which an AI or ML system may exhibit behaviour apparently (in)consistent with previous explanations. We will attempt to be more precise than the fairly open definition above, but ultimately this is about human judgement or impressions of what seems to be coherent.

We will first look at situations where different inputs to the same model give rise to apparent incoherence; that is, an AI medical advice system said that grapefruit was good to eat for one kind of cancer, but not for another. We will then consider cases where the model has changed, say, owing to new training data; perhaps analogous to the doctor changing their opinion based on a new article in The Lancet.

2.1. Notation

We will use the following semi-formal notation for the AI cases:

- X, d_X, e_X – previous input X, decision and explanation for a model M
- Y, d_Y, e_Y – current input Y, decision and explanation for the same
- d'_X, e'_X – decision for input X and explanation for this following a model change to M'

The precise meaning of these differ depending on the kind of input data (e.g. images, medical test results, user interface logs), the kinds of output (e.g. medical diagnosis, classification, automated action) and explanation (e.g. SHAP-style feature importance, linear discriminant, decision tree).

Inconsistency

We will use the symbol $\nsim$ to mean 'apparently contradicts' for different kinds of comparisons. In some cases this is effectively 'not equal', but in others, for example looking at the relation between a feature-importance explanation and particular decision, this is a more complex relationship.

Explanation as function

In many cases explanations are expected to be local [2, 3, 4], that is only operative in a neighbourhood of the particular input. However, the explanation can often be applied in relation to other inputs; we will write $e_X[Y]$ for the explanation given for input X interpreted in the context of input Y. Crucially some explanations can be treated as functions that give a decision for a particular input, in these cases we can think of $e_X[Y]$ as the decision that would be taken given input Y treating e_X as a function.

2.2. Fixed model

First let's consider the case of a fixed model that has been trained or constructed beforehand and does not learn further during the period of use (see Figure 1, upper). We have two main cases:

Inconsistency of decisions – Is the current decision inconsistent with past explanation(s): $d_Y \nsim e_X[Y]$? Is the past decision consistent with current explanation: $d_X \nsim e_Y[X]$?

Inconsistency of explanations – Do the explanations agree in terms of decisions on the inputs, but with different reasoning: $d_Y \sim e_X[Y]$ while $e_X \nsim e_Y$?

A human example of the first case would be if Tommaso said that a Fiat 500 was a good car because it was small and then later said he would like to have a Humvee. An example of the second case would be if he said he liked a blue Fiat 500 because it was small and then later said he liked a blue Mini because it was blue.

This incoherence might be for valid reasons. For example, in the first, Tommaso might prefer a small car for ease of parking at work, but if not for that would really like the idea of driving the Humvee – that is they are local explanations. In the second case, it might be that the explanation of the Fiat 500 had been made in comparison to a SUV whereas that for the (all) blue Mini was in contrast to a red, white and blue striped Mini. Note that the latter, contrastive explanations, need special treatment, which, as noted, we leave for a future paper.

As with Tommaso's reasoning, an AI model might be working well and be a justified inconsistency, albeit initially appearing incoherent. Alternatively, the incoherence may represent a genuine problem in reasoning:

- one or other decision or explanation was simply wrong as the model generalises poorly;
- the act of finding the later explanation effectively opened up ways of looking at the data that would have been better applied to first input (related to model change);
- the explanation finding mechanism has stochastic elements or stability issues in certain areas and simply gives different explanations by chance. In contrast to one being wrong, each might have validity.

2.3. Changed model

Now consider cases where the model has changed due to new training data (see Figure 1, lower). All the above apply, that is we might have presented X to the old model M and Y to the new model M', and found apparent incoherence. These are not pictured for reasons of space, but would be represented by $d'_Y \nsim e'_X[Y]$ and $e'_Y \nsim e_X$, etc. In practice, this may result from new training examples 'near' the old decision points.

In addition, we may see non-monotonic reasoning, that is issues of consistency with the same input X in different models M and M':

Inconsistency of decisions – Has the decision changed? $d'_X \neq d_X$

Inconsistency of explanations – Has the explanation for the same decision changed? $d'_X = d_X$, but $e'_X \nsim e_X$

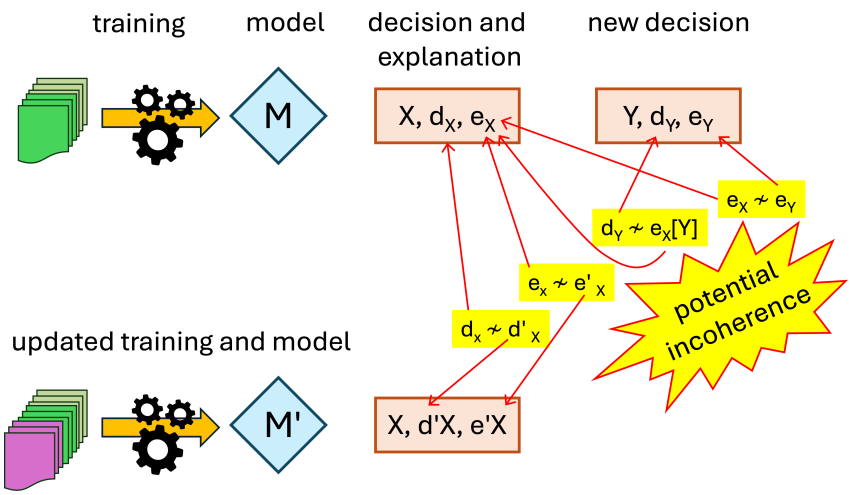

Figure 1: Types of incoherence.

Again similar issues can arise for human–human interactions. The earlier example of a doctor changing their diagnosis or treatment based on a new Lancet article is an example of the first case. A health-related example of the second would be a nutritionist who has always recommended a varied diet in order to ensure a broad range of vitamins and nutrients, but based on recent studies, now makes the same recommendation but emphasising the way a varied diet encourages a diverse gut biome with an ensuing wide impact on mental and physical health.

3. Strategies to improve Coherence

There are several different ways in which we can ensure coherence between decisions and explanations.

highlight inconsistency with previous explanations: "I know I said A before, but this is a different kind of situation". This doesn't ensure consistency, but it maintains a claim to coherence.

explain inconsistency with previous explanations: "I know I said A before, but this is different because of B". This justifies the claim to coherence.

constrain consistency with previous explanations by adding each previous explanation e_i as a constraint when making future decisions. This continually uses past explanations to update, or manage the model, but may run into limits and may only be possible with some kinds of machine learning algorithms.

ensure consistency by using each previous explanation e_i as a local decision rule when the current situation is sufficiently close to the input that gave rise to the explanation. That is completely replace the model rule locally.

We'll look at each of these in a little more detail.

3.1. Highlight Inconsistency

Here the system needs to keep track of previous decisions and explanations and simply detect that there is an apparent inconsistency. The exact form of this detection will vary depending on the form of ML and XAI. As an example, the FRANK system [17] is used during interactive human training, but adopts a mechanism that applies rules based on previous decisions to verify new user input:

"*At first Frank applies the Ideal Rule Check (IRC), checking if the record is covered by one of the given rules*" [18, p.17]

This approach is being used to monitor the consistency of *human* training examples in a process of 'skeptical learning' (Fig. 2). However, the underlying mechanisms for checking aspects of training data is similar to those that would be required to monitor future model decisions/advice. Rather than a human labelling, we would instead check a new AI decision against previous rules.

Detecting and highlighting inconsistency is a fairly minimal strategy, but may help retain user confidence.

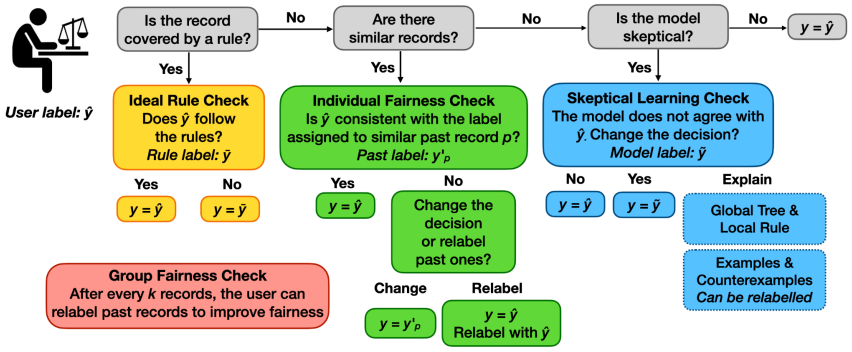

Figure 2: Skeptical Learning (from [18, fig.1]).

3.2. Explaining Inconsistency

Where there is justified inconsistency of any of the types discussed in Section 2.2, we ideally need to explain why this is occurring.

Counterfactual-style explanations are already being used in some XAI contexts [15, 19] where decisions d_X and d_Y differ. For example, given two inputs X and Y that look similar, but are given different decisions d_X and d_Y, we can try to locate training examples t_X and t_Y with labels d_X and d_Y, and close to X and Y respectively (ideally also both 'between' X and Y), thus justifying the different decision.

In a similar way, if $d_Y \nsim e_X[Y]$, we can find a training example t_Y that is close to X or 'between' X and Y, but where the label on t_Y is not what one would expect with $e_X[t_Y]$, thus justifying the limits of the explanation e_X. The same technique can be used with multiple training examples to justify $e_X \nsim e_Y$.

In many ways if the model changes, as described in Section 2.3, less justification is needed as the new training examples quite reasonably will have changed the model. However, if the new examples appear to be very different to an input X, it may still seem odd that the decision d_X or explanation e_X changes so that *change-oriented explanations* are needed. In some cases we may be able to find new training examples that are close to the past example, that is a new t_X that is close to X, but with a different label or incompatible with the old explanation for d_X. In others there may be non-local changes, for example, in a CNN (convolutional neural network) new training data might change low-level features that have impacts on very different input data. This highlights the general XAI challenge of in some way surfacing these intermediate emergent features.

3.3. Constrain Inconsistency

In some cases the underlying algorithm may be able to be constrained to continue to be consistent with a previous explanation. For example, the Query-by-Browsing system uses a variant of ID3 to build decision trees and SQL queries [1], which can be thought of as a single global 'explanation'. However, the top-down nature of ID3 means that small changes in training data may give rise to a completely different decision tree. For this reason, one variant of Query-by-Browsing used genetic algorithms to evolve the decision tree [20]; thus favouring smaller changes to the tree where this is possible consistent with previous data.

A more model agnostic method would be to generate synthetic training examples t_i that are distant from existing training data, but close to a previous example X. If each new training example is labelled

to be consistent with e_X, this effectively cements the explanation for the locality. This is similar to the techniques used to generate privacy-preserving synthetic data in [21].

3.4. Ensure Consistency

As noted, in some cases we can interpret a decision d_X and e_X as a rule 'WHEN in locality L_X APPLY rule R_X'. For example, LIME creates a linear discriminant model by looking at training examples in the region of the input [4]; this both incorporates an existing idea of locality of the explanation (L_X) and an executable rule (R_X).

This collection of locality–rule pairs, (L_i,R_i) can then be used in a two stage process as illustrated in Figure 3: given a new unseen input Y, we first check if it is within a patch, and if so return the result of the rule; if there is no matching patch the original model is used to generate the decision and explanation.

```
Initially a model M, and empty set of patches P.
for each new example X
  1.  look for (L[i],R[i]) in P such that X in L[i]
  2.  if found
     2.1 give decision and explanation by applying R[i] to X
  3.  if not found
     3.1  let dX = decison of M at X
     3.2  let eX = explanation of M at X
     3.3  let L = a locality of X
     3.4  let R = eX interpreted as a rule
     3.5  add (L,R) to P
     3.6  give decision dX and explanation eX
```

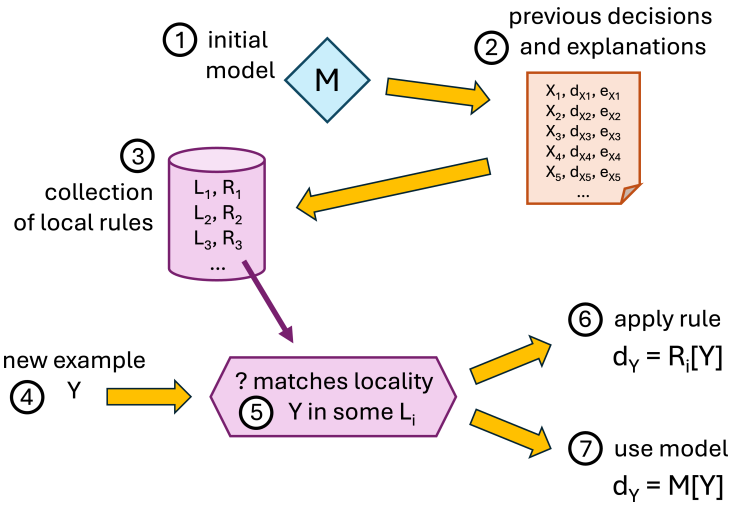

Figure 3: Ensuring consistency with previous explanations

This is rather like ensemble methods where one has multiple models and then meta-learning to create a decision rule to determine which is to be used. In this case the rule set consists of the original model M and a series of example–explanation pairs, (X, e_X), (Y, e_Y), etc. Effectively one is doing ensemble learning on M, e_X, e_Y, but where we have the anchor points X, Y to help.

We can think of these local rules as comprising a partial *patch model* as depicted in Figure 4. The figure also highlights several issues of patch models:

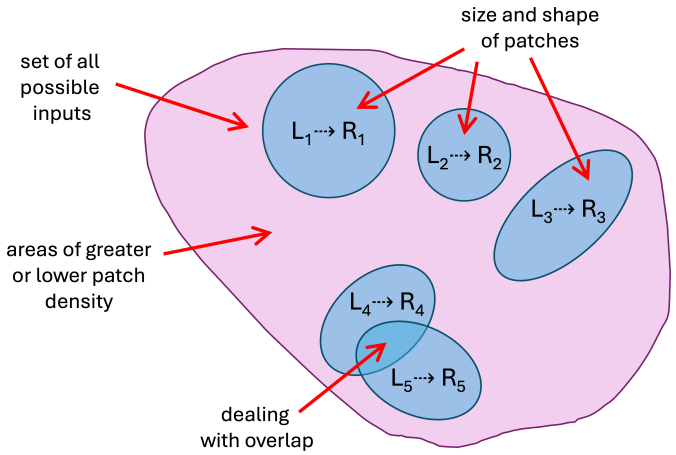

Figure 4: Patch model with important issues

varying density of patches – Some areas of the input space may be densely covered in patches, others relatively empty. For the purposes of coherence, this is not a problem, merely reflecting the distribution of previous user inputs and associated explanations.

varying size and shape of patches – The localities defining the patches may differ in size and (highly multidimensional) shape. The consequent issues of what to regard as 'near' will be discussed in Section 4.

overlapping patches – If two localities overlap and the decisions implied by their rules differ, then there clearly needs to be a meta-decision rule or adjustments of the localities to disambiguate the decision. However, even if the rules agree in the intersection, the explanations will be different, so there still needs to be some adjustment to one or both localities.

The method above is iterative, building a secondary model patch-by-patch. It is also a *partial* patch model, as the patches (initially at least) do not fully cover all inputs and the original model is still used in the gaps.. The GLocalX method [5] is similar, but performs this whole process 'upfront', that is by exploring the entire space, creating local explanations everywhere and then using this to create a complete patch model (global explanation) all before ever encountering any unseen examples.

This method can also be used when there is underlying model change. If new training examples are not 'close' to a patch, the patch can be retained between models, thus also dealing with between-models coherence at a single input point for both decisions and explanations. However, if the underlying model change has not also been limited using some form of 'constrain inconsistency' approach, then this could lead to increased instability at patch boundaries.

4. Nearness and locality

In multiple places we have needed to think about some form of nearness or locality. In Section 2, which considered the ways in which incoherence may occur, explanations for inputs X and Y would only be seen to be in conflict if X and Y are sufficiently close. Similarly, the patch models in Section 3 depend on defining a locality over which each rule operates, typically defined in terms of closeness to a defining example. Indeed 'local' explanation methods such as SHAP and LIME [3, 4] have to have some measure of what is close to a particular input in order to perform perturbations.

Note there are three senses of closeness, one might want to consider:

closeness of input vectors – XY – For binary features this might be Hamming distance or for continuous features some form of Euclidean distance in feature space normalised by individual feature variance.

closeness of outputs/classifications/decisions – d_X d_Y – This might be a binary agree/disagree, but could be a more complex metric of the output such as a set of classifications with weightings.

coherence of explanations – e_X e_Y – This is the metric that is critical for *instability* in XAI [16]. For feature importance explanations this might be a euclidean distance, cosine similarity or Spearman or Kendall rank correlation coefficient . For symbolic explanations, this may be some form of inter-formulae edit distance.

It is the first that we are dealing with in this section, but all are important in different circumstances.

4.1. Localised feature importance

Initially, closeness can be based on a global metric of closeness of input vectors. However, once we have local explanations these can be used to help define more localised metrics. For example, as noted previously, local explanations are often created by looking at training data close to the input; to be 'local' these will have adopted a measure of nearness, which can then be used to create the locality for a patch model.

In addition, the explanation will often create some form of feature importance which can be used to create localised nearness metrics. In the case of perturbation and hotspot methods this is very direct, as each feature is given a direct measure of importance, which can then be used to weight the feature differences in a local Euclidean metric; that is:

$$d(X,Y) = \sum_f w_f (X_f - Y_f)^2$$

where w_f are weights based on the feature importance vector. Note that smaller differences are considered significant where they have higher feature importance, whereas even quite large differences in unimportant features may still be considered 'close'.

In the case of more algorithmic explanations such as decision trees, the fact that a feature is mentioned can be used as metric of feature importance weighting these more highly than others. If the explanation includes derived features (e.g boolean 'SALARY > 50000'), then these can be used to give a scale to the feature. If the SALARY in the input that gave rise to this explanation is 60000, then we would expect the locality of the rule to extend at least to some inputs that are otherwise similar, but with SALARY less then 50000 so that the locality defines a region within which the rule is meaningfully applicable.

4.2. Explaining using measures of nearness

As well as being important for constraining inconsistency or creating patch models, measures of nearness can be used as part of explanations themselves. This might be vague, something like, "X and Y are similar in many ways, but differ in ways which are particularly important for the decision making". More convincing explanations could give the precise metrics being used to make the distinction, for example, "while employee X and Y have similar experience and skills, their jobs differ in terms of risk".

Local measures of nearness could also be used in counterfactual generation. If we are looking for a training example Z to explain the difference between decisions and/or explanations for X and Y, then Z should be 'between' X and Y. The weighted locality metrics for X and Y are likely to be better measures of 'between' than global feature distances.

4.3. Explanations of measures of nearness

Of course, if metrics of nearness contribute to the coherence of explanations and decisions, they must themselves be explainable to end users. For example, rather than simply saying "while X and Y differ substantially in feature F, this is considered unimportant", instead the explanation could be "while X and Y differ substantially in feature F, this feature does not appear in the explanations for X and Y and is therefore considered unimportant".

5. Discussion and Conclusion

This paper has outlined several strategies for achieving coherent explanations in AI systems, particularly in response to temporal or contextual shifts. It has identified several promising directions for future work, including the development of mechanisms for explaining changes in reasoning; the use of patch models that retain and reuse prior explanations; and the exploration of nearness metrics, which determine when explanations can be meaningfully applied to new inputs. More broadly, the diversity of model architectures and explanation types suggests a rich design space for experimenting with coherence strategies, from algorithmic constraints to user interface representations and feedback mechanisms.

From a human-centred AI perspective, coherence in explanations is not merely a technical attribute but a vital social and cognitive affordance that underpins trust and mutual understanding. Just as people rely on consistent reasoning to interpret intentions and anticipate actions, AI systems — all of which operate within socio-technical settings — should treat coherence as a primary design objective alongside accuracy and fairness. For instance, if an AI system revises its reasoning, a transparent shift in rationale (e.g., "Given the user's recent preferences, I now recommend slower but more scenic routes") can maintain user confidence even amidst change. Thus, explanation strategies supporting *temporal narrative continuity* — where decisions are justified in the moment and situated within a comprehensible arc of system behaviour — are key to fostering durable human-AI combination.

While much of this paper focuses on model-specific XAI techniques, large language models (LLMs) increasingly act as *explanation interfaces* — either as direct decision-makers or as natural language layers over other AI systems. In these contexts, their internal consistency over a series of decisions becomes a crucial dimension of explanation quality.

Recent work has shown that large language models (LLMs) often exhibit sycophantic behaviour — a tendency to align their outputs with user biases, personas, or perceived preferences—at the cost of logical consistency and rational argumentation. This phenomenon introduces both intra-response and inter-response inconsistencies, undermining the expectation of coherent explanatory reasoning over time. For example, models have been shown to shift or abandon previously correct reasoning chains when faced with user disagreement or subtle framing changes [22, 23, 24, 25, 26]. Benchmarks such as SycEval [27] and BeHonest [28] quantify how sycophancy can persist across turns, leading to regressive reasoning where models rationalize incorrect answers to maintain agreement. This behaviour poses a direct challenge to explanation consistency, particularly in human-AI collaboration where users expect stable, accountable rationales for system behaviour. Techniques such as bias-augmented training or pinpoint tuning have shown promise in mitigating these effects [22, 26], but the deeper issue remains: without mechanisms to preserve a coherent explanatory stance, models risk eroding user trust even when individual outputs appear plausible. Addressing sycophancy is thus central to ensuring that explanations remain reliable not just in the moment, but across the evolving arc of user interaction.

In related work we have been looking at the way humans can explain their decisions/labels to AI on order to improve ML and XAI [29]. A key way in which these human explanations can be used is to constrain the ML system to create decision systems that respect the human explanation as well as the decision/label. This differs from the work in this paper in that the human explanations are effectively additional input to the model, whereas the coherent use of XAI explanations is in some way a feedback loop based on the current model. However, the algorithmic requirements for ensuring XAI consistency turn out to be very similar to those for the use of human explanations as input.

Coherence in AI explanations — whether provided through structured XAI methods or natural language interfaces — remains an open challenge with significant implications for trust, usability, and long-term human-AI hybridity. As AI systems become increasingly embedded in interactive settings, the ability to provide stable, transparent, and revisitable justifications will be as important as the correctness of individual decisions. This paper has presented potential strategies to implement coherence within XAI, but it is not intended to offer a final solution; our aim is to provide clarification of the area and a partial roadmap for future research. We hope this initial exploration encourages further work on algorithms, interaction strategies, and evaluation frameworks that treat coherence not as an afterthought, but as a central goal of explainable AI.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

Acknowledgments

This work has been supported by the HORIZON Europe projects TANGO - Grant Agreement n. 101120763 and SoBigData++ Grant Agreement n. 871042. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

[1] A. Dix, Human issues in the use of pattern recognition techniques, in: R. Beale, J. Finlay (Eds.), Neural Networks and Pattern Recognition in Human Computer Interaction, Ellis Horwood, 1992, pp. 429–451. URL: https://alandix.com/academic/papers/neuro92/.

[2] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013). (ICLR 2014, Workshop Poster).

[3] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[4] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[5] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, GLocalX – from local to global explanations of black box AI models, Artificial Intelligence 294 (2021) 103457.

[6] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 1–42.

[8] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–18.

[9] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, Explainable artificial intelligence: an analytical review, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1424.

[10] X. Wang, M. Yin, Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making, in: Proceedings of the 26th International Conference on Intelligent User Interfaces, 2021, pp. 318–328.

[11] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: a review on explainable artificial intelligence, Cognitive Computation 16 (2024) 45–74. URL: https://doi.org/10.1007/s12559-023-10179-8.

[12] H. Vainio-Pekka, M. O.-O. Agbese, M. Jantunen, V. Vakkuri, T. Mikkonen, R. Rousi, P. Abrahamsson, The role of explainable AI in the research field of ai ethics, ACM Transactions on Interactive Intelligent Systems 13 (2023) 1–39. URL: https://dl.acm.org/doi/10.1145/3599974.

[13] S. Myers, N. Chater, Mutual understanding initial theory, TANGO Deliverable D1.1, 2024.

[14] S. Myers, N. Chater, Interactive explainability: Black boxes, mutual understanding and what it

would really mean for AI systems to be as explainable as people, 2024. URL: osf.io/preprints/psyarxiv/ha37x_v1. doi:10.31234/osf.io/ha37x.
[15] R. Guidotti, A. Monreale, S. Ruggieri, F. Naretto, F. Turini, D. Pedreschi, F. Giannotti, Stable and actionable explanations of black-box models through factual and counterfactual rules, Data Mining and Knowledge Discovery 38 (2024) 2825–2862.
[16] F. Gawantka, F. Just, M. Savelyeva, M. Wappler, J. Lässig, A novel metric for evaluating the stability of XAI explanations, Advances in Science, Technology and Engineering Systems Journal 9 (2024) 133–142. doi:10.25046/aj090113.
[17] F. Mazzoni, R. Guidotti, A. Malizia, A Frank System for Co-Evolutionary Hybrid Decision-Making, in: International Symposium on Intelligent Data Analysis, Springer, 2024, pp. 236–248.
[18] A. Monreale, S. Teso, Cognition-aware explanations for HML, TANGO Deliverable D2.1, 2024.
[19] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery 38 (2024) 2770–2824.
[20] A. Dix, Interactive querying-locating and discovering information, in: Second Workshop on Information Retrieval and Human Computer Interaction, Glasgow, 11th September 1998, 1998. https://www.alandix.com/academic/papers/IQ98/.
[21] F. Naretto, Explainable AI methods and their interplay with privacy protection, Ph.D. thesis, Scuola Normale Superiore, 2023. URL: https://ricerca.sns.it/handle/11384/133984.
[22] J. Chua, E. Rees, H. Batra, S. R. Bowman, J. Michael, E. Perez, M. Turpin, Bias-augmented consistency training reduces biased reasoning in chain-of-thought, 2024. doi:10.48550/arXiv.2403.05518.
[23] J. Liu, A. Jain, S. Takuri, S. Vege, A. Akalin, K. Zhu, S. O'Brien, V. Sharma, TRUTH DECAY: Quantifying multi-turn sycophancy in language models, 2025. doi:10.48550/arXiv.2503.11656.
[24] Q. Xie, Z. Wang, Y. Feng, R. Xia, Ask again, then fail: Large language models' vacillations in judgment, 2024. URL: https://arxiv.org/abs/2310.02174. arXiv:2310.02174.
[25] B. Wang, X. Yue, H. Sun, Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate, 2023. URL: https://arxiv.org/abs/2305.13160. arXiv:2305.13160.
[26] W. Chen, Z. Huang, L. Xie, B. Lin, H. Li, L. Lu, X. Tian, D. Cai, Y. Zhang, W. Wang, X. Shen, J. Ye, From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning, 2024. doi:10.48550/arXiv.2409.01658.
[27] A. Fanous, J. Goldberg, A. A. Agarwal, J. Lin, A. Zhou, R. Daneshjou, O. Koyejo, SycEval: Evaluating LLM sycophancy, 2025. doi:10.48550/arXiv.2502.08177.
[28] S. Chern, Z. Hu, Y. Yang, E. Chern, Y. Guo, J. Jin, B. Wang, P. Liu, Behonest: Benchmarking honesty in large language models, 2024. URL: https://arxiv.org/abs/2406.13261. arXiv:2406.13261.
[29] A. Dix, T. Turchi, B. Wilson, A. Monreale, M. Roach, Talking Back – human input and explanations to interactive AI systems, in: Workshop on Adaptive eXplainable AI (AXAI), IUI 2025, Cagliari, Italy, 24th March 2025, 2025. URL: https://alandix.com/academic/papers/AXAI2025-talking-back/.