

Explaining Changes in Evolving Systems

Eirini Ntoutsi
Bundeswehr University Munich
Germany
Email: eirini.ntoutsi@unibw.de

Mykola Pechenizkiy
Eindhoven University of Technology
Netherlands
Email: m.pechenizkiy@tue.nl

Jerzy Stefanowski
Poznan University of Technology
Poznan, Poland
Email: jerzy.stefanowski@put.poznan.pl

Abstract—AI systems increasingly operate in non-stationary environments where data distributions, user behaviour, and system dynamics evolve over time. However, most Explainable AI (XAI) methods are designed for static settings, assuming fixed data distributions and models. We advocate for a shift from explaining predictions to explaining and understanding change. This requires explaining how and why model behaviour evolves over time to support diverse stakeholders: transparency for end users, actionable decisions such as model updates or retraining for developers, and auditing and compliance assessment over time. We propose a temporally-aware perspective and introduce a three-layer framework (data, models, explanations) for explaining change and understanding explanation drift in evolving systems.

I. INTRODUCTION

AI systems are increasingly deployed in non-stationary environments, where data distributions, user behaviour, and system dynamics evolve over time. These changes arise in a wide range of real-world applications and are commonly studied under the notion of *concept drift* [22], [8], which refers to an online supervised learning setting where the relationship between input data and the target variable changes over time.

Concept drift has been a central topic in data stream mining for over three decades¹. Research focused on *maintaining predictive performance* under data changes through incremental learning, sliding-window methods, and adaptive modifications of models. Moreover *drift detection* was studied, with methods such as DDM, EDDM, and ADWIN to determine when adaptation should occur [2]. Attention subsequently expanded to *model management*, including recurring concept detection and adaptive memory mechanisms that reuse previously learned knowledge instead of relearning models from scratch. Standardized benchmarks, prequential protocols, and frameworks such as MOA [3] enabled rigorous, reproducible comparisons, complemented by comprehensive surveys [8].

Recently, research has shifted from detecting and adapting to understanding change itself. In many applications, maintaining predictive performance is insufficient: practitioners also need to understand what changed, why it changed, which attributes or subpopulations are affected, how severe the change is, and whether adaptation is appropriate. This direction connects concept drift with explainable AI, causal

inference, uncertainty estimation, and visualization, recognizing that changes should be characterized before adaptation decisions are made [11], [25]. The latest developments extend this perspective toward trustworthy and responsible adaptation, where adaptive systems are expected to preserve not only predictive accuracy but also robustness, calibration, fairness, interpretability, privacy, and safety. Consequently, concept drift research is increasingly intersecting with continual learning, MLOps, AI governance, and Responsible AI.

Overall, the field has evolved from asking how models should adapt, to *when adaptation should occur, how knowledge should be managed, how adaptive systems should be evaluated, and ultimately what has changed and why*. Despite this progress, most research remains centered on adaptation, with comparatively limited attention to explaining changes and assessing their broader implications. This motivates the following discussion of the key limitations and failure modes of current concept drift research, highlighting the need to move beyond detecting and adapting toward explaining, evaluating, and governing adaptive AI systems.

Explainable AI (XAI) [9] has become central to ensuring transparency and trust in AI systems [6], serving not only end users but also developers and auditors responsible for the development, monitoring and governance of these systems. However, most XAI methods are fundamentally designed for static settings, where models are trained once and explanations are generated for a fixed data distribution and model [9]. When applied in evolving systems, this assumption breaks down: explanations may become stale, temporally inconsistent, or misleading as the underlying data and models evolve, while providing limited support for actionable decisions.

Existing attempts to extend XAI to evolving settings include snapshot-based analyses that repeatedly apply static explanation methods at different time points, as well as online variants of feature-importance measures [7] or wider use of the results from drift detectors [21]. While these approaches represent important first steps, they largely transfer static notions of explanation to dynamic environments rather than explicitly explaining how and why system behaviour changes over time. As a result, the temporal dynamics underlying evolving system behaviour remain largely unexplained.

We take the position that *static XAI is fundamentally misaligned with evolving systems*. Explaining predictions at a single point in time is not sufficient when both data and models

¹Several ICDM papers have significantly shaped the field, including the Best Paper Award-winning SAM-kNN [15], which introduced adaptive memory management for heterogeneous concept drift.

evolve. Instead, what is needed is *the ability to explain and understand change* - capturing *how* and *why* model behaviour evolves over time and enabling *actionable insights* for diverse stakeholders. This requires considering not only the evolution of data and models, but also explanations as first-class artifacts for transparency, auditing, compliance, and decision-making. Building on recent work on *explanation drift* in continual learning [5], we view it more broadly as systematic changes in explanations over time, arising from data evolution, model adaptation, or the explanation method itself.

To support this shift, we propose a temporally-aware, multi-layer perspective on XAI and introduce a *three-layer framework* spanning data, models, and explanations. We identify key failure modes of existing approaches and outline a research agenda for temporally-aware and actionable explanations in evolving AI systems. Our central thesis is that explaining change and understanding explanation drift should become a first-class objective of XAI research.

II. LIMITATIONS OF STATIC XAI FOR EVOLVING DATA

Most existing XAI methods are designed for static learning settings, where models are trained on fixed data distributions and explanations capture a single snapshot of model behaviour [9]. In evolving systems, however, both data and models may change over time [8], [1], giving rise to several limitations of static XAI. We identify these as failure modes (FMs), which motivate the key desiderata for temporally-aware XAI.

FM1: Temporal inconsistency (Can explanations generated at different points in time be meaningfully compared?)

Static XAI methods generate explanations independently at each point in time, without explicitly modelling temporal relationships across them. Consequently, these explanations are often difficult to compare and interpret, even when the underlying system evolves gradually.

FM2: Staleness of explanations (Does the explanation still reflect the current behaviour of the system?)

Explanations generated at a particular point in time may become outdated as data and models evolve [5], [17]. While predictive models are often updated to accommodate changing environments [8], many existing XAI approaches still treat explanations as static artifacts. This can lead stakeholders to rely on outdated interpretations of system decisions, potentially undermining trust, auditing, and decision making.

FM3: Misattribution of change (What caused the observed change in the explanation?)

Even when changes in explanations are observed, static XAI methods provide limited support for understanding their underlying causes. Similar explanation changes may arise from fundamentally different sources, including data shifts, model updates, or changes in the explanation method itself, as illustrated by the instability of feature importance [24].

FM4: Snapshot myopia (How and why did behaviour evolve between two points in time?)

Most existing XAI methods focus on explaining model behaviour at a single point in time and therefore do not capture

how or why it changes. This is particularly problematic in non-stationary environments, where understanding transitions and longer trajectories is often more important [10], [19].

FM5: Lack of actionability (What should stakeholders do in response to the observed change?)

Static explanations are typically designed to improve understanding of model decisions rather than support interventions in evolving systems. Consequently, they may provide limited guidance on how stakeholders should respond to observed changes, including whether model retraining, adaptation, monitoring, or further investigation is warranted. Thus, explanations may be informative but not actionable for a particular application [14], [16].

FM6: Adaptation lag (Can explanations keep pace with system evolution?)

In evolving environments, explanations should be generated and updated continuously. While incremental approaches such as iPFI [7] address this challenge for feature importance, many implementations of static XAI methods (e.g., those discussed in [21]) may not meet the computational and operational requirements of rapidly changing data and adaptive models. As a result, methods designed for offline analysis may become impractical in evolving environments.

FM7: Loss of longitudinal accountability (Can XAI inform why did the system change over the past period?)

Static XAI cannot answer when behaviour changed, how quickly, which update introduced the change, or whether accuracy, calibration, fairness, or other performance properties deteriorated gradually. While related to FM3, this failure mode concerns accountability for the system's evolution rather than attribution of change.

FM8: Model multiplicity (Which explanation should be trusted? Can explanations uniquely characterize system behaviour when multiple equally accurate models exist?)

In evolving environments, multiple models may achieve comparable predictive performance while relying on substantially different decision logic. Retraining, stochastic optimization, changes in the training data, or alternative adaptation strategies may therefore produce explanations that differ considerably despite similar predictive accuracy. Static XAI methods typically explain a single selected model and provide little support for reasoning about explanation variability arising from model multiplicity. Consequently, observed changes in explanations may reflect alternative but equally valid models rather than meaningful changes in the underlying concept. FM8 naturally relates to FM1–FM3: model multiplicity introduces ambiguity in attributing observed explanation changes (FM3), complicates temporal comparison when models at different time points are equally valid but structurally different (FM1), and may make explanations appear stale when an equally valid model replaces the previous one (FM2). While model multiplicity is receiving increasing attention in trustworthy AI, e.g. [12], [13], it has barely been connected to temporal explanations under concept drift.

These failure modes highlight a fundamental limitation of static XAI in evolving environments. They suggest that explaining and understanding change requires more than generating explanations at individual points in time. Instead, explanations must explicitly account for temporal dynamics and interactions across data, models, and explanations. This motivates the temporally-aware framework introduced next.

III. A TEMPORALLY-AWARE FRAMEWORK FOR EXPLAINING CHANGE

To move beyond static explanations, we require a framework that *explicitly* accounts for temporal dynamics in evolving AI systems. In such systems, changes arise not only from shifts in data distributions, but also from model updates, feedback loops, and evolving system objectives. Explanations must therefore reflect how these components evolve over time, rather than treating them as isolated, static artifacts.

We propose a *temporally-aware, multi-layer framework* for explaining change in evolving systems (Figure 1). The framework captures evolution across three *interdependent* layers: data, models, and explanations. These layers correspond to different sources and manifestations of change, and their interactions provide a more comprehensive understanding than analyzing each in isolation.

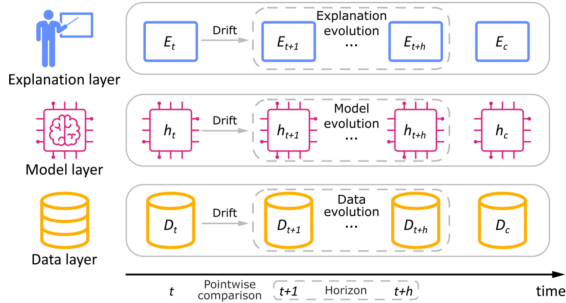


Fig. 1. Three-layer framework for explaining change, capturing the evolution of and interactions across data, models, and explanation layers.

A. Three layers of evolution

We distinguish three complementary and interdependent layers at which evolution occurs:

- **Data layer:** This layer captures changes in the input data distribution over time, including shifts in feature values, class priors, or the emergence of new patterns. Such changes are commonly studied under *concept drift* [8], [11], including changes in the input distribution (often referred to as *virtual drift*) or in the relationship between inputs and targets (*real drift*) [25]. Such changes can be analyzed using statistical and probabilistic methods, including measures of dataset similarity [20], as well as clustering and analysis of sub-concepts within classes. Understanding these changes is essential for identifying *what in the environment is changing*.

- **Model layer:** This layer reflects how predictive models evolve in response to changing data, including parameter updates, retraining, or structural modifications. Changes may manifest as shifts in model predictions, decision boundaries, or internal representations. While prediction changes are traditionally monitored using drift detectors [21], other forms of model evolution, such as structural changes, require additional tools. This layer captures *how the system adapts to change*.
- **Explanation layer:** This layer captures how explanations themselves evolve over time, for instance through changes in feature importance, decision rules, or counterfactuals. Explanations at this layer provide insight into *why model behaviour changes*, linking observed model dynamics to underlying causes. As first-class artifacts for transparency, auditing, compliance, and decision-making, explanations warrant dedicated analysis and are therefore treated as a distinct layer rather than a meta-layer.

Together, these layers capture *what changes*, *how the system adapts*, and *why behaviour evolves*.

B. Interactions across layers

While each layer can be analysed independently, changes in evolving data and systems are inherently cross-layer phenomena. Data shifts, in particular concept drifts, influence model updates, which in turn affect explanations. Conversely, changes observed in explanations may signal underlying shifts in data or model behaviour. For example, a change in feature importance (explanation layer) may result from a shift in the data distribution (data layer) or a model update or retraining (model layer) in the absence of a data shift. Identifying this requires combining signals from all layers.

Cross-layer analysis is therefore a key distinctive feature of our framework, enabling more precise attribution and a deeper understanding of evolving system behaviour.

Initial examples of joint cross-layer analysis have already emerged [23], [18]. The first combines changes in group counterfactuals with data-level information to distinguish global from local drift and identify changes in class sub-concepts, supporting subsequent modification of the decision boundary and model update [23]. The more recent STRIDE system [18] integrates feature-importance and prototype explanations with prediction-drift detectors, descriptive statistics, and clustering to interpret changes in several data-stream benchmarks.

C. Explanation drift

A central implication of our framework is that explanations may also evolve. We refer to systematic changes in explanations as *explanation drift*. Analogous to concept drift, explanation drift captures changes in how model behaviour is characterized and interpreted over time. This moves beyond "snapshot explanations" at a single point in time toward considering changes in explanations between time points or over longer horizons.

Explanation drift may manifest as shifts in representations of explanations. Such changes may arise from changes in the

underlying data, model adaptation, or instability of the explanation method itself. Recent work has investigated explanation evolution and stability in continual learning. For instance, [5] quantifies deviations between explanations produced by continual and offline models, showing that models with similar predictive performance may nevertheless produce substantially different explanations. These findings highlight the importance of explanation-aware monitoring in dynamic Denvironments.

IV. DESIDERATA FOR XAI IN EVOLVING SYSTEMS

The following properties define key requirements for XAI in evolving systems, some addressing the failure modes identified in Section II, while others reflecting stakeholder needs and operational constraints encountered in real-world deployments.

- **Temporal consistency**, enabling meaningful comparison across time, so that changes in explanations reflect actual system dynamics rather than artifacts of the explanation process itself.
- **Causal attribution of change**, identifying whether changes arise from data, models, the explanation method itself, or their interactions, supporting understanding and diagnosis of evolving behaviour.
- **Cross-layer coherence**, aligning explanations with observed data and model dynamics, ensuring that explanations remain consistent with system-level changes.
- **Actionability**, supporting decisions such as model updates, monitoring, retraining, or other interventions in response to change, enabling stakeholders to respond effectively to observed changes.
- **Granularity of explanation**, supporting multiple levels, from individual instances and groups to model regions and global behaviour, and different temporal scopes, from snapshots to change explanations and long-term profiles. This allows explanations to match different stakeholder and application needs.
- **Computational efficiency and adaptability**, ensuring that explanations can be generated and updated under system-specific constraints (e.g., limited memory and real-time processing in streams), while remaining responsive to evolving data and models and timely enough for practical use.

These properties operationalize the goal of explaining change in evolving systems, while supporting diverse stakeholders, including end users, developers, and auditors. Moreover, existing forms of explanation (e.g., attribution methods, rules, prototypes, counterfactuals, visualizations [4]) are primarily designed for static settings and may not adequately capture the temporal dynamics of explanations in evolving systems, highlighting the need for *new explanation paradigms*.

Realizing these desiderata raises several open research challenges:

- **Evaluating explanations over time**: Existing evaluation measures such as fidelity and completeness primarily target static explanations. New measures and evaluation scenarios are needed to assess explanation changes across multiple time points.

- **Cross-layer attribution of changes**: Observed changes may originate from data shifts, model adaptation, explanation instability, or their interactions. Distinguishing between these causes remains largely unexplored. Future research should investigate methods capable of attributing observed changes to their underlying sources while reasoning across data, model, and explanation layers. This includes developing causal frameworks for change attribution, identifying interactions between layers, and understanding how local changes propagate through evolving systems.
- **New tools for joint cross-layer analysis**: Analysing interdependencies among data, model, and explanation changes requires integrated tools for visualization, monitoring, and joint analysis.
- **Beyond single model explanations**: Emerging systems increasingly involve multiple interacting components, including ensembles, adaptive pipelines, foundation-model ecosystems, and multi-agent systems. Explaining change must therefore account for both the evolution of individual models and interactions among multiple adaptive components.

V. CONCLUSIONS

We argued that static XAI is fundamentally misaligned with evolving systems. While most existing explanation methods are designed to explain predictions at a single point in time, real-world AI systems continuously evolve through changing data, models, and environments. Understanding the behaviour of evolving AI systems therefore increasingly requires explaining change rather than merely explaining predictions.

To support this shift, we introduced a temporally-aware three-layer framework spanning data, model, and explanations, and outlined a research agenda for temporally-aware XAI. More fundamentally, we argue that explaining change and understanding explanation drift should become first-class objectives of XAI research. Achieving this vision will require new explanation paradigms, evaluation methodologies, and interaction mechanisms capable of supporting transparency, accountability, and actionable decision-making in evolving AI systems.

ACKNOWLEDGMENT

Jerzy Stefanowski's work was supported by the National Science Centre (Poland) grant No. 2023/51/B/ST6/00545

REFERENCES

- [1] S. Arora, R. Rani, and N. Saxena. A systematic review on detection and adaptation of concept drift in streaming data using machine learning techniques. *WIREs Data Mining and Knowledge Discovery*, 14(4):e1536, 2024.
- [2] R. S. M. Barros and S. G. T. C. Santos. A large-scale comparison of concept drift detectors. *Information Sciences*, 451-452:348–370, 2018.
- [3] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: massive online analysis. *J. Mach. Learn. Res.*, 11:1601–1604, 2010.
- [4] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, 2023.

- [5] A. Cossu, F. Spinnato, R. Guidotti, and D. Bacciu. Drifting explanations in continual learning. *Neurocomputing*, 597:127960, 2024.
- [6] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act), 2024. Official Journal of the European Union.
- [7] F. Fumagalli, M. Muschalik, E. Hüllermeier, and B. Hammer. Incremental permutation feature importance (ipfi): towards online explanations on data streams. *Machine Learning*, 112(12):4863–4903, 2023.
- [8] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37, 2014.
- [9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.
- [10] F. Hinder, V. Vaquet, J. Brinkrolf, and B. Hammer. Model-based explanations of concept drift. *Neurocomputing*, 555:126640, 2023.
- [11] F. Hinder, V. Vaquet, and B. Hammer. One or two things we know about concept drift—a survey on monitoring in evolving environments. part a: detecting concept drift. *Frontiers in Artificial Intelligence*, 7, 2024.
- [12] E. Hsu, T. Cao, L. Semenova, and C. Zhong. The rashomon set has it all: Analyzing trustworthiness of trees under multiplicity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026.
- [13] E. Hsu, H. Chen, C. Zhong, and L. Semenova. The double-edged nature of the rashomon set for trustworthy machine learning, 2025.
- [14] D. Komnick, K. Lammers, B. Hammer, V. Vaquet, and F. Hinder. Causal explanation of concept drift – a truly actionable approach. In I. Koprin-ska, J. Mendes-Moreira, and P. Branco, editors, *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 396–412, Cham, 2026. Springer Nature Switzerland.
- [15] V. Losing, B. Hammer, and H. Wersing. KNN classifier with self adjusting memory for heterogeneous concept drift. In F. Bonchi, J. Domingo-Ferrer, R. Baeza-Yates, Z. Zhou, and X. Wu, editors, *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pages 291–300. IEEE Computer Society, 2016.
- [16] M. Moshkovitz, S. Srinivas, L. Semenova, N. Frost, C. Rashtchian, V. Boreiko, S. Zhang, H. Lakkaraju, C. Rudin, and J. W. Vaughan. Position: Explainability research must prioritize foundations over ad-hoc methods. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.
- [17] M. Muschalik, F. Fumagalli, B. Hammer, and E. Hüllermeier. Agnostic explanation of model change based on feature importance. *KI-Künstliche Intelligenz*, 36(3):211–224, 2022.
- [18] W. Nagórka, M. Redmer, K. Czech, D. Aksoy, and J. Stefanowski. Stream insight and drift explanation (stride): A python toolkit for concept drift detection and explanation. In M. Baratchi and et.al, editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track, Demo Track and Industrial Track ECML PKDD 2027 Naples*, pages 348–352, Cham, 2027. Springer Nature Switzerland.
- [19] E. Ntoutsi, M. Spiliopoulou, and Y. Theodoridis. Fingerprint: Summarizing cluster evolution in dynamic environments. *International Journal of Data Warehousing and Mining (IJDWM)*, 8(3):27–44, 2012.
- [20] I. Ntoutsi, A. Kalousis, and Y. Theodoridis. A general framework for estimating similarity of datasets and decision trees: exploring semantic similarity of decision trees. In *Proceedings of the 2008 SIAM international conference on data mining*, pages 810–821. SIAM, 2008.
- [21] D. Pelosi, D. Cacciagrano, and M. Piangerelli. Explainability and interpretability in concept and data drift: A systematic literature review. *Algorithms*, 18(7), 2025.
- [22] J. Read and I. Zliobaite. Supervised learning from data streams: An overview and update. *ACM Comput. Surv.*, 57(12), July 2025.
- [23] I. Stepka and J. Stefanowski. Explaining concept drift through the evolution of group counterfactuals. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2025, Porto, Portugal, September 15–19, 2025, Revised Selected Papers, Part IV*, pages 413 – 427. Springer Nature, 2026.
- [24] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo. Statistical stability indices for LIME: obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73:91 – 101, 2022.
- [25] G. I. Webb, R. Hyde, H. Cao, H. Nguyen, and F. Petitjean. Characterizing concept drift. *Data Min. Knowl. Discov.*, 30(4):964–994, 2016.