Resilience-based framework for enhancing NaTech risk management in industrial critical infrastructures

Beyond Input Attribution: A Hands-On Tutorial to Concept-Based Explainable AI and Mechanistic Interpretability

Eliana Pastor
eliana.pastor@polito.it
Politecnico di Torino, Italy

Eleonora Poeta
eleonora.poeta@polito.it
Politecnico di Torino, Italy

André Panisson
andre.panisson@centai.eu
CENTAI Institute, Torino, Italy

Alan Perotti
alan.perotti@centai.eu
CENTAI Institute, Torino, Italy

Gabriele Ciravegna
gabriele.ciravegna@centai.eu
CENTAI Institute, Torino, Italy

Abstract

As deep learning systems become pervasive, the demand for trustworthy and transparent AI continues to grow. Traditional feature attribution methods, however, often lack robustness and alignment with human reasoning. This tutorial moves beyond feature attribution by introducing participants to two complementary interpretability paradigms: Concept-Based Explainable AI (C-XAI) and Mechanistic Interpretability. C-XAI provides explanations grounded in high-level, human-interpretable concepts, bridging the gap between model reasoning and human understanding. In parallel, mechanistic interpretability—a quickly emerging field—focuses on reverse-engineering neural networks to uncover and disentangle the internal mechanisms that give rise to human-understandable representations. Through interactive coding sessions and hands-on exercises, attendees will gain practical experience implementing, evaluating, and comparing a variety of C-XAI and mechanistic interpretability techniques. By the end of the tutorial, participants will be equipped with a modern interpretability toolbox and a deeper understanding of how to apply them in real-world scenarios.

Keywords

Concept-based Explainable AI, C-XAI, XAI

ACM Reference Format:
Eliana Pastor, Eleonora Poeta, André Panisson, Alan Perotti, and Gabriele Ciravegna. 2025. Beyond Input Attribution: A Hands-On Tutorial to Concept-Based Explainable AI and Mechanistic Interpretability. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3711896.3737606

1 Tutorial outline

- **Introduction to Explainable AI (XAI)** (15 minutes). We start with an overview of XAI and a discussion of the limitations of traditional input attribution methods, motivating the need for more human-understandable explanations.
- **Concept-Based XAI (C-XAI)** (75 minutes). We cover both by-design approaches, where models are explicitly trained to reason with concepts, and post-hoc techniques that extract concepts from pre-trained models.
- **Mechanistic Interpretability** (75 minutes). We explore methods for reverse-engineering neural networks to uncover emergent representations, often without requiring predefined human concepts or labeled data.
- **Wrap-up and open discussion** (15 minutes). We conclude with key takeaways, addressing practical considerations, and outlining future research directions.

In the following, we detail the core approaches the participants will experience during the hands-on. The material is available at https://cxai-mechint-htutorial-kdd2025.github.io/.

1.1 C-XAI Techniques

Concept-Based Explainable-by-design Models A straightforward way to create an interpretable deep learning model is to directly train it to represent human-defined concepts as intermediate representations [3, 10]. As these models are forced to directly use concepts, they provide explanations that accurately reflect their reasoning. They also support interventions, allowing users to manipulate concept values and observe the effects on predictions—a tool for debugging and building trust. In this part of the tutorial, participants will train a **Concept Bottleneck Model (CBM)** [6], where intermediate concept predictions directly influence the final output. As the requirement for concept-level annotation on training data is frequently unfeasible, researchers have started to explore the existing knowledge in pretrained large language models to automatically provide concept annotation. These models allow the creation of an explainable-by-design concept-based model without extra human annotation effort. As an example of this approach, participants will explore a **Label-Free CBMs** [9], which overcome the need for annotated concept labels by exploiting CLIP [11].

Post-Hoc Concept-Based Explanation Methods Since training a model from scratch or fine-tuning could be impractical, it is crucial to also offer explanations of existing models. Post-hoc C-XAI methods achieve this goal by projecting examples of concepts defined by humans in the latent space of the model. They analyze how these concepts align with internal representations and influence the predictions of the model. As part of this tutorial, participants will implement post-hoc techniques such as **TCAV** (Testing with Concept Activation Vectors) [5] to quantify the directional influence of human-defined concepts on model decisions.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737606

1.2 Mechanistic Interpretability

In some scenarios, we must satisfy two constraints simultaneously: working with pretrained models and avoiding manual concept annotation, while still ensuring interpretability. **Mechanistic Interpretability** addresses this challenge by revealing how neural networks internally encode high-level, symbolic concepts without requiring labeled data or retraining One key tool is **Sparse Autoencoders (SAEs)**, which learn compact and sparse representations. By enforcing sparsity, SAEs highlight the most important features, especially in unsupervised settings where labeled data is limited. To bridge the gap between the model's internal representations and human concepts, we introduce the notion of *superposition* and *monosemanticity*, which describe how SAEs map complex patterns to interpretable features. The first approach we explore is **Discover-Then-Name** [12], where latent features are first uncovered and then annotated with human-readable labels. This enables mapping of high-dimensional activations to understandable concepts—e.g., identifying objects like "cat," "car," or "tree" in image models. The second approach, **SAEuron** [2], extends this framework by enabling not only concept discovery and labeling but also *steering* of model behavior. By adjusting internal representations, SAEuron allows fine-grained control, such as refining the model's recognition of a "cat", thus supporting real-time, interpretable intervention.

2 Scope, Engagement, and Impact

Target audience and prerequisites. This tutorial is designed for researchers, machine learning engineers, and data scientists interested in advancing their understanding of explainable AI. Participants should have basic knowledge of machine learning and deep learning, familiarity with Python, and experience with frameworks like PyTorch. Hands-on exercises will be conducted using pre-prepared Python notebooks on Google Colab, ensuring an accessible and installation-free experience.

Audience Engagement. To ensure an engaging and interactive experience, the tutorial will include live coding demonstrations using Jupyter Notebooks, allowing participants to follow along with step-by-step implementations. We will encourage active participation, where attendees will experiment with models and their explanations. We will hold discussions and Q&A sessions between sections, where participants can share insights and ask questions.

Related events. Several past workshops and tutorials on XAI have been held at major conferences (e.g., [1, 4, 7, 8]). This tutorial differentiates itself by focusing on hands-on sessions on C-XAI and mechanistic interpretability, bridging the gap between high-level conceptual explanations and low-level neural mechanisms.

Societal impact. This tutorial promotes trustworthy AI by providing practical skills in concept-based and mechanistic interpretability. C-XAI improves human understanding by aligning model explanations with domain-relevant concepts, aiding in bias detection, fairness, and intervention. Mechanistic interpretability reveals how neural networks process information, helping identify learned behaviours, biases, and vulnerabilities. This contributes to safety and robustness, ensuring models operate as intended.

3 Tutors and contributors

Eliana Pastor (in-person presenter) is an assistant professor at Politecnico di Torino, Italy. Her research interests are trustworthy AI, explainable AI, and fairness in AI. She is the lecturer of the 'Explainable and Trustworthy AI' course at Politecnico di Torino.

Eleonora Poeta (contributor) is a PhD student in Trustworthy Artificial Intelligence at the Politecnico di Torino, Italy. Her research focuses on Trustworthy AI, with particular interests in explainable AI, concept-based explainability, and robustness in AI.

André Panisson (in-person presenter) is a Principal Researcher at CENTAI Institute, Italy. He leads the *Responsible AI Team* for enhancing the explainability, fairness, and transparency of Artificial Intelligence systems.

Alan Perotti (contributor) is a Senior Researcher at the CENTAI Institute, Italy. He focuses on explainability from both fundamental and applied research perspectives. He leads the development of the XAI library for Intesa Sanpaolo Bank. He will chair the XAI session at IJCNN 2025.

Gabriele Ciravegna (contributor) is a Researcher at CENTAI Institute, focused on enhancing the comprehensibility, reliability, and robustness of neural networks. Since 2019, he has been regularly publishing and reviewing for top conferences and journals like Neurips, ICML, and IEEE TPAMI.

References

[1] Gabriele Ciravegna, Mateo Espinosa Zarlenga, Pietro Barbiero, Francesco Giannini, Zohreh Shams, Damien Garreau, Mateja Jamnik, and Tania Cerquitelli. 2024. Workshop on Human-Interpretable AI. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6708–6709.

[2] Bartosz Cywiński and Kamil Deja. 2025. SAeUron: Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders. arXiv:2501.18052

[3] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. 2022. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems* 35 (2022), 21400–21413.

[4] Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. 2020. Explainable AI in industry: practical challenges and lessons learned: implications tutorial. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 699–699.

[5] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.

[6] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International conference on machine learning*. PMLR, 5338–5348.

[7] Miriam Cindy Maurer, Jacqueline Michelle Metsch, Philip Hempel, Theresa Bender, Nicolai Spicher, and Anne-Christin Hauschild. 2024. Explainable Artificial Intelligence on Biosignals for Clinical Decision Support. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[8] Mia C. Mayer, Muhammad Bilal Zafar, Luca Franceschi, and Huzefa Rangwala. 2023. Hands-on Tutorial: "Explanations in AI: Methods, Stakeholders and Pitfalls". In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[9] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. 2023. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129* (2023).

[10] Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. 2023. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936* (2023).

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[12] Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. 2024. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*. Springer, 444–461.