# Explainable Interactive Machine Learning @ UNITN

**Andrea Passerini, Stefano Teso**
DISI - University of Trento
{name.surname}@unitn.it

## Abstract

Explainable Interactive Machine Learning (XIL) aims at moving the focus of XAI from post-hoc explainability to human-in-the-loop machine learning. The idea is that the machine interacts with the user providing explanations for its predictions and receiving feedback on them, so as to guide the model toward generating increasingly better explanations and as a consequence better and more reliable predictions. XIL can substantially contribute to achieve a truly human-centric AI, paving the way to cognition-aware personal assistants, hybrid human-machine classifiers, interactive and counterfactual decision making and cooperative AI in general. The research on XIL at the University of Trento is driven by the Structured Machine Learning group (sml.disi.unitn.it) and the KnowDive group (knowdive.disi.unitn.it) in collaboration with the Mobile and Social Computing Lab at Fondazione Bruno Kessler, and it is funded by the European Projects "TAILOR: Foundations of Trustworthy AI Integrating Reasoning, Learning and Optimization" and "WeNet - Internet of Us".

Most research on explainable AI focuses on extracting explanations from black-box models, where the aim is to enrich the output of a machine learning system with information that can help explaining the reasons behind a given prediction. Typical examples include saliency maps [Adebayo *et al.*, 2018] for object detection or local interpretable surrogate models [Ribeiro *et al.*, 2016]. While such explanations can help in building trust in model predictions and occasionally identify biases that the model could have learned [Lapuschkin *et al.*, 2019], they cannot prevent the model from *learning* these biases.

Explainable Interactive Machine Learning (XIL) [Teso e Kersting, 2019; Schramowski *et al.*, 2020] has recently emerged as a paradigm aimed at filling this gap. The idea is to cast model training as an interactive process, in which the model periodically queries a domain expert providing explanation-enriched predictions and receives a feedback that is useful to correct wrong explanations in addition to wrong predictions. Consider the (in)famous example of the husky classified as a wolf because of the snow in the background. The CAIPI XIL framework [Teso e Kersting, 2019] manages to prevent this type of bias at training time, by identifying the snow as a wrong explanation for predicting wolves, and augmenting the training set with examples of wolves with the snow background replaced by random noise.

We are actively working on extending the XIL framework and adapting it to a number of relevant scenarios.

## 1 Explainable Interactive Skeptical Learning

AI-empowered personal assistants are a formidable tool that can boost our productivity, reduce our stress and improve our well-being. In order to achieve these goals, however, they need to be able to understand the specificity of each individual and adapt to the changing conditions that characterize our lives. A certain amount of human feedback is inevitable to learn accurate personalized models. While human supervision is inherently noisy, and machine learning techniques are robust to moderate amounts of noise, feedback on human behaviour is substantially more noisy and unreliable [Sorokin e Berger, 1939; Tourangeau *et al.*, 2000], and can seriously affect the quality of predictive models learned from it. We recently developed a framework named *skeptical learning* [Zeni *et al.*, 2019; Bontempelli *et al.*, 2020], that aims at interactively cleaning human feeback in sequential learning settings. The idea is that when the machine learning algorithm becomes confident enough in its own predictions, it starts challenging user feedback that conflicts with its own beliefs, potentially discovering labeling errors and inaccuracies. Explainability comes into play when the machine can motivate its suspicion with explanations supporting its own prediction. We have shown [Teso *et al.*, 2021] how influence functions can be used to identify counter-examples that conflict with the feeback received, highlighting potential inconsistencies that, if corrected, contribute to improving the quality of the data and of the learned model.

The plan is to generalize this approach to structured learning and interaction, in which the machine learning system progressively acquires structured knowledge in addition to low-level perception abilities, and interacts with the user in a structured way so as to refine the acquired knowledge and possibly adapt it to changing conditions.

## 2 XIL for Interpretable Neural Networks

One of the approaches that is gaining popularity as an alternative to post-hoc explanation of black-box models is the direct learning of interpretable neural networks [Alvarez-Melis e Jaakkola, 2018; Chen *et al.*, 2019; Koh *et al.*, 2020]. ProtoPNets [Chen *et al.*, 2019] are a successful example of this paradigm. They learn to identify prototypical parts of training images (e.g. a beak, a wing) and classify an image in terms of a linear combination of (part-)prototype similarities. A caveat of this solution is that if confounders are present in the training set, they can be easily learned as prototypes. This is a common situation in the medical domain, where this behaviour has been repeatedly observed for different types of neural architectures [Barnett *et al.*, 2021; DeGrave *et al.*, 2021]. We developed a general framework [Bontempelli *et al.*, 2021] for interactive debugging of interpretable neural networks that accommodates the most common architectures and identifies several directions for research. Our preliminary results indicate that interactive feedback on part-prototypes, coupled with constraint-enforcement strategies, can indeed prevent ProtoPNet from learning and using confounders.

Building on these initial findings, we plan to develop a general strategy capable of interactively debugging interpretable neural networks during training, and guiding them to learn semantically meaningful concepts and combine them into reliable predictions while minimizing human effort.

## 3 Disentangling Interpretable Neural Networks

Interpretable neural networks make use of specific regularization and learning techniques that encourage the latent representations or concepts learned by the model to be human-interpretable. One of the main desiderata is that these representations are *disentangled*, i.e., that the concepts themselves are causally independent so as to properly capture causally different elements of the input. The intuition is that disentangled representations make it easier to associate understandable semantics to the machine's predictions and explanations [Locatello *et al.*, 2019]. Motivated by this observation, we have designed a novel architecture that integrates approaches to disentangled representation learning – which has been explored in the context of causal representation learning [Schölkopf *et al.*, 2021] – with interpretable neural networks. Preliminary results suggest that disentanglement has the potential of helping the network to acquire more interpretable representations, which is not always the case for existing architectures [Mahinpei *et al.*, 2021], without compromising predictive performance.

Future work involves assessing the degree to which techniques for learning disentangled representations influence quality of the explanations output by interpretable neural networks, and – vice versa – to evaluate to what extent XIL can help with disentangling the learned representations.

## 4 Explainable Counterfactual Interventions

Being able to provide counterfactual interventions – sequences of actions we would have had to take for a desirable outcome to happen – is essential to explain how to change an unfavourable decision by a black-box machine learning model (e.g., being denied a loan request). Existing solutions have mainly focused on generating feasible interventions without providing explanations on their rationale. Moreover, they need to solve a separate optimization problem for each user. We took a different perspective and cast the problem as a program synthesis task. We leverage program synthesis techniques, reinforcement learning coupled with Monte Carlo Tree Search for efficient exploration [De Toni *et al.*, 2021], and rule learning to extract explanations for each recommended action. This allows us to learn a program that outputs a sequence of explainable counterfactual actions given a user description and a causal graph, and can be queries in real-time.

The next step consists in bringing the human in the loop, so as to elicit causal relationships and action costs directly from the user rather assuming to know them beforehand, thus maximizing the chance of generating the most appropriate intervention for a given user.

## 5 Challenges

Our research on XIL aims at providing useful theoretical and practical tools for boosting human-machine collaboration. On the long run, this research area has the potential to substantially change the approach people have towards AI systems, and technology in general. In order for this to happen, however, a number of challenges have to be addressed, including:

- **Countering user's ignorance** of the machine's internals, for instance by leveraging machine guidance in the form of global explanations [Popordanoska *et al.*, 2020], which will help annotators to understand what tasks the machine performs better or worse at, and therefore to select appropriate supervision or corrective feedback.
- **Cognition-aware explainability** to adapt communication strategies to the characteristics of human reasoning and its biases and to adjust them based on the specificity of the person being interacted with, from the machine learning expert to the domain expert to the man of the street.
- **Joint human-machine learning**, i.e., developing learning protocols in which *both* the human and the machine increase their knowledge of the environment and their mutual understanding, including their respective skills and limitations, so as to boost the capabilities of the resulting hybrid human-machine system.
- **Lifelong interactive learning**, in which the interaction between the human and the machine spans over multiple tasks, evolving conditions and long time horizons, with a lifespan as a final goal. While this is clearly a highly ambitious goal, with substantial technological challenges to be addressed and ethical implications to be investigated, we believe time is ripe to start thinking of human-centric AI systems with this goal in mind.

## References


[Adebayo *et al.*, 2018] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, e Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[Alvarez-Melis e Jaakkola, 2018] David Alvarez-Melis e Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[Barnett *et al.*, 2021] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, e Cynthia Rudin. IAIA-BL: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3:1061–1070, 2021.

[Bontempelli *et al.*, 2020] Andrea Bontempelli, Stefano Teso, Fausto Giunchiglia, e Andrea Passerini. Learning in the Wild with Incremental Skeptical Gaussian Processes. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 2020.

[Bontempelli *et al.*, 2021] Andrea Bontempelli, Fausto Giunchiglia, Andrea Passerini, e Stefano Teso. Toward a unified framework for debugging gray-box models. 2021.

[Chen *et al.*, 2019] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, e Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32:8930–8941, 2019.

[De Toni *et al.*, 2021] G. De Toni, L. Erculiani, e A. Passerini. Learning compositional programs with arguments and sampling. In *AIPLANS*, 2021.

[DeGrave *et al.*, 2021] Alex DeGrave, Joseph Janizek, e Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3, 07 2021.

[Koh *et al.*, 2020] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, e Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348, 2020.

[Lapuschkin *et al.*, 2019] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, e Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

[Locatello *et al.*, 2019] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, e Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[Mahinpei *et al.*, 2021] Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, e Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

[Popordanoska *et al.*, 2020] Teodora Popordanoska, Mohit Kumar, e Stefano Teso. Machine guides, human supervises: Interactive learning with global explanations. *arXiv preprint arXiv:2009.09723*, 2020.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, e Carlos 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*, page 1135–1144, 2016.

[Schölkopf *et al.*, 2021] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, e Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[Schramowski *et al.*, 2020] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, e Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

[Sorokin e Berger, 1939] Pitirim Aleksandrovich Sorokin e Clarence Quinn Berger. *Time-budgets of human behavior*, volume 2. Harvard University, 1939.

[Teso *et al.*, 2021] Stefano Teso, Andrea Bontempelli, Fausto Giunchiglia, e Andrea Passerini. Interactive label cleaning with example-based explanations. In *Proceedings of NeurIPS*, 2021.

[Teso e Kersting, 2019] Stefano Teso e Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, page 239–245, 2019.

[Tourangeau *et al.*, 2000] Roger Tourangeau, Lance J Rips, e Kenneth Rasinski. *The psychology of survey response*. Cambridge University Press, 2000.

[Zeni *et al.*, 2019] Mattia Zeni, Wanyi Zhang, Enrico Bignotti, Andrea Passerini, e Fausto Giunchiglia. Fixing mislabeling by human annotators leveraging conflict resolutio n and prior knowledge. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 2019.