Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Doctoral Consortium Track

Trustworthy AI for Post-Stroke Care

Victor F. Lopes de Souza
LIRMM, Université de Montpellier, CNRS, Montpellier, France
Euromov Digital Health in Motion, Univ. Montpellier, IMT Mines Alès, Montpellier, France
SyCoIA, IMT Mines Alès, Alès, France
victor-fernando.lopes-de-souza@umontpellier.fr

Abstract

As AI systems are increasingly deployed in high-stakes settings, two concerns become particularly important: cautiousness (awareness of model limitations for risk-aware decisions) and explainability (providing human-understandable justifications). My PhD research develops methods that enable both, from theoretical and application perspectives. On the theoretical side, I formalize cautious and interpretable models for multiple uncertainty frameworks, with algorithms for explanation and preference elicitation under uncertainty. On the application side, I implement these methods in post-stroke rehabilitation, designing AI systems that analyze multimodal clinical data to support individualized rehabilitation planning.

1 On the Research Problem

As AI systems are increasingly deployed in high-stakes settings, their decisions can have significant consequences for individuals and society. Ensuring the **trustworthiness of Artificial Intelligence** (AI) is therefore a central concern [European Comission, 2019]. Two key dimensions of trustworthy AI are **cautiousness**, awareness of model limitations that supports risk-aware decision-making [Dubois and Prade, 2009], and **explainability**, the ability to provide human-understandable justifications for model decisions [Barredo Arrieta *et al.*, 2020]. Both dimensions are especially critical in safety-sensitive applications. My PhD addresses this challenge along two complementary axes: theory and application.

On the theoretical side, **I develop methods that enable cautiousness and explainability in models, both through mathematical formalisation and by implementing new algorithms**. Cautiousness relies on uncertainty quantification, which can be modelled in several frameworks, including imprecise probability theory [Walley, 1991], possibility theory [Dubois and Prade, 1988], rough sets [Pawlak, 1982], fuzzy sets [Zadeh, 1965], and Dempster–Shafer evidence theory [Shafer, 1976]. Explainability enables users to understand, critique, and improve models. A common taxonomy distinguishes intrinsic methods—models designed to be interpretable—from post-hoc methods—explanations produced for already-trained black boxes [Carvalho *et al.*, 2019]. Post-hoc techniques mainly fall into two families [Barredo Arrieta *et al.*, 2020]: feature-relevance methods, which rank or quantify the influence of input features [Baehrens *et al.*, 2010], and simplification methods, which approximate black-box classifiers with interpretable surrogates such as decision trees, rule lists, or linear models [Guidotti *et al.*, 2018].

On the application side, the focus of my work is on **post-stroke rehabilitation**, where trustworthy AI can directly support clinical decisions. My research is conducted within the ReArm project, a quadruple-blind, randomised, sham-controlled, bi-centre, two-arm interventional study involving fifty-eight chronic stroke patients, designed to assess a treatment that combines virtual reality therapy with anodal transcranial direct current stimulation. The collected data include clinical assessments, EEG and fNIRS recordings, and continuous upper-limb monitoring in daily life using wearable accelerometers [Muller *et al.*, 2021]. The goal of the application part of my PhD is to design trustworthy AI systems that exploit this multimodal information to produce personalised insights and support individualised rehabilitation planning.

2 Contributions

I summarize the main contributions of my PhD research:

- **Explaining evidential clusterings with IEMM** (best poster award at BFTA 2025; paper under review at ECML-PKDD 2026, preprint available [de Souza *et al.*, 2025]). We formalise *representativeness*, a utility-based criterion that encodes decision-makers' preferences over explanation misassignments, and *evidential Mistakeness*, a loss tailored to credal partitions. Building on these notions, the Iterative Evidential Mistakeness Minimization (IEMM) algorithm learns decision-tree surrogates for evidential clustering; we also provide conditions for effective explanations in both hard and evidential settings and validate the approach on synthetic and real-world datasets.
- **Personalized Activity Profiling in Post-Stroke Care via Explainable Evidential Clustering** (accepted at FUZZ-IEEE WCCI 2026). We design an explainable clustering pipeline for wrist-accelerometer data: five-minute windows are transformed into clinically mean-

ingful arm-use features and clustered using either (i) Evidential C-Means with an IEMM decision-tree surrogate or (ii) K-means with a Belief Rule-Based Classification System. We then combine clustering results with patient-reported activity logs using evidence theory to derive activity-level profiles with associated uncertainty, producing clinically interpretable insights on real-world post-stroke data. We compare the two pipelines and discuss their respective advantages and limitations.

- **Eliciting utilities under imprecision** (under review at UAI 2026). We propose practical elicitation procedures that quantify how much imprecision a user is willing to accept when receiving set-valued predictions, i.e., how much precision to trade for robustness and accuracy. Our proposed iterative algorithm converges to the target parameter on both tabular and image-based benchmarks with a limited number of queries.
- **Counterfactuals and adversarial examples for set-valued classifiers** (accepted at IPMU 2026). We generalize counterfactual explanations through *foiled predicates* that specify conditions the current set-valued prediction fails to satisfy. We also define adversarial examples for set-valued outputs, distinguishing *possible* adversarial examples that induce uncertainty about the true class from *certain* adversarial examples that guarantee misclassification, and we illustrate the framework on different datasets.

3 Impact and Next Steps

Trustworthy AI ultimately aims to support human-machine cooperation: models should not merely output decisions but provide the elements required for users to critique them. In this view, mistakes are not only risks to mitigate, but can also trigger a deeper understanding of the data, modelling assumptions, and domain knowledge. Designing systems that can be interpreted, meaningfully challenged, and ultimately trusted is therefore a prerequisite for responsible deployment in safety-sensitive settings.

In post-stroke rehabilitation, this requirement is particularly important. Stroke remains a major cause of long-term disability and reduced autonomy in daily activities, and decision-support tools can only be useful if they are clinically sound and aligned with medical practice. For this reason, continuous dialogue with clinical experts is central in my work, both to validate the relevance of the extracted patterns and to ensure that the proposed explanations and uncertainty representations match practitioners' needs.

Next, I will strengthen the methodological foundations and broaden the scope of the application. This includes advancing the elicitation of cautious utilities, extending cautious and explainable approaches to multimodal rehabilitation data, and developing cautious explainable regression methods to identify factors that are associated with better post-stroke recovery outcomes.

References

[Baehrens *et al.*, 2010] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.*, 11:1803–1831, August 2010.

[Barredo Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020.

[Carvalho *et al.*, 2019] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, August 2019.

[de Souza *et al.*, 2025] Victor F. Lopes de Souza, Karima Bakhti, Sofiane Ramdani, Denis Mottet, and Abdelhak Imoussaten. Explainable Evidential Clustering, July 2025.

[Dubois and Prade, 1988] Didler Dubois and Henri Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4(3):244–264, 1988.

[Dubois and Prade, 2009] Didier Dubois and Henri Prade. Formal Representations of Uncertainty. In Denis Bouyssou, Didier Dubois, Marc Pirlot, and Henri Prade, editors, *Decision-making Process*, pages 85–156. Wiley, 1 edition, January 2009.

[European Comission, 2019] EU European Comission. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019. Accessed: 2026-06-17.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5):93:1–93:42, August 2018.

[Muller *et al.*, 2021] Camille O. Muller, Makii Muthalib, Denis Mottet, Stéphane Perrey, Gérard Dray, Marion Delorme, Claire Duflos, Jérôme Froger, Binbin Xu, Germain Faity, Simon Pla, Pierre Jean, Isabelle Laffont, and Karima K. A. Bakhti. Recovering arm function in chronic stroke patients using combined anodal HD-tDCS and virtual reality therapy (ReArm): A study protocol for a randomized controlled trial. *Trials*, 22(1):747, October 2021.

[Pawlak, 1982] Zdzisław Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11(5):341–356, October 1982.

[Shafer, 1976] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, April 1976.

[Walley, 1991] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall, 1991.

[Zadeh, 1965] Lotfi Aliasker Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, June 1965.