Towards Cognitive Representatives

Research Proposal · June 2023

DOI: 10.13140/RG.2.2.30989.72161

2 authors:

Loizos Michael

126 PUBLICATIONS **903** CITATIONS

SEE PROFILE

Vasileios Theodoros Markos

Open University of Cyprus

8 PUBLICATIONS **10** CITATIONS

SEE PROFILE

Towards Cognitive Representatives

Vassilis Markos[1] , **Loizos Michael**[1,2]
[1]Open University of Cyprus
[2]CYENS Center of Excellence
vasileios.markos@st.ouc.ac.cy, loizos@ouc.ac.cy

Abstract

Given the ever-growing use of Artificial Intelligence (AI) in everyday life, being capable of explaining a model's reasoning has become of crucial importance. While in most cases AI serves as a *cognitive replacement* for humans, a paradigm shift to *cognitive representation*, where AI representatives express their representee's views on a certain context in a faithful manner, seems necessary to accommodate any ethical and legal demands. In that regard, we present our plan of research, our progress so far and what remains to be completed.

1 Problem and Motivation

From recognizing cats in videos (Le et al. 2011) to analyzing and searching through big text corpora (Christensen et al. 2017) and mastering the games of Chess and Go (Silver et al. 2017), Artificial Intelligence (AI) has found applications in an overwhelming variety of domains, focusing increasingly more on our everyday lives (e.g., Finance (Sadgali, Sael, and Benabbou 2019), Healthcare (Ahmad, Teredesai, and Eckert 2018), Education (Ciolacu et al. 2017)). In most cases, AI adopts the role of a *cognitive replacement*, substituting humans in tasks they used to do themselves. The way such replacements are usually built, by training on massive datasets, makes them difficult to interpret and remain faithful to their representees, being usually oblivious to their domain expertise, making a case for a paradigm shift towards *cognitive representation*, i.e., the use of AI models as agents that act on behalf of their representees, respecting their preferences and being amenable to changes when required to. To that direction, we present three relevant research questions we aim to explore:

Q1. *Is it possible to design a formally guaranteed dialectical process of machine knowledge adaptation, with the aims of boosting performance, learning speed, and explainability?* In other words, it is of our interest to explore how one could implement and design cognitive representatives by utilizing and adapting eXplaibable AI (XAI) solutions in that regard.

Q2. *How can a process as described in Q1 be supported in terms of interface design?* Apart from exploring theoretical perspectives related to the design of cognitive representation, it is of major importance to determine how those are reflected in design choices, with the aims of improving user experience, and usability, in general.

Q3. *How could one integrate feedback from multiple resources in a diversity-aware manner to build collective representatives?* There are cases where eliciting individual knowledge and expertise does not suffice to achieve an acceptable performance. To that extent, it is worth exploring how one could integrate potentially conflicting knowledge to a single corpus in a way that respects the diversity of its contributing sources.

In the spirit of McCarthy's "Advice Taker" (McCarthy 1959), we propose *Machine Coaching* (Michael 2019), an advice-based interaction, which relies on bilateral explanation exchange between a human coach and a machine learner with the aim of explicating the former's preferences to the latter and fine-tuning its performance. The structure of the rest of this summary is as follows: (i) in Section 2 we briefly review related literature; (ii) in Section 3 we present our plan of research; (iii) in Section 4 we present our progress to date.

2 Related Literature

Works in Explainable AI (XAI) might be split into two broad categories with respect to how they bind explanations to learning: (i) *post-hoc* methodologies seek to explain a black-box's outputs after the learning process has been completed, by tracing back the root causes of that output while; (ii) *interactive* methodologies seek to interwine explanation giving / taking with the learning process. Probably the most popular XAI methodologies come from the first category, including LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017), both being capable of providing localized explanations for virtually any machine learning model. This versatility has led to the widespread adoption of post-hoc over interactive XAI methodologies, which however comes at a significant cost. Typically, offering explanations means they are *compatible* with and *comprehensible* to their consumers, in the sense of resembling a theoretically or empirically expected (human-like) behavior (Albini et al. 2020; Bach et al. 2015; Barredo Arrieta et al. 2020), seeking to be *stable*, in the sense of not varying significantly on similar contexts (Tsang et al. 2018; Alvarez-Melis and Jaakkola 2018), and boost users' *trust* on the model (Gan et al. 2015; Montavon, Samek, and Müller

2018). On top of those, post-hoc methodologies also need to maximize *faithfulness* to the underlying model (Balog, Radlinski, and Arakelyan 2019), i.e., any explanation offered should reflect the model's internal reasoning up to some significant extent and not be just a sound but potentially irrelevant justification of its decision (Ribeiro, Singh, and Guestrin 2016; Barredo Arrieta et al. 2020).

Echoing ideas from McCarthy's "Advice Taker" (McCarthy 1959), one expects that any difficulties arising from introducing post-hoc explanations in opaque machine learning models should be lifted by allowing for a machine to be tutored by a more experienced human coach through iterative advice taking. As discussed above, there have been attempts to formalize such a procedure (Michael 2017; Michael 2019) resulting to *Machine Coaching*, a learning framework where a human coach progressively refines a machine's knowledge by providing advice to it on certain occasions. Learnability in this context is conceptualized within the Probably Approximately Correct (PAC) semantics (Valiant 1984) while explainability is achieved by allowing the machine to directly expose (parts of) its internal reasoning process as faithful explanations of its decisions. Thus, the machine is inherently transparent with respect to its internal reasoning, while its explanations, relying on human-provided input, are good candidates for compatible and comprehensible explanations.

Despite the prevalence of post-hoc explanation methodologies, the inherent merits of interactive XAI models have led to a growing interest in the last years. Explainable Interactive Learning (XIL) (Teso and Kersting 2019) is a framework that considers a human supervisor in the role of a learning "coach", as with Machine Coaching. In that context, the supervisor provides advice by improving predicted labels and / or the corresponding explanations, effectively distinguishing, thus, between wrong responses and correct responses, but for the wrong reasons. While Machine Coaching relies on Computational Argumentation to provide the underlying reasoning machinery, XIL assumes black-box access to an arbitrary active learning algorithm (Settles 2009) accompanied by a local post-hoc explainer, such as LIME (Ribeiro, Singh, and Guestrin 2016). As a result, any explanations generated by XIL are not part of the underlying black-box model, but essentially provide additional labels to the already labeled training data, leaving the issue of faithfulness unaddressed for XIL.

Apart from ensuring faithfulness and effectively building user trust, ingesting explanations within the training phase of a model has been found to improve its final performance (Selvaraju et al. 2019), while it has also been reported that bilateral explanations from and towards the learner may also speed up learning (Dong et al. 2017). Similar boosts have also been observed in cases where a model is retrained utilizing explanations (Rieger et al. 2019), while post-hoc fine tuning of a model's functionality has also been observed to improve its performance in (Ribeiro, Singh, and Guestrin 2016). Also, closely related to Machine Coaching at an abstract level, *Coactive Learning*, as described in (Shivaswamy and Joachims 2015), describes a learning process where the learner is offered sub-optimal advice by its supervisor with the aim to provide just a "slight improvement" compared to the learner's former predictive accuracy. Its relative simplicity allows for Coactive Learning to be accommodated by most widely used supervised machine learning algorithms, allowing one to maintain their merits, while also making the resulting models more transparent.

It is also the case that the provision of human advice during learning also reduces any demand for labeled data, as shown in (Raghavan and Allan 2007), where human supervisors' advice is embedded into the training of Support Vector Machines. In a similar fashion, in (Settles 2011), an interactive learning framework that allows the labeling of certain data points chosen by human supervisors also appears to minimize any needs for labeled data. Another approach in the same direction, with advice in the form of user-crafted explanations, is presented in (Zaidan, Eisner, and Piatko 2007). There it is shown that, in certain contexts, providing access to richer data points can lead to better performance while resulting to a more explainable and, thus, trustworthy model, compared to providing access to a significantly larger volume of training less expressively labeled data.

3 Plan of Research

Our plan of research is split in seven tasks, as follows:

T1. **Exploration.** This task includes the theoretical and empirical exploration of Machine Coaching as described in (Michael 2019) and the implementation of any related functionalities. Also, T1 includes the development of a language that can accommodate knowledge representation, as described in (Michael 2019).

T2. **In vitro assessment.** This task includes the assessment of any functionalities developed in T1 in artificial settings, i.e., using artificial data and simulated coaches in the stead of human ones. The main goal of T2 is to investigate the efficiency, efficacy and performance of any developed infrastructure.

T3. **Refinement I.** Any infrastructure developed in T1 and assessed during T2 will be revised and improved under the lights of any obtained results.

T4. **Application.** Preparing for T5, we consider realistic settings in which Machine Coaching could provide added value compared to other AI methodologies.

T5. **In situ assessment.** Having determined candidate fields of application, we design solutions that rely on Machine Coaching for them. Then, we proceed to assess them *in situ*, focusing mainly on efficacy and efficiency, as in T2, but also on user experience and overall usability, given the prevalent role human coaches play in Machine Coaching.

T6. **Refinement II.** For each determined field of application, we aim to properly utilize any results obtained in T5 with the aim of improving our approach locally (i.e., for that specific application) as well as globally (i.e., for Machine Coaching as an abstract learning framewok).

T7. **Large-scale deployment.** Given the completion of all previous tasks, T1-6, we plan to offer any developed

functionalities in a uniform way through APIs, to facilitate the inclusion of Machine Coaching in applications, as well as developing an ecosystem to allow the design of and interaction between cognitive representatives by non-expert users.

Tasks T1-3 are mostly related to research question Q1, mainly addressing the theoretical and implementation aspects stressed there. Similarly, tasks T4-6 are related to Q2, putting more focus on how any Q1-related findings can influence the design of an interface facilitating the development of cognitive representatives in certain contexts. Lastly, T7 is related to Q3, since extending our efforts at a broader scale requires taking more complex dynamics into account.

4 Progress to Date

At the time of writing this, tasks T1 and T2 have been completed to their most part. Namely, we have designed *Prudens*, an argumentation-based language allowing for reasoning as described in (Michael 2019) to take place (Markos and Michael 2022b). Prudens offers the following syntactic constructs: (i) ***constants***, which correspond to entities of the universe of discourse; (ii) ***variables***, which serve as placeholders for constants; (iii) first-order ***predicates*** (with variables and/or constants as arguments) and propositional ones, capturing relations and conditions about the universe of discourse, respectively; (iv) ***literals***, which are either predicates themselves or negated, with negation being used in the classical sense, (two literals corresponding to the same predicate but with opposite signs are said to be ***conflicting***); (v) if-then ***rules***, which connect a set of premises, the rules' ***body***, with a single literal, the rules' ***head*** (as with literals, two rules with conflicting heads are said to be ***conflicting***); (vi) ***policies***, which are lists comprising of rules alongside a priority relation defined over all pairs of conflicting rules, which is by default determined by the rules' order of appearance, i.e., the later a rule appears in the policy, the higher its priority over conflicting ones is; (vii) ***contexts***, which are sets of *pairwise non-conflicting* literals, corresponding to a set of facts being known at the beginning of the reasoning process.

Regarding T2, Prudens's efficiency and efficacy have been systematically assessed in artificial settings (Markos, Thoma, and Michael 2022). Namely, we have designed and implemented artificial coaches using other machine learning models and paradigms (decision trees / forests and evolutionary learning), and have had them coach agents on artificially generated policies. Results from those experiments provided empirical validation for the theoretically guaranteed efficiency and efficacy of Machine Coaching under certain assumptions. Also, given that all policies in (Markos, Thoma, and Michael 2022) were propositional, we have positive evidence regarding scalability in that setting. What we still have to accomplish within T2 is an assessment of scalability when it comes to first-order policies, on which, we are currently working on with simulations on synthetic data.

Regarding T3, we are currently working on improving Prudens in terms of time efficiency regarding first-order policies, while also making certain syntactic extensions that have proven useful throughout T2. Regarding T4, we have chosen Othello (Reversi), a two-player board game, as a first domain of application of Machine Coaching. We have implemented an interface[1] that allows coaching in the context of the game through patterns, expected to facilitate a broader and non-expert audience to contribute as coaches. We have also contacted game experts to get feedback about interface design and improving the coaching process.

While there has not been any work on T5 and T6 yet, we have started working on building any needed infrastructure, based on our currently available functionalities related to Prudens, as a part of T7. Namely, we are in the process of building the corresponding APIs and gradually integrating any new features there, as they come. In parallel, we have also been working on diversity-related issues, including how one could efficiently diversify a ranked list to balance between user preferences and inclusion (Markos and Michael 2022a), and, recently, diversity perception. Regarding the latter, we have conducted a study to explore factors that might affect perceived diversity in given settings, unveiling that most widely used diversity metrics are oblivious to several factors that influence individual perceptions of diversity, making a case for the need for more informed metrics to be designed. To that extent, it might be interesting to pursue ways in which Machine Coaching could facilitate the elicitation of an individualized metric of diversity either by explicit coaching, or implicitly, by first passively observing user behavior and then using gathered data as a proxy coach.

Acknowledgements This work was supported by funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements no. 739578 and no. 823783, and from the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation, and Digital Policy.

References

Ahmad, M. A.; Teredesai, A.; and Eckert, C. 2018. Interpretable Machine Learning in Healthcare. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 447–447.

Albini, E.; Lertvittayakumjorn, P.; Rago, A.; and Toni, F. 2020. DAX: Deep Argumentative eXplanation for Neural Networks. *ArXiv* abs/2012.05766.

Alvarez-Melis, D., and Jaakkola, T. S. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* 10(7):1–46.

Balog, K.; Radlinski, F.; and Arakelyan, S. 2019. Transparent, Scrutable and Explainable User Models for Personalized Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, 265–274. New York, NY, USA: Association for Computing Machinery.

[1] Available at vmarkos.github.io/coachello.

Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58:82–115.

Christensen, K.; Nørskov, S.; Frederiksen, L.; and Scholderer, J. 2017. In Search of New Product Ideas: Identifying Ideas in Online Communities by Machine Learning and Text Mining. *Creativity and Innovation Management* 26(1):17–30.

Ciolacu, M.; Tehrani, A. F.; Beer, R.; and Popp, H. 2017. Education 4.0 — Fostering Student's Performance with Machine Learning Methods. In *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)*, 438–443.

Dong, Y.; Su, H.; Zhu, J.; and Zhang, B. 2017. Improving Interpretability of Deep Neural Networks with Semantic Information.

Gan, C.; Wang, N.; Yang, Y.; Yeung, D.-Y.; and Hauptmann, A. G. 2015. DevNet: A Deep Event Network for multimedia event detection and evidence recounting. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2568–2577.

Le, Q. V.; Ranzato, M.; Monga, R.; Devin, M.; Chen, K.; Corrado, G. S.; Dean, J.; and Ng, A. Y. 2011. Building High-level Features Using Large Scale Unsupervised Learning.

Lundberg, S., and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions.

Markos, V., and Michael, L. 2022a. Post-hoc diversity-aware curation of rankings. In *International Conference on Agents and Artificial Intelligence*.

Markos, V., and Michael, L. 2022b. Prudens: An argumentation-based language for cognitive assistants. In Governatori, G., and Turhan, A.-Y., eds., *Rules and Reasoning*, 296–304. Cham: Springer International Publishing.

Markos, V.; Thoma, M.; and Michael, L. 2022. Machine coaching with proxy coaches. In *1st International Workshop on Argumentation and Machine Learning — COMMA 2022*. CEUR-WS.

McCarthy, J. 1959. Programs with Common Sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, 75–91.

Michael, L. 2017. The Advice Taker 2.0. In *Proceedings of the 13th International Symposium on Commonsense Reasoning*, volume 2052.

Michael, L. 2019. Machine Coaching. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI @ IJCAI 2019)*, 80–86.

Montavon, G.; Samek, W.; and Müller, K.-R. 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing* 73:1–15.

Raghavan, H., and Allan, J. 2007. An Interactive Algorithm for Asking and Incorporating Feature Feedback into Support Vector Machines. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, 79–86. New York, NY, USA: Association for Computing Machinery.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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, 1135–1144. New York, NY, USA: Association for Computing Machinery.

Rieger, L.; Singh, C.; Murdoch, W. J.; and Yu, B. 2019. Interpretations are Useful: Penalizing Explanations to Align Neural Networks With Prior Knowledge.

Sadgali, I.; Sael, N.; and Benabbou, F. 2019. Performance of Machine Learning Techniques in the Detection of Financial Frauds. *Procedia Computer Science* 148:45–54. The Second International Conference On Intelligent Computing In Data Sciences, ICDS2018.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128(2):336–359.

Settles, B. 2009. Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences.

Settles, B. 2011. Closing the Loop: Fast, Interactive Semi-Supervised Annotation With Queries on Features and Instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1467–1478. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Shivaswamy, P., and Joachims, T. 2015. Coactive Learning. *Journal of Artificial Intelligence Research* 53:1–40.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; and Hassabis, D. 2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm.

Teso, S., and Kersting, K. 2019. Explanatory Interactive Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 239–245. New York, NY, USA: Association for Computing Machinery.

Tsang, M.; Sun, Y.; Ren, D.; and Liu, Y. 2018. Can I trust you more? Model-Agnostic Hierarchical Explanations.

Valiant, L. G. 1984. A Theory of the Learnable. *Communications of the ACM* 27(11):1134–1142.

Zaidan, O.; Eisner, J.; and Piatko, C. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 260–267. Rochester, New York: Association for Computational Linguistics.

Curriculum Vitae

Personal Information

First Name:	Vasileios
Middle Name:	Theodoros
Last Name:	Markos
Date of Birth:	1996, March 2nd
Address:	Odemisiou 16-20, 16231
Country:	Greece

Studies

2020 — today	PhD, Cognitive Systems, Open University of Cyprus (OUC)
2018 — 2020	MSc, Cognitive Systems, (OUC)
2014 — 2018	BSc, Mathematics, University of Athens (UoA)

Work Experience

2019 — today	Research & Development on XAI related topics (Project: WeNet)
2021 — 2022	Research & Development on XAI related topics (Project: MARI-Sense)
2018 — 2021	Mathematics & Computer Science Teacher (private sector after-school support)
2017 — today	Mathematics & Computer Science Teacher (*volunteering* on social structures, after-school support)

Personal Skills

Languages	Greek (Mother Tongue)
	English (C2)
	German (C1)
Programming	Java
	JavaScript
	HTML/CSS
	Python
	Prolog
Soft Skills	Team coordination
	Presentation skills
	General communication skills

Published Work

Markos, V., & Michael, L. (2022b). Prudens: An Argumentation-Based Language for Cognitive Assistants. In G. Governatori & A.-Y. Turhan (Eds.), Rules and Reasoning (pp. 296–304). Cham: Springer International Publishing.

Markos, V., Marios, T., & Michael, L. (2022). Machine Coaching with Proxy Coaches. 1st International Workshop on Argumentation and Machine Learning — COMMA 2022. CEUR-WS.

Markos, V., & Michael, L. (2022a). Post-hoc Diversity-aware Curation of Rankings. International Conference on Agents and Artificial Intelligence.

Markos, V. (2020). Application of the Machine Coaching Paradigm on Chess Coaching. School of Pure & Applied Sciences, Open University of Cyprus.

Miscellaneous

Competitions	Blue Innovation 2021 Hackathon (website), member of the winning team
Hobbies	Music (violin/viola)
	Creative writing

Letter of Recommendation

This is to state that Vassilis Markos is a Ph.D. student under my supervision at Open University of Cyprus, and that I support his participation in the KR'23 doctoral consortium.

Loizos Michael
Associate Professor
Open University of Cyprus