Towards Robust Contrastive Explanations for Human-Neural Multi-Agent Systems

Extended Abstract

Francesco Leofante
Imperial College London
London, United Kingdom
f.leofante@imperial.ac.uk

Alessio Lomuscio
Imperial College London
London, United Kingdom
a.lomuscio@imperial.ac.uk

ABSTRACT

Generating explanations of high quality is fundamental to the development of trustworthy human-AI interactions. We here study the problem of generating contrastive explanations with formal robustness guarantees. We formalise a new notion of robustness and introduce two novel verification-based algorithms to (i) identify non-robust explanations generated by other methods and (ii) generate contrastive explanations augmented with provable robustness certificates. We present an implementation and evaluate the utility of the approach on two case studies concerning neural agents trained on credit scoring and image classification tasks.

KEYWORDS

Explainable AI, Formal Verification

ACM Reference Format:
Francesco Leofante and Alessio Lomuscio. 2023. Towards Robust Contrastive Explanations for Human-Neural Multi-Agent Systems: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

The forthcoming adoption of AI in modern societies has lead to the emergence of sophisticated multi-agent systems in which humans and artificial agents interact and collaborate [2, 14, 24]. Advances in deep learning [19] have facilitated the development of neural agents governed by neural networks (NNs) synthesised from data [1]. We call Human-Neural Multi-Agent System (HNMAS) a system composed by humans and neural agents interacting and communicating in view of achieving common goals. While HNMAS may offer rapid gains in terms of performance and generalisation, neural agents are known to produce outputs that are not normally intelligible to humans, thus hindering the development and deployment of HNMAS that can be trusted by human agents.

The area of Explainable AI (XAI) is concerned with making NNs, and other learned models, more understandable to humans. A widely recognised factor contributing towards this goal is the availability of *contrastive explanations* (CEs), i.e., semi-factual (SF) [17] and counterfactual (CF) [21] arguments supporting or contrasting the decisions taken by an NN. Crucially, CEs are typically used to provide recourse to individuals that have been impacted by the decisions of an AI. Several approaches have been proposed to compute CEs for NNs according to different quality criteria, such as *validity* and *proximity* [29], *plausibility* [17] and *actionability* [28]. Our focus here is the criterion of *robustness* [4, 7, 8, 15, 25, 27]; in particular, we study *robustness to noisy execution*. Current algorithms generate explanations under the assumption that the human receiving a recourse recommendation will implement it exactly. However, several studies have reported that this rarely happens in practical applications [3, 22]. The noise introduced may jeopardise the validity of CEs, ultimately reducing the trust humans put into their neural agent counterpart. To remedy this, we draw from the literature on (local) robustness verification of NNs [5, 10, 12, 12, 16, 18, 23, 30] and propose a novel approach to generate CEs that are robust to noisy execution.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

2 ROBUST EXPLANATIONS VIA VERIFICATION

The notion of robustness that we study is tied to variations of a CE's classification with respect to changes applied to the CE itself. Intuitively, if a CE is modified slightly then the classification provided by the classifier for that new input should not change radically. If this property does not hold, it is likely to signify that the CE is an artefact of the NN and does not represent, nor explain, its underlying classification logic [11]. Furthermore, this lack of robustness is not in line with human intuition and expectations, which ultimately weakens the power of CEs within HNMAS.

Contributions. We begin by formalising the notion of robustness to noisy execution which we target in this work.

Definition 1. *Consider an input x_F and a binary neural network classifier f such that $f(x) = 0$. Let x be a CF (resp. SF) explanation computed for x_F s.t. $f(x) = 1$ (resp. $f(x) = 0$). The CF (resp. SF) explanation x is said to be* robust to noisy execution up to magnitude *δ if for all inputs x' such that $\|x' - x\|_\infty \leq \delta$, we have that $f(x') = 1$ (resp. $f(x') = 0$).*

In a nutshell, Def. 1 requires that explanations remain valid across a (reasonably-sized) neighbourhood. This is to ensure that they cannot be invalidated by small noise introduced by humans when implementing recourse recommendations.

We then propose an approach based on formal verification of neural networks to mechanise the analysis and discovery of contrastive explanations and associated robustness. The approach relies on solving the following two problems:

(1) **Prove that a (possibly non-robust) CE exists for a given input**. We show that answering this question yields an NP-complete problem, which can be recast as a verification problem and solved using any (complete) verification procedure.

(2) **Prove that a CE is robust to noisy execution**. We show that answering this question yields a coNP-complete problem, which again can be solved using verification techniques.

3 EXPERIMENTAL EVALUATION

We use the results above to derive algorithms to *(i)* determine whether explanations generated by other methods are robust for user-defined δ's and *(ii)* generate robust explanations. We apply these algorithms on different input data types (tabular and images) and neural architectures (fully-connected and convolutional). Our approach can be instantiated with any complete neural network verifier; the current implementation leverages Venus [5] and VeriNet [13]. Both verifiers are used as black-boxes from their user interface; we refer to the respective papers for more details. We evaluated our approach on two case studies involving neural agents trained to perform credit scoring and traffic sign recognition.

Automated credit scoring. We consider the verification and generation of robust CF explanations for a neural agent trained to perform credit scoring tasks based on the HELOC dataset [9].

Experiment 1. We used verification techniques to check the robustness of heuristically-computed explanations. For these experiments we considered Contrastive Explanation Method (CEM) [6], a popular algorithm which uses a gradient-based search to compute SF and CF explanations. Our aim is to understand the extent to which explanations provided by CEM are robust. Given a contrastive explanation x, a neural classifier f and a robustness threshold δ, we formulate a verification query to establish whether f satisfies local robustness for x and δ. If local robustness is satisfied, then the explanation is guaranteed to be δ-robust. Otherwise, a counterexample can be returned that identifies an input for which the explanation is invalidated. Overall, we observed that both SFEs and CFEs obtained via CEM are robust for small δ's. However, their robustness decreases when considering larger domains, revealing that most of the explanations proposed by the tool are not robust. This is understandable as CEM is not designed to generate robust explanations; however users have no way to determine the extent to which an explanation is robust.

Experiment 2. We also present a procedure to generate robustness guarantees for CF explanations. The algorithm receives an input x_F, a neural classifier f and a robustness threshold δ. The overall aim of the algorithm is to first decide whether a CF explanation x exists; if one can be found, the algorithm operates further steps to quantify its robustness. More specifically, a binary search is performed to find the largest δ across which x is robust. At each step of the search, a verification query checking the robustness of the explanations is performed until either the largest robust δ is found or a termination condition is reached. Our experiments reveal that the explanations generated are characterised by small robustness thresholds on average; this information can be used by regulators and users alike to select only explanations that are robust for larger δ's, and filter out others that may be problematic.

Experiment 3. We performed additional experiments and modified our procedure to account for actionability (as well as robustness). Actionability is typically enforced by allowing changes only on input features which are classified as mutable a-priori (e.g., the education level of an applicant may change while their ethnicity may not). Such domain knowledge can be seamlessly incorporated into our framework. We validated the ability of our approach to generate robust, actionable CF explanations for the HELOC dataset. Our results have important practical implications: our explanations suggest changes that are achievable in practice and are formally guaranteed to yield the expected outcome for any slight perturbation of magnitude less than the robustness threshold identified. We see these results as an important contribution toward complementing existing formal approaches for XAI [20].

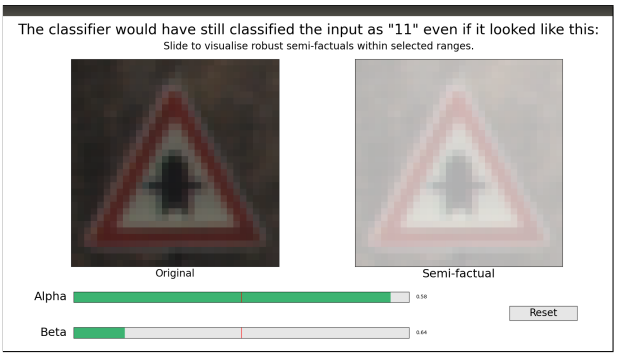

Figure 1: Our GUI to generate robust SF explanations.

Traffic sign recognition. We also considered agents dealing with traffic sign recognition tasks based on the GTSRB dataset [26], which contains images of traffic signs collected under strong variations in visual appearance due to, e.g., illumination and weather conditions. Given such variability, a neural agent may fail to provide robust decisions: a correctly classified image may cease to be so if small photometric changes were applied to it. We then show how SF explanations, augmented with robustness guarantees, can provide formal assurances to demonstrate that images will be classified correctly even in presence of photometric changes.

Experiment 4. We trained a convolutional neural network to solve the GTSRB classification task. We use a verification-based procedure to check whether classifications are robust across a set of photometric changes suggested by the user. When this is the case, the user is given the possibility to generate several SFEs by using GUI shown in Fig. 1, where sliders can be used to navigate the space of parameters controlling photometric changes (α for contrast and β for brightness). Otherwise, an explanation is returned to exemplify the circumstances under which the neural classifier fails.

4 FUTURE WORK

Our preliminary results motivate several further research directions, including: (i) extending our approach to other learned models (ii) investigating further synergies between XAI and VNN, aiming to improving the user-friendliness of our explanations (iii) conducting user studies to evaluate the implications that robustness (or a lack thereof) may have on human trust within HNMAS.

ACKNOWLEDGMENTS

Work partially supported by the DARPA Assured Autonomy programme (FA8750-18-C-0095), the UK Royal Academy of Engineering (CiET17/18-26) and the Imperial College Research Fellowship.

REFERENCES

[1] M. Akintunde, E. Botoeva, P. Kouvaros, and A. Lomuscio. 2022. Formal Verification of Neural Agents in Non-deterministic Environments. *Journal of Autonomous Agents and Multi-Agent Systems* 36, 1 (2022).

[2] S. Barrett, A. Rosenfeld, S. Kraus, and P. Stone. 2017. Making friends on the fly: Cooperating with new teammates. *Artif. Intell.* 242 (2017), 132–171.

[3] D. Björkegren, J. Blumenstock, and S. Knight. 2020. Manipulation-Proof Machine Learning. *arXiv preprint 2004.03865* (2020).

[4] E. Black, Z. Wang, and M. Fredrikson. 2022. Consistent Counterfactuals for Deep Models. In *Proceedings of the International Conference on Learning Representations, ICLR22*. OpenReview.net.

[5] E. Botoeva, P. Kouvaros, J. Kronqvist, A. Lomuscio, and R. Misener. 2020. Efficient Verification of Neural Networks via Dependency Analysis. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI20)*. AAAI Press, 3291–3299.

[6] A. Dhurandhar, P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam, and P. Das. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS18)*.

[7] R. Dominguez-Olmedo, A. Karimi, and B. Schölkopf. 2022. On the Adversarial Robustness of Causal Algorithmic Recourse. In *Proceedings of the 39th International Conference on Machine Learning (ICML22)*, Vol. 162. PMLR, 5324–5342.

[8] S. Dutta, J. Long, S. Mishra, C. Tilli, and D. Magazzeni. 2022. Robust Counterfactual Explanations for Tree-Based Ensembles. In *Proceedings of the International Conference on Machine Learning (ICML22)*, Vol. 162. PMLR, 5742–5756.

[9] FICO Community. 2019. Explainable Machine Learning Challenge. https://community.fico.com/s/explainable-machine-learning-challenge.

[10] D. Guidotti, L. Pulina, and A. Tacchella. 2021. pyNeVer: A Framework for Learning and Verification of Neural Networks. In *Prooceedings of the 19th International Symposium on Automated Technology for Verification and Analysis (ATVA21) (Lecture Notes in Computer Science, Vol. 12971)*. Springer, 357–363.

[11] L. Hancox-Li. 2020. Robustness in machine learning explanations: does it matter?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*20)*. ACM, 640–647.

[12] P. Henriksen and A. Lomuscio. 2020. Efficient Neural Network Verification via Adaptive Refinement and Adversarial Search. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI20)*. IOS Press, 2513–2520.

[13] P. Henriksen and A. Lomuscio. 2021. DEEPSPLIT: an Efficient Splitting Method for Neural Network Verification via Indirect Effect Analysis. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI21)*. ijcai.org, 2549–2555.

[14] N. Jennings, L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, and A. Rogers. 2014. Human-agent collectives. *Commun. ACM* 57, 12 (2014), 80–88.

[15] J. Jiang, F. Leofante, A. Rago, and F. Toni. 2023. Formalising the Robustness of Counterfactual Explanations for Neural Networks. In *Proocedings of the 37th AAAI Conference on Artificial Intelligence (AAAI23)*. AAAI Press.

[16] G. Katz, C. Barrett, D. Dill, K. Julian, and M. Kochenderfer. 2017. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In *Proceedings of the 29th International Conference on Computer Aided Verification (CAV17) (Lecture Notes in Computer Science, Vol. 10426)*. Springer, 97–117.

[17] E. Kenny and M. Keane. 2021. On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI21)*. AAAI Press, 11575–11585.

[18] P. Kouvaros and A. Lomuscio. 2021. Towards Scalable Complete Verification of ReLU Neural Networks via Dependency-based Branching. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI21)*. ijcai.org, 2643–2650.

[19] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[20] J. Marques-Silva and A. Ignatiev. 2022. Delivering Trustworthy AI through Formal XAI. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI22)*. AAAI Press, 12342–12350.

[21] T. Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.

[22] M. Pawelczyk, T. Datta, J. van-den Heuvel, G. Kasneci, and H. Lakkaraju. 2022. Let Users Decide: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. *CoRR* abs/2203.06768 (2022).

[23] L. Pulina and A. Tacchella. 2010. An Abstraction-Refinement Approach to Verification of Artificial Neural Networks. In *Proceedings of the 22nd International Conference on Computer Aided Verification (CAV10) (Lecture Notes in Computer Science, Vol. 6184)*. Springer, 243–257.

[24] A. Rosenfeld and A. Richardson. 2019. Explainability in human-agent systems. *Auton. Agents Multi Agent Syst.* 33, 6 (2019), 673–705.

[25] S. Sharma, J. Henderson, and J. Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES20)*. ACM, 166–172.

[26] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. 2011. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN11)*. IEEE, 1453–1460.

[27] S. Upadhyay, S. Joshi, and H. Lakkaraju. 2021. Towards Robust and Reliable Algorithmic Recourse. In *Advances in Neural Information Processing Systems 34 (NeurIPS21)*. 16926–16937.

[28] B. Ustun, A. Spangher, and Y. Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*19)*. ACM, 10–19.

[29] S. Wachter, B. Mittelstadt, and C. Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* abs/1711.00399 (2017).

[30] S. Wang, H. Zhang, K. Xu, X. Lin, S. Jana, C. Hsieh, and J. Kolter. 2021. Beta-crown: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification. *arXiv preprint arXiv:2103.06624* (2021).