Trust topologies in verification of social explainable AI

Wojciech Jamroga[1], Damian Kurpiewski[2], Łukasz Mikulski[3] and Teofil Sidoruk[4]

[1,2,4]Polish Academy of Sciences, Poland
[2,3]Nicolaus Copernicus University, Poland
[1]University of Luxembourg, Luxembourg
[4]Warsaw University of Technology, Poland

[1]w.jamroga@ipipan.waw.pl
[2]d.kurpiewski@ipipan.waw.pl
[3]l.mikulski@mat.umk.pl
[4]t.sidoruk@ipipan.waw.pl

1. Introduction

In the era of pervasive artificial intelligence, elements of AI are deeply entrenched in facets of our daily lives, ranging from social media interactions to car navigation, and even in curating our music and movie preferences. Beyond personal use, AI technologies drive the backbone of many business operations, ushering in profound societal and economic shifts. Recently, the trajectory of AI research has shifted towards a paradigm termed as *Social Explainable AI (SAI)*, a movement that prioritizes decentralization, human-centric design, and explicability [14, 2]. This trend is in response to the broader pushback against traditional, centralized machine learning systems — a pushback grounded not only in technical challenges like scalability but also in ethical imperatives surrounding data transparency and computational trustworthiness [3, 11].

SAI, while a budding concept, is a fertile ground for research exploration. As we embark on this journey, it's crucial to ascertain whether SAI delivers on its promises of efficacy, transparency, and user mindfulness. A holistic understanding of SAI requires studying its envisioned properties and potential unintended interactions, especially in a nuanced environment comprising both AI entities and human users [1, 15, 4]. One pressing concern is the susceptibility of SAI to adversarial threats, such as impersonation or more intricate man-in-the-middle attacks. In these cases, rogue entities could seize control of communication nodes to corrupt or manipulate data, or propagate misinformation. The vitality of Social AI hinges on its resilience to such threats; otherwise, its vulnerabilities will inevitably be targeted. While the discourse around adversarial onslaughts on traditional machine learning models is not novel, discussions surrounding SAI have primarily zeroed in on its anticipated functions. This might be attributed to the inherent complexities—conceptual, computational, and social—associated with SAI. Delving into the potential *unintended* behaviors of these systems undeniably poses significant challenges.

In our prior work [8], we advocated for the utilization of *formal methods for multi-agent systems* [16, 12] as a robust framework for an all-encompassing analysis of Social Explainable AI. Taking this exploration a step further, the current article shines a spotlight on introducing and integrating *trust topologies* into these models. Trust topologies offer a detailed map of trust relations between agents using directional graphs. In essence, an agent's trust in another is signified by its willingness to accept messages from the latter.

2. Models

The *Social Explainable AI* (SAI) framework [14, 2, 4] seeks to rectify the limitations of current centralized AI systems. Modern machine learning (ML)-driven AIs often operate as “black boxes”, challenging even for experts to interpret. This, combined with the challenges of managing vast datasets and concerns over privacy and centralized data storage, underscores the need for alternatives.

While many SAI initiatives employ *gossip learning* as their ML approach for PAIVs [13, 5, 6] and offer simulation tools for assessment [10], our approach carves out a unique direction. Our objective is to modify or entirely reshape our models by introducing trust topologies and novel network-like communication structures. We’re keen on enhancing multi-agent interactions in learning, leveraging these new topologies to foster more robust and transparent systems.

A cornerstone for optimizing interactions in Social Explainable AI (SAI) involves meticulously examining the foundational protocol. While our prior research provided a framework for this protocol [8], our current focus pivots to refining the AI model’s sharing phase, particularly by embedding trust topologies.

2.1. Agents and their phases

The spotlight of this exploration is the learning facet of the SAI protocol. Each device fortified with an AI module is portrayed as an individual agent. This agent functions in two primary cycles: the learning phase and the sharing phase.

Learning Phase In this phase, the agent meticulously trains its native AI model. This training’s proficiency hinges upon the agent’s inherent specifications and prowess, which mirrors the capabilities of the embedded components, such as the CPU. Repeated iterations of training refine the model’s quality, leading to potential outcomes where the model can be overtrained, undertrained, or aptly trained. Upon culmination of this phase, an agent is mandated to disseminate its model to its peers.

Sharing Phase The crux here is the model-sharing process among agents. This dissemination operates based on a simple protocol, designed around a ring topology. Each agent receives a model from its predecessor (in terms of ID) and bequeaths its model to its successor. This cycle culminates when the final agent circles back, bestowing its model upon the initial agent. Trust topologies come into play during model reception. An agent, based on its established trust parameters, elects to either embrace or dismiss an incoming model. Once accepted, this external model is amalgamated with the agent’s original model. Post this phase, agents can revert to learning, enhancing their models further.

For a granular depiction of this procedure, we intend to harness the capabilities of the open-source experimental model checker STV [9], which has been efficaciously employed in modeling real-world protocols such as Selene [7].

2.2. Potential threats in the ecosystem

Ideally, every agent in this ecosystem acts with integrity and stringently adheres to the protocol. However, real-world scenarios aren’t always utopian. Machines can falter, or agents might get tainted by malicious entities, prompting two pivotal threats: the “man-in-the-middle” and “impersonator” attacks.

Man-in-the-Middle This threat scenario introduces a rogue agent, termed the 'intruder'. While the intruder remains dormant during data assimilation and learning, its menace is palpable during sharing. It can stealthily intercept any model in transit and divert it to any agent of its choosing.

Impersonator Here, an AI agent, compromised by malignant code, manifests erratic behavior. While it's incapacitated from partaking in data collection or learning, it's fully equipped to share its model, adhering to the established protocol. The treachery lies in its ability to misrepresent its model's quality, beguiling the subsequent agent into erroneously endorsing it.

Trust topologies are central to mitigating these threats. By defining and adhering to a network of trust relationships, agents can make informed decisions on accepting models, bolstering the system's resilience against adversarial attacks.

3. Conclusions

In our preliminary exploration of Social Explainable AI (SAI), we emphasize the paramountcy of decentralization, transparency, and a human-centric approach. Recognizing the nascent stage of SAI research, we identify potential vulnerabilities that may be overlooked. Our tentative approach, integrating strategic ability models from multi-agent systems, offers a novel lens to address these gaps. The introduction of trust topologies, though still a work in progress, aims to discern its impact on verification results, laying groundwork for future rigorous examinations in SAI's formal modeling.

Acknowledgement. The work has been supported by NCBR Poland and FNR Luxembourg under the PolLux/FNR-CORE project SpaceVote (POLLUX-XI/14/SpaceVote/2023), by NCN Poland under the CHIST-ERA grant CHIST-ERA-19-XAI-010 (2020/02/Y/ST6/00064), and by the CNRS IEA project MoSART.

References

[1] M. Conti and A. Passarella, The internet of people: A human and data-centric paradigm for the next generation internet, *Comput. Commun.*, 131:51–65, 2018.

[2] P. Contucci, J. Kertesz and G. Osabutey, Human-AI ecosystem with abrupt changes as a function of the composition, *PLOS ONE*, 17(5):1–12, 2022.

[3] G. Drainakis, K. V. Katsaros, P. Pantazopoulos, V. Sourlas and A. Amditis, Federated vs. centralized machine learning under privacy-elastic users: A comparative analysis. In *Proceedings of NCA*, 1–8, IEEE, 2020.

[4] A. Fuchs, A. Passarella and M. Conti, Modeling human behavior — Part I: Learning and belief approaches, *CoRR*, abs/2205.06485, 2022.

[5] I. Hegedüs, G. Danner and M. Jelasity, Gossip learning as a decentralized alternative to federated learning. In *Proceedings of IFIP DAIS*, volume 11534 of *Lecture Notes in Computer Science*, 74–90, Springer, 2019.

[6] I. Hegedüs, G. Danner, and M. Jelasity, Decentralized learning works: An empirical comparison of gossip learning and federated learning, *J. Parallel Distributed Comput.*, 148:109–124, 2021.

[7] D. Kurpiewski, W. Jamroga, L. Masko, L. Mikulski, W. Pazderski, W. Penczek and T. Sidoruk, Verification of multi-agent properties in electronic voting: A case study, In *Proceedings of AiML 2022*, 2022.

[8] D. Kurpiewski, W. Jamroga and T. Sidoruk, Towards modelling and verification of social explainable AI. In A. P. Rocha, L. Steels and H. Jaap van den Herik (eds.), *Proceedings of the 15th International Conference on Agents and Artificial Intelligence, ICAART 2023, Volume 1, Lisbon, Portugal, February 22-24, 2023*, 396–403, SCITEPRESS, 2023.

[9] D. Kurpiewski, W. Pazderski, W. Jamroga and Y. Kim, STV+Reductions: Towards practical verification of strategic ability using model reductions, In *Proceedings of AAMAS*, 1770–1772. ACM, 2021.

[10] V. Lorenzo, C. Boldrini and A. Passarella, SAI simulator for social AI gossiping, 2022. `https://zenodo.org/record/5780042`

[11] A.-R. Ottun, P. C. Mane, Z. Yin, S. Paul, M. Liyanage, J. Pridmore, A. Yi Ding, R. Sharma, P. Nurmi and H. Flores, Social-aware federated learning: Challenges and opportunities in collaborative data training, *IEEE Internet Computing*, 1–7, 2022.

[12] Y. Shoham and K. Leyton-Brown, *Multiagent Systems – Algorithmic, Game-Theoretic, and Logical Foundations*, Cambridge University Press, 2009.

[13] Social AI gossiping. Micro-project in Humane-AI-Net, Project website, 2022. `https://www.ai4europe.eu/research/research-bundles/social-ai-gossiping`

[14] Social Explainable AI, CHIST-ERA, Project website, 2021–24. `http://www.sai-project.eu/`

[15] M. Toprak, C. Boldrini, A. Passarella and M. Conti, Harnessing the power of ego network layers for link prediction in online social networks, *CoRR*, abs/2109.09190, 2021.

[16] G. Weiss (ed.) *Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence.* Cambridge, Mass: MIT Press, 1999.