Towards an XAI World: Problems and Solutions

Penglin Cai
Yuanpei College
Peking University
cpl@stu.pku.edu.cn

Abstract

The explainability of AI has aroused much interest and attention for a long time. With the boosting of model size and parameters, it seems to be increasingly difficult to give a concrete explanation of how models work, and AI models are acting as if "black-boxes". However, unexplainable AI techniques may bring potential risks, including the misalignment between AI and humans, the weird phenomena such as hallucination, and the uncontrollable value systems of AIs along with the resulting behaviors. Therefore, how to design explainable AI (XAI) systems has become a serious problem to be solved. In this essay, we introduce some major problems in designing XAI systems and propose possible solutions accordingly.

1 Introduction

With the increasing model size and the deepening network structure, the explainable AI (XAI) technique has become a serious problem to be resolved. Although existing deep neural networks and huge models have achieved great performance in solving tasks and making decisions, these models are usually lacking in explainability [11]. Most of the times we do not know what is happening in the "black-boxes", nor do we understand how they make decisions from those latent variables.

The existing unexplainability has brought many concerns and problems. One early example is that vanilla deep neural networks are lacking in adversarial robustness [27]. Another is that the misalignment between AI and humans makes it difficult for those embodied agents to think and act like human, and to make decisions in humans' situations. Moreover, trustworthy large models are far from being realized due to some drawbacks and shortcomings of large models. For instance, the phenomenon of hallucination of large language models (LLMs) make them not stable and reliable sometimes, thus preventing humans from trusting them.

In this essay, we discuss the major problems and challenges remaining to be resolved in current research. We also try to propose possible solutions to tackle these challenges, with both tentative approaches and potential strengths or limitations. Our discussion will revolve around the aim towards an XAI world.

2 Problems, Challenges, Solutions, and Opportunities

In this section, we list several problems and challenges relevant to XAI in the existing studies, and propose some potential solutions.

2.1 The Adversarial Robustness in Deep Learning

A typical example of adversarial attacking happened in the setting of classification of visual images. Ian *et al*. [10] tried to attack a trained classifier using a specially designed noise. The original image added with the noise showed no obvious abnormalities from the perspective of human observations, but would lead to the classifier's wrong classifications (Figure 1).

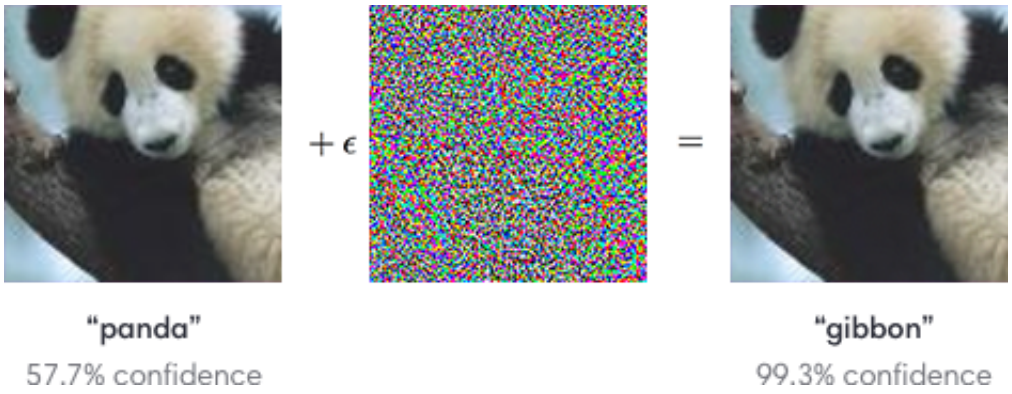

Figure 1: An attacking example from Ian *et al*. [10]. The trained model would classify a normal image into a correct category, but would classify the image added with aggressive noises into a totally wrong category, with a significant confidence.

In the aspect of optimization, there have been much previous work aimed at designing defensive algorithms [21, 31, 33], most of which focused on mathematical derivation. Another novel method, adversarial training [4], was proposed to intrinsically enhance the robustness. However, such methods can be weak in generalizability, which is an unresolved issue.

From our perspective, the instability when facing an adversarial attack can actually reflect the model's sensitivity to small disturbances. The reason why these models cannot distinguish these images after small perturbations is that they do not truly understand and master the abstract concepts of these objects. Hence, there should be an alignment between these concepts and these images (or other carriers), but not the implicit variables (such as the tensor used to store the image).

In terms of such alignment, one idea is to leverage the emergence capability of large models with massive amount of data. For instance, large language models and large vision generative models have demonstrated marvelous ability of emergence [1, 29]. Another idea is to design a proper computational model to learn the representations of the concepts in daily lives. Although there has been much previous work in the human-level learning [15, 35], few of the studies focus on computational frameworks, remaining an unresolved challenge. Following the principle in physiology, learning a new concept may be accompanied by the formation of new neurons and new connections [8]. Similarly, the connection and interruption of neurons in a neural network can be reflected in the weight of the parameters. Through learning and training, the magnitude of these parameters can demonstrate the tightness of each connection. This may provide a possible way of explaining the deep neural networks.

2.2 The Alignment between Humans and Embodied AI

In an embodied environment, intelligent agents are asked to act as humans in the real world, including navigation, exploration, interaction with objects, and making decisions. However, why do agents engage in such behaviors still needs explaining. Existing embodied environments [24, 26] can only achieve alignment between vision, language, and physical simulation, which is a low-level alignment. By contrast, a higher-level alignment asks the alignment between agents and humans, *i.e*., the agents should determine all the behaviors standing at the same situation as humans do.

There has been a few researches on human-robot alignment. Li *et al*. [18] modeled human-AI alignment in teaming using a dynamic game theory approach. Harland *et al*. [12] studied the alignment in goal-conditioned reinforcement learning when there were multiple objects as goals. Gabriel *et al*. [9] focused on the cases where agents should be in line with humans in the value system.

In our proposal, the alignment in embodied environments mainly consists of two parts - the alignment in the level of cognition (physically) and the alignment in the value system (mentally). The cognitive alignment can be considered as inferring the intention of humans, which requires strong insight and the ability of reasoning. On the other hand, the alignment in the value system may require learning from massive amount of common sense, obtaining a value system similar to humans.

2.3 Towards a Unified Human-AI Collaboration

In robotics and the similar areas, one of the abilities that intelligent agents must be equipped with is to collaborate with humans [28]. However, how agents can cooperate with humans in a teamwork remains unexplained.

In this area, previous typical attempts [7, 23] were to design intelligent agents to assist humans playing in the Overcooked environment [6] (Figure 2). However, in these attempts, agents usually served as assistants to aid humans, without their subject status. In another example, RoCo [20] was a proposed benchmark constructing a setting of multi-robot collaboration, which introduced robotics into multi-agent systems. However, RoCo utilized LLMs as upstream decision-makers, while the downstream used a module of motion planning [16]. This further reduces explainability.

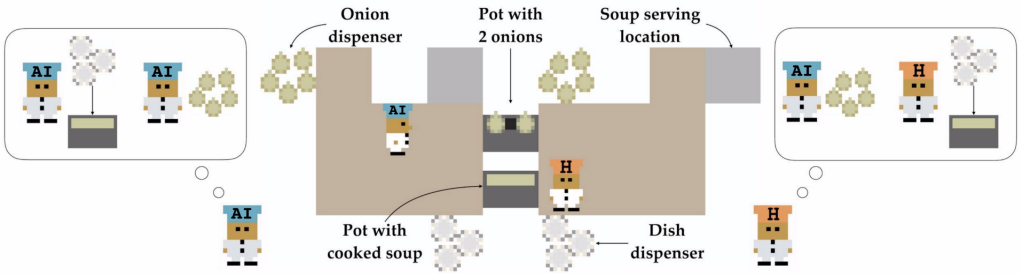

Figure 2: The Overcooked environment used in Carroll *et al*. [7]. "H" denotes human and "AI" denotes agents, leveraging human-AI collaboration to solve multiple tasks in the environment.

Accordingly, we propose that the subject status of intelligent agents should be emphasized in a unified human-AI collaboration scenario. In other words, we should let the agents take the initiative, but not serving as assistants following instructions. What we should do is to enhance the emotional quotient of agents, making them to find things to be done themselves.

Another track lies in the methods used in training these agents. Existing methods focused on optimizing a specific and even pre-defined function, thus lacking explainability. Although the designed optimization objectives can be explained by human prior knowledge, the results of the optimization can be far from our imagination. One idea is to make each agent explore for their own implicit potential function, which is used to represent their certain features, such as personality. With these explored representations, agents can develop their own personalities, explain their own decisions, and further learn how to collaborate with humans.

2.4 Satisfactory and Trustworthy LLMs

In the era of AIGC, more and more people begin to seek help from large generative models. However, sometimes the performances of these generative models are not satisfactory, making them not reliable enough. In this section, we mainly discuss the unsatisfactory of large language models (LLMs) as an example, with the already-taken and future-possible solutions.

Large language models have been found to have many capability boundaries in various domains. For instance, Arkoudas *et al*. [2] claimed that GPT-4 [22], which stands for the most capable model among existing LLMs, cannot truly reason. By contrast, GPT-4 was found to solve many other tasks pretty well, such as decision making, information searching, and deliberation [5]. Another common but serious phenomena is that LLLMs have the tendency of hallucination [3]. A recent work aimed to benchmark Large Language Models as AI research agents, but encountered the cases where LLMs were hallucinating [13]. Sometimes Large Language Models would make bad plans during some certain steps in the research process. After being pointed out where they did it poorly, their next decision was claimed to have been improved, but with no modification at all.

To tackle such problems, a natural idea is to make LLMs report their process of critical thinking. Therefore, Chain-of-Thought [30] was proposed to add an intermediate thought as prompting, aimed to enhance the emergence capacity of LLMs. In terms of the scenarios in which LLM-based agents interact with the environment, ReAct [34] and Reflection [25] could help LLMs reflect during the interaction.

Besides, limited by some closed-source black-box models, researchers have been exploring new methods towards prompt-tuning [17], which intends to optimize the input prompt to achieve better results. For instance, P-tuning [19] was universally effective across a wide range of model scales on natural language understanding tasks. Later, such methods were generalized to other relevant domains, such as prompt tuning in decision-transformers [14, 32]. One of the great advantages is that these methods are parameter-efficient, thus saving resources for training.

In our perspective, working on the input prompts can indeed be the most effective way to tackle such challenges, especially for those closed-source black-box models. Therefore, how to design plausible and effective prompts to make LLMs explain their behaviors and enhance the performances, can be a vital research area in the foreseeable future.

3 Conclusion

In this essay, we reviewed the development of explainable AI (XAI) techniques. With the discussion of the existing problems and major challenges, we propose several possible solutions based on marvelous previous work. Through these discussions, we aim to tackle the problems and challenges via such solutions, stepping towards an XAI world with great explainability of AI techniques.

References

[1] Alexander Andonian. *Emergent Capabilities of Generative Models:"Software 3.0" and Beyond*. PhD thesis, Massachusetts Institute of Technology, 2021. 2

[2] Konstantine Arkoudas. Gpt-4 can't reason. *arXiv preprint arXiv:2308.03762*, 2023. 3

[3] Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):1–2, 2023. 3

[4] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 2

[5] Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. 3

[6] Justin Bishop, Jaylen Burgess, Cooper Ramos, Jade B Driggs, Tom Williams, Chad C Tossell, Elizabeth Phillips, Tyler H Shaw, and Ewart J de Visser. Chaopt: a testbed for evaluating human-autonomy team collaboration using the video game overcooked! 2. In *2020 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6. IEEE, 2020. 3

[7] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019. 3

[8] Wei Deng, James B Aimone, and Fred H Gage. New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory? *Nature reviews neuroscience*, 11 (5):339–350, 2010. 2

[9] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437, 2020. 2

[10] Ian Goodfellow, Nicolas Papernot, Sandy Huang, Yan Duan, Pieter Abbeel, and Jack Clark. Attacking machine learning with adversarial examples. *OpenAI Blog*, 24, 2017. 1, 2

[11] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120, 2019. 1

[12] Hadassah Harland, Richard Dazeley, Bahareh Nakisa, Francisco Cruz, and Peter Vamplew. Ai apology: interactive multi-objective reinforcement learning for human-aligned ai. *Neural Computing and Applications*, pages 1–14, 2023. 2

[13] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Benchmarking large language models as ai research agents. *arXiv preprint arXiv:2310.03302*, 2023. 3

[14] Jikun Kang, Romain Laroche, Xindi Yuan, Adam Trischler, Xue Liu, and Jie Fu. Think before you act: Decision transformers with internal working memory. *arXiv preprint arXiv:2305.16338*, 2023. 4

[15] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 2

[16] Jean-Claude Latombe. *Robot motion planning*, volume 124. Springer Science & Business Media, 2012. 3

[17] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 4

[18] Mengyao Li and John D Lee. Modeling goal alignment in human-ai teaming: a dynamic game theory approach. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 66, pages 1538–1542. SAGE Publications Sage CA: Los Angeles, CA, 2022. 2

[19] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022. 4

[20] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. *arXiv preprint arXiv:2307.04738*, 2023. 3

[21] Alexander Matyasko and Lap-Pui Chau. Improved network robustness with adversary critic. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[22] OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2023. 3

[23] Andres Rosero, Faustina Dinh, Ewart J de Visser, Tyler Shaw, and Elizabeth Phillips. Two many cooks: Understanding dynamic human-agent team communication and perception using overcooked 2. *arXiv preprint arXiv:2110.03071*, 2021. 3

[24] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 2

[25] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[26] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020. 2

[27] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020. 1

[28] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–6, 2020. 3

[29] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 2

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 3

[31] Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018. 2

[32] Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, pages 24631–24645. PMLR, 2022. 4

[33] Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. In *International Conference on Machine Learning*, pages 7025–7034. PMLR, 2019. 2

[34] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[35] Dagmar Zeithamova, Michael L Mack, Kurt Braunlich, Tyler Davis, Carol A Seger, Marlieke TR Van Kesteren, and Andreas Wutz. Brain mechanisms of concept learning. *Journal of Neuroscience*, 39(42):8259–8266, 2019. 2