Open the Black Box:
the Journey Toward Responsible Artificial Intelligence

Jingze Zhang
Department of Automation
Tsinghua University
jz-zhang21@mails.tsinghua.edu.cn

Abstract

Enabling AI systems to collaborate safely and efficiently with humans is one of the ultimate goals in AI research. However, the majority of state-of-the-art intelligent systems today operate as black-box models. This means that while we are aware of their impressive capabilities, we remain unaware of the causal chains and underlying logic driving their reasoning. This limitation hinders the advancement of artificial intelligence and makes it challenging for humans to fully trust these autonomous decision-making intelligent agents. Given these considerations, in this essay, we offer a profound exploration of explainable artificial intelligence, reviewing its fundamental paradigms, and contemplating potential pathways toward responsible artificial intelligence in the future.

Figure 1: **Our ultimate goal of XAI is responsible AI.** Nowadays, when we talk about human robot value alignment, or when we reflect on the U-V theory, one significant thing that we cannot forget is that we still on our journey to open the black-box of AI systems. The foundation of responsible AI systems such as Baymax in *Big Hero 6* is explainable AI system.

1 Introduction

> *"Machine learning is becoming ubiquitous in basic research as well as in industry. But for humans to trust it, they first need to understand what the machines are doing."*[2]
>
> ———By David Castelvecchi

Artificial intelligence technology has been one of the most rapidly developing fields in recent years. A decade ago, intelligent systems were only capable of handling basic AI tasks such as image classification[7, 5], object detection[10] or other simple tasks far from human-level creativity. However, today's intelligent systems can manage complex multi-modal tasks and have achieved unimaginable successes in areas like generating models[11, 6, 13]. Following this tendency, it's hard to imagine what the next decades will witness in the general artificial intelligence industry, especially in cognitive reasoning and robotics whose ability is still far from human intelligence currently.

With the swift evolution of AI technology, more and more judgments and decisions are closely dependent on intelligent systems. In this context, a completely inexplicable system is difficult for us to accept, which highlights the urgency of addressing the safety and interpretability issues in artificial intelligence.

However, the domain of *Explainable Artificial Intelligence* or XAI, is a widely discussed but still controversial topic. It encompasses a wide range of topics for scholarly inquiry and reflection. The diversity of these topics often results in a lack of clarity regarding the precise definitions, testing methodologies, and potential future technological directions within the domain of XAI[2].

Based on the aforementioned motivations and my limited literature review about XAI, we seperate our essay into the following 4 parts. In section 2, we will give an explanation about the philosophy foundation of explanation. In section 3, we will divide current XAI into two paradigms: transparent models and post-hoc models, and we will also discuss the virtue, drawback and relationship of these 2 paradigms. Furthermore, we will discuss the journey towards responsible artificial intelligence, which serves as our utimate goal of explainable artificial intelligence(XAI).

2 What is explanation: a reflection from philosophy perspective

In Malle [8]'s book, the authors provided a general reason about the philosophy and psychology view of explanation. From their aspect of view, the starting point of explanation is ***share meaning***, which is a significant driven-power of meaning finding and social interaction management. The ability to share meaning unambiguously is a crucial foundation for successful cooperation between intelligent agents[9]. Therefore, explainable artificial intelligence is an essential condition for the integration of intelligent agents into human society and for their safe and efficient collaboration with humans.

From a philosophy aspect, the explanation process can also be recognised as a series Q&A process. For instance, when others want to challenge the robustness of our models, we may be challenged with the following questions: Why does object a have property P? Why do you make the decision a, rather to obtain the action b? The answer process can be comprehended as a kind explanation process and can viewed as a test standard of XAI, where our ultimate goal is to dig into the artificial intelligent system decision making process and find out the casual chain during the inference process of the model.

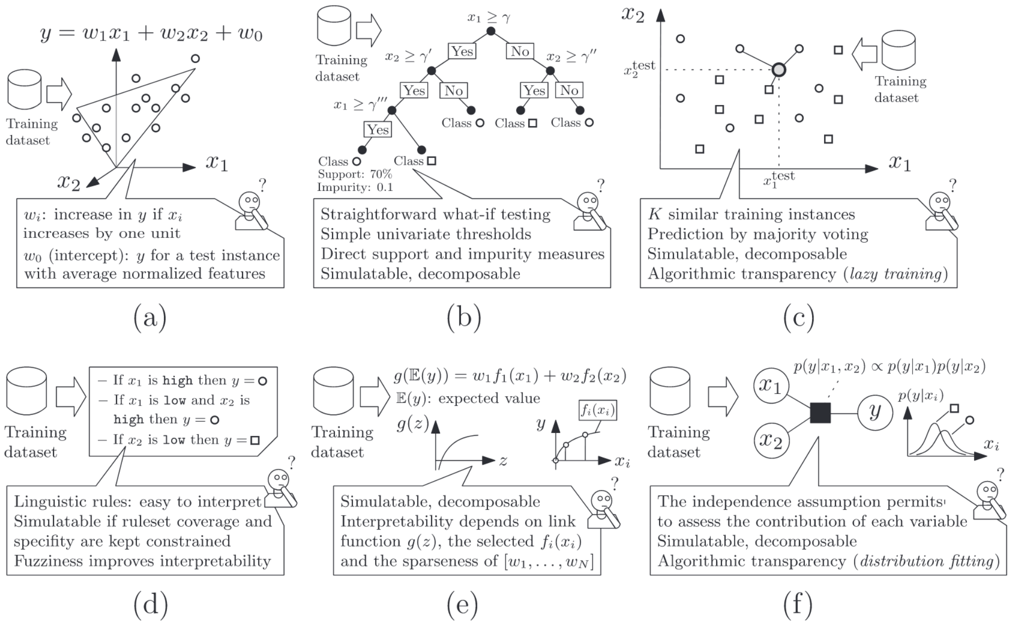
From the above analysis, we can infer that an interpretable model does not necessarily have to be a white-box model. For instance, human intelligence, to this day, remains an unopened black box, yet this does not hinder the interpretability of the brain as an intelligent system. Specifically, the focus of our subsequent discussion will be on the interpretable aspects of both white-box and black-box models.

3 Transparent models or Post-hoc explainabilty model?

Once we have defined explainable artificial intelligence (XAI) and established our ultimate goal, we can reconsider the paradigms through which we aim to achieve explainable AI. Presently, research primarily falls into two paradigms: one involves designing models with interpretability in mind from the outset, while the other relies on explaining why an effective model works well after validation. After briefly reviewing these two paradigms, we will delve into the connections between them and explore the possibility of integrating these two approaches to build a more general model.

3.1 Transparent models

A viable approach to achieve intelligence models with interpretability is to design them as transparent models from the outset. In such models, the behavior of each layer is readily comprehensible to humans. Prior to the dominance of *deep neural networks* (DNN) in artificial intelligence research, these white-box models held sway in the realm of model exploration. In the following figure 2, we will find that these models are primarily composed of statistical components and owe their high interpretability due to the fact that their structures are directly designed based on statistical principles and mathematical intuition.


$y = w_1x_1 + w_2x_2 + w_0$
Training dataset
x_1
x_2
?
$x_1 \geq \gamma$
Yes
No
$x_2 \geq \gamma'$
$x_2 \geq \gamma''$
$x_1 \geq \gamma'''$
Class ○
Class □
Support: 70%
Impurity: 0.1
x_1^{test}
x_2^{test}
– If x_1 is high then $y =$ ○
– If x_1 is low and x_2 is high then $y =$ ○
– If x_2 is low then $y =$ □
$g(\mathbb{E}(y)) = w_1f_1(x_1) + w_2f_2(x_2)$
$\mathbb{E}(y)$: expected value
$g(z)$
z
y
$f_i(x_i)$
x_i
$p(y|x_1, x_2) \propto p(y|x_1)p(y|x_2)$
$p(y|x_i)$

w_i: increase in y if x_i increases by one unit
w_0 (intercept): y for a test instance with average normalized features

(a)

Straightforward what-if testing
Simple univariate thresholds
Direct support and impurity measures
Simulatable, decomposable

(b)

K similar training instances
Prediction by majority voting
Simulatable, decomposable
Algorithmic transparency (*lazy training*)

(c)

Linguistic rules: easy to interpret
Simulatable if ruleset coverage and specifity are kept constrained
Fuzziness improves interpretability

(d)

Simulatable, decomposable
Interpretability depends on link function $g(z)$, the selected $f_i(x_i)$ and the sparseness of $[w_1, \ldots, w_N]$

(e)

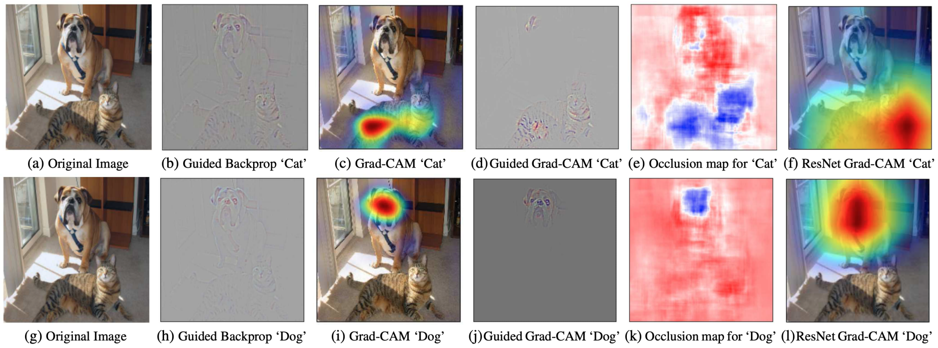
The independence assumption permits to assess the contribution of each variable
Simulatable, decomposable
Algorithmic transparency (*distribution fitting*)

(f)

Figure 2: **A graphical illustration of dominant transparent models.**The resource is obtained from Arrieta et al. [1]. The figure showcases some visualization of white-box model. The models from (a) to (f) are linear model, decision tree, k-nearest neighbor, rule-based method, generailized additive models and bayesian models.

3.2 Post-hoc explainability models

Another widely adopted paradigm in explainable artificial intelligence(XAI) is known as *Post-hoc explainability models*. The advantage of this approach lies in not incorporating excessive prior knowledge into the model's design, thereby ensuring the model's capabilities[1]. Post-hoc explainability aims to interpret black-box models through methods like statistical analysis and alignment, performed after the fact. This approach elucidates the underlying mechanisms behind the model's impressive performance, thereby deepening our understanding and trust in the model. Additionally, it offers potential methods to enhance the model's effectiveness.


(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d)Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j)Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (l)ResNet Grad-CAM 'Dog'

Figure 3: **A visualization of post-hoc explainability models.** The figure is obtained from Selvaraju et al. [12]. In this work, the researchers find what is happening in the latent layers of deep neural network. Similar resuls can also be found in attention-based or transformer-structured models.

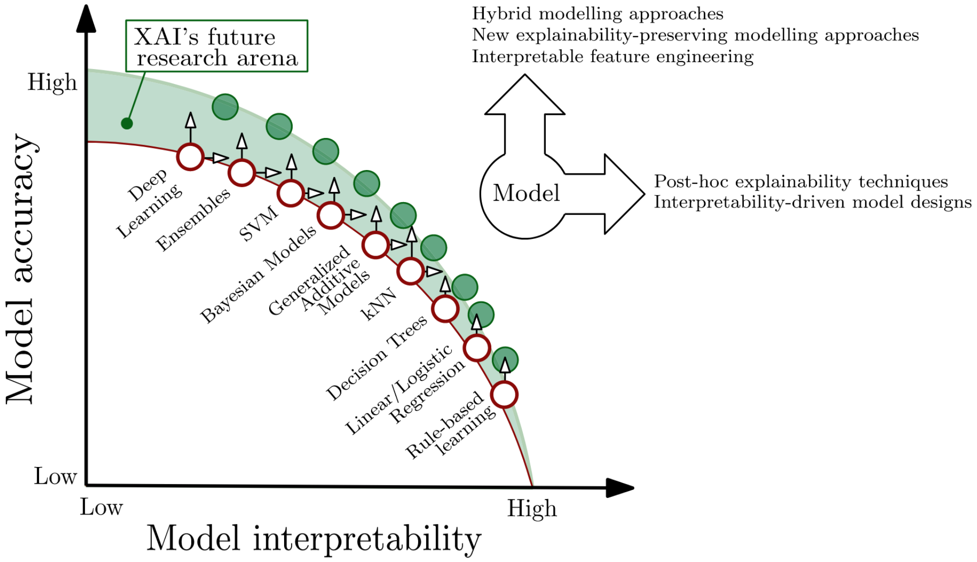


Figure 4: Trade-off of the model accuracy and model interpretability. XAI's future lies in the green area. The image is obtained from [1]

4 Towards responsible artificial intelligence

4.1 World model aligned with human values

Humans have a world model in their mind to comprehend and inference about the complex world. The world model can be described as a system that learns a compressed spatial and temporal representation of the world via an unsupervised manner[4]. The concept of the world model was first proposed to utilize as a compressed environment to train reinforcement learning agents[4]. And the concept of world model has been generalized in the *Large Modal Model* age, the concept of world model has been generalized.

Recently, a study conducted by MIT found large modal models learn a representation of the real world in the spatial and temporal dimension[3]. It shows that the large modal models draws the conclusion via have the model project geographic locations onto the world map. Results also showcase that similar to the information processing procedure in our brains, when processing with temporal information and spatial information, different parts of the neural network are activated.

Numerous experiments have indicated that current large language models have developed mechanisms akin to a world model. Our understanding of these internal world models within language models remains vague. For the advancement of explainable and responsible artificial intelligence, it is crucial to make these internal world models transparent and align them with human behaviors and value systems.

4.2 Machine ToE

In a previous essay, we delved deeply into the concept of the "Machine Theory of Mind" and its significance. The "Theory of Mind" refers to the human capacity to construct a model of oneself and others' minds, enabling the rational explanation and inference of others' behaviors. This idea offers a potential technological pathway for developing explainable and responsible artificial intelligence. Future responsible artificial intelligence will need to possess a "Theory of Mind" capable of understanding both the self and others, and this "Theory of Mind" should be a subset of the human Theory of Mind.

5 Summary

In this essay, we delve into the definition and philosophical interpretation of explainable artificial intelligence models. We present several prevalent methods currently used to implement explainable intelligence and explore potential approaches to explainability in the context of future general-purpose AI and responsible AI systems. Through literature review and analysis, we have come to understand that there is still a long way to go in achieving truly explainable intelligence. Our future aim is to design AI systems that are not only highly interpretable but also highly effective.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 3, 4

[2] Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016. 1, 2

[3] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023. 4

[4] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023. 1

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`. 1

[8] Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press, 2006. 2

[9] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Cornell University - arXiv,Cornell University - arXiv*, Jun 2017. 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[12] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[13] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1), 2023. 1