What's in an AI's Mind's Eye? We Must Know

Moshe Sipper[1*] and Raz Lapid[1,2]

[1]Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva, 8410501, Israel
[2]DeepKeep, Tel-Aviv, Israel

[*]To whom correspondence should be addressed; E-mail: sipper@bgu.ac.il.

Preprint, to appear in *IEEE Computer*, July 2024

Abstract

We discuss the crucial importance of explainability and understandability in artificial intelligence, in addition offering a small, insightful experiment, followed by a discussion of responses, challenges, and obstacles. We believe the pursuit of AI explainability and understandability is crucial, to be ignored at out peril.

AI Explainability is Hard—Yet Crucial

One of the major challenges of modern-day artificial intelligence (AI) is the striving to understand the "mind" of an AI—in particular deep networks—i.e., comprehend how such networks "think" (some would argue the quotation marks are superfluous).

We humans often ask "why"—we *need* explanations. This is not just some abstract desire. It is part of our makeup, part of our evolutionary ancestry whose survival depended on modeling and understanding the world. Thus, it is natural that we seek explanations from AIs. Why did the AI reject a college application? Why did the AI classify the image as an elephant? What are the reasons? Are they sound? Or, do they evidence bias or some other form of erroneous thinking? Why did the AI recommend a certain medical procedure? Indeed, in the medical domain explainability is legally required.

An entire subfield of *eXplainable Artificial Intelligence (XAI)* has arisen in recent years [1], focusing on finding explanations for the reasoning of deep networks. However, we find that usually such explanations are "shallow" in some sense, when compared with deeper explanations that humans offer upon being questioned about their thoughts and reasoning.

A typical example of XAI in images involves a so-called "explanation map", which shows the pixels most responsible for producing a network's output, for example, classification of a bird in a nature image (Figure 1a). Essentially, these are the pixels that had the most influence on the network's gradients [2]. While such commonly used explanation maps offer insight into a network's workings, we think the explanations are shallow and local, being, as they were, about low-level pixels and not about high-level concepts, such as: "This is a bird because of its beak and feathers". XAI is used not only to explain images, but also for other forms of data: tabular, time series, linguistic, and more. Yet from what we have seen there is always some sense of not being quite up to par with human explanations.

There have been interesting attempts to go beyond shallow explanations. For example, Feather et al. [3] focused on metamers—"stimuli that produce the same responses at some stage of a network's representation"—showing that metamers from early network layers were recognizable to human observers but those from deeper layers were not.

A recent work introduced the idea of a "probe"—a neural network that is simpler than the one under study, trained to decode the original network's internal activations [4]. While undoubtedly a step forward, this still does not quite provide a full-blown, humanlike, explanation. Very recently, thanks to advances in multimodal AI [5], new XAI approaches began delivering conceptual explanation capabilities. Another recent approach used a large language model (GPT-4) to explain neurons in another large language model (GPT-2XL), focusing on what they termed "explanation score": a measure of a language model's ability to compress and reconstruct neuron activations using natural language [6]. Schwettmann et al. [7] introduced FIND (Function INterpretation and Description), a benchmark suite for evaluating the building blocks of automated interpretability methods. These approaches are resource intensive since they rely on large, deep networks.

Explanation at Eye Level: An Applied Gedankenexperiment

We wish herein to advocate a higher level of explanation modeling and understanding. Towards this end we designed and performed a simple yet thought-provoking setup focusing on faces in the CelebA dataset. We deployed the Deepface software package, which includes both facial recognition and facial-attribute analysis [8]. The package offers several models, trained *independently* and on *different* datasets. For facial recognition we chose three of the available models: Google FaceNet, OpenFace, and ArcFace. For facial-attribute analysis Deepface offers four models: gender, age, facial expression, and ethnicity.

Crucially, we now have access to completely different models, trained on different datasets, performing different tasks. We then perform the following steps:

- Select a random image from CelebA—designate it the "original".
- Call `DeepFace.find` with the original image and the entire CelebA dataset. This function finds the most-similar images to the original by generating latent representations (embeddings) of all dataset images and then comparing those with the embedding of the original image through the cosine-similarity measure.
- The most-similar image should be the same as the original. For the next 5 most-similar images, call `DeepFace.analyze`, a function that deploys the four models that asses gender, age, expression, and ethnicity.

This simple procedure is repeated to produce multiple outputs. Note that `find` (facial recognition) and `analyze` (facial attributes) use different models, as explained above.[1]

Figure 1b shows a sample output panel. There are three rows, per three facial-recognition models. For each model we considered the 6 most-similar images. The most similar is the original, followed by five additional images to the right. *Independently* of face recognition, now come the attribute models and analyze the images. The analysis results are given below each image, showing the outputs produced by the gender model, the age model, the expression model, and the ethnicity model.

As we've emphasized above, there are several independent models at work here. Face recognition is done separately from face analysis, and within each of these two categories the models are different.

Observing the sample panel of Figure 1b we note that the analysis of similar images tend to agree to some extent or other. Indeed, to gather statistics we ran 1000 random images, which amounted to 15,000 images (1000 x 3 models x 5 images). For each image we then asked whether the analyzed attributes agreed with those of the original (for 3 attributes this is a simple true/false assessment; for age, we defined 'agreement' as being within 3 years either way). The results were: age: 59% agreement with original, emotion: 49%, gender: 88%, ethnicity: 68%.

We find it interesting that when one model outputs images it considers similar, a completely different model tends to view *high-level concepts* (and human at that)—gender, age, emotion, ethnicity—similarly.

Another intriguing phenomenon we observed is that now and again similar images found by the recognition models caused the analysis models—again, independently—to make similar mistakes. This is demonstrated in Figure 1c: The analysis models seems to have misjudged the original image with respect to gender and age. We then obtain similar images through the recognition model. They do not look quite similar—like humans, AI is not

[1] The code is available online at [9].

perfect—yet, curiously, when you hand them over to the analysis models, gender and age coincide with the original mistakes.

Responses, Challenges, Obstacles

In response to the fundemental challenge of insufficient to no explainability in most contemporary AI solutions, the literature presents several strategic avenues. Each approach comes with pros and cons.

Interpretable models, such as decision trees and linear models, inherently offer transparency in the decision-making process. This transparency, however, comes at the expense of a trade-off between interpretability and predictive accuracy: The more interpretable the model, the simpler it needs to be, and thus its predictive accuracy declines. That said, for some tasks, these oft-overlooked models are the perfect choice.

Rule-based aystems define decision rules explicitly, thus offering inherent transparency. However, manual crafting of rules may be impractical for complex tasks, and automated rule generation encounters challenges in capturing subtle decision boundaries.

Explainability techniques for black-box models, such as Local Interpretable Model-agnostic Explanations (LIME) provide locally good explanations for complex, black-box models. However, global interpretability is not guaranteed at all, and fidelity with respect to the overall model behavior may be compromised.

Visualizations, such as saliency maps, offer intuitive insights into model decisions. Challenges lie in designing effective visualizations, and interpretation by humans may greatly vary, potentially leading to misconceptions. Further, it has been shown that it is possible to manipulate these maps—so-called adversarial attacks [2].

The pursuit of explainable AI is often driven by the desire to enhance trust and understanding in AI systems. However, while XAI holds the potential to address these concerns, it is important to recognize possible unforeseen challenges and unintended consequences. We think there are (at least) four key obstacles that warrant careful consideration:

Trade-off between accuracy and interpretability is always an intricate balancing act. A more accurate model will usually tend to be less interpretable, and vice versa.

Security concerns. Explanations generated by XAI systems can be powerful tools for understanding and communicating AI decisions. However, they also carry the risk of being misused or misinterpreted. For example, explanations could be used to manipulate users by framing decisions in a biased or misleading way, or to justify biased decisions by providing a veneer of objectivity. Additionally, users may oversimplify or misinterpret explanations, leading to inaccurate or incomplete understanding of AI decisions.

Fairness and robustness. XAI explanations should not only provide insights into AI decisions but also be fair and robust to potential biases. This means that explanations should not perpetuate or reinforce existing biases in the data or the model itself. Moreover, explanations should be robust to adversarial attacks or attempts to manipulate them to achieve specific outcomes. Ensuring fairness and robustness in XAI is particularly crucial in sensitive applications where AI decisions have significant impacts on individuals or groups (e.g., the medical domain).

Illusion of understanding. XAI can provide valuable insights into the inner workings of AI models, but it is important to avoid creating an illusion of complete understanding. AI models, especially complex ones, often involve intricate relationships between features, non-linear dependencies, and stochastic processes. While XAI can help unravel some of these complexities, it is essential to recognize that explanations may not capture the full extent of the model's behavior. Overreliance on XAI explanations without critical evaluation could hinder a deeper understanding of AI systems and their limitations.

... And What is Explainability?

The point at which we shall be content with an explanation is unclear. Is "because it has two ears" enough? Why does the network output this explanation? Can it dig further to produce, e.g., "because most mammals have two ears"? Is *that* a sufficient explanation? Here we seem to be delving into the philosophical nature of explanations—but we may have to, given the rise of AI. As recently noted by Prince [10]: "There is also an ongoing debate about what it means for a system to be explainable, understandable, or interpretable... there is currently no concrete definition of these concepts."

We believe the pursuit of AI explainability and understandability is crucial, to be ignored at out peril. Perhaps task completion and its explanation should be fully integrated, as recently shown by Sipper [11]. Deep learning pioneer Geoffrey Hinton said in a recent interview (CBS News, 10/8/2023): "What we did was we designed the learning algorithm. That's a bit like designing the principle of evolution. But when this learning algorithm then interacts with data, it produces complicated neural networks that are good at doing things. But we don't really understand exactly how they do those things."

References

[1] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022.

[2] Snir Vitrack Tamam, Raz Lapid, and Moshe Sipper. Foiling explanations in deep neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=wvLQMHtyLk.

[3] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ac27b77292582bc293a51055bfc994ee-Paper.pdf.

[4] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

[5] Nikolaos Rodis, Christos Sardianos, Georgios Th Papadopoulos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Iraklis Varlamis. Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions. *arXiv preprint arXiv:2306.05731*, 2023.

[6] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models, 2023. URL https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

[7] Sarah Schwettmann, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. FIND: A function description benchmark for evaluating interpretability methods, 2023.

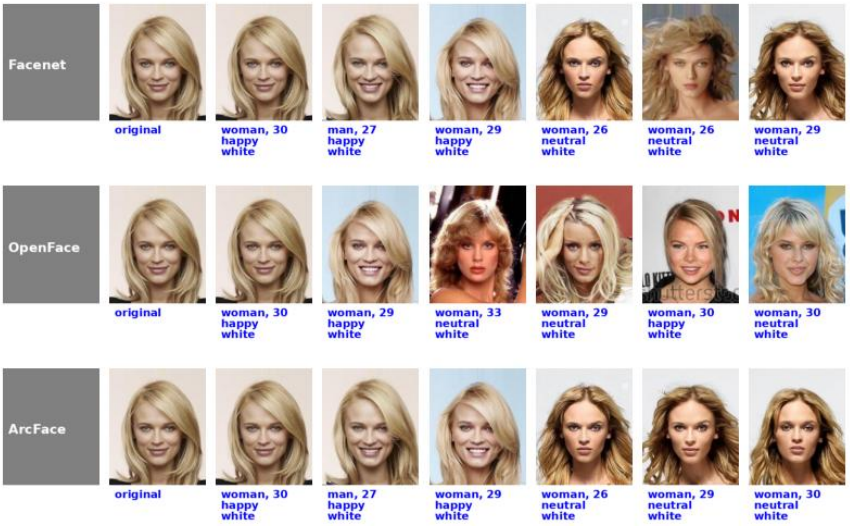
[8] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. doi: 10.1109/ICEET53442.2021.9659697. URL https://doi.org/10.1109/ICEET53442.2021.9659697.

[9] Moshe Sipper. A minimal example of face recognition and facial analysis, using Deepface, with an accompanying Colab notebook, 2023. URL https://medium.com/ai-mind-labs/a-minimal-example-of-face-recognition-and-facial-analysis-ce4024da30d8.

[10] Simon JD Prince. *Understanding Deep Learning*. MIT Press, 2023.

[11] Moshe Sipper. Task and explanation network. arXiv:2401.01732, 2024.

(a) Image (left) with sample explanation map (right).



(b) Sample output panel. Each row shows 1+1+5 images: original, original again, 5 most similar to original.

(c) Example of an interesting error.

Figure 1: Understanding AI thinking.