An Analysis on Explainability and Interpretability in Artificial Intelligence

Ryan Ahrari

Baskin School of Engineering, University of California, Santa Cruz, USA

Contact the Author

Abstract

As computers have become ubiquitous in industry, AI has developed rapidly in conjunction to maximize efficiency. In particular, as of late, Large-Language Models (LLMs), and the tools that adapt them for common use (ChatGPT [1]) have exploded in popularity. The result of this is that AI will soon be a component of computers that practically all workers will need to understand how to utilize effectively. There then comes a need to understand how AI works and for what purposes it should be used. If an intelligent agent is capable of describing why it has made a decision, it will make it easier for the user to determine whether the agent should be used to automate a specific decision or not. To that end, in this project, I will review and analyze a group of selected papers on the fields of explainability and interpretability in AI. In addition to my own personal written review, I will be including and describing a sentiment analysis that I have done on all of the papers using the Flair NLP tool[2]. The optimistic way in which the topics (explainability, interpretability) are referred to in the papers demonstrates the need for further investment into AI. The goal of this is to develop a standard for designing AI with high levels of interpretability and explainability in mind.

[1] `https://openai.com/blog/chatgpt`
[2] `https://github.com/flairNLP/flair`

Introduction

In the modern world, the use of AI and ML is rapidly spreading throughout all aspects of daily life. Crucially, in the medical field, and in government agencies, intelligent agents and ML models are being used to make vital decisions with greater frequency [AET18]. Even in industries unrelated to technology, such as screenwriting, AI is starting to be seen as a viable option to replace human work[3]. As this bleeding-edge technology becomes critical to many integral human systems and deals intimately with situations that potentially have a mortal impact, it is clear that, as humans, we would want to have a firm understanding of why an agent or model makes a particular decision in a given circumstance.

This is where explainability and interpretability come into play. These terms refer to the reasoning given by an agent or model as to why a decision is made and the understandability of that reasoning for a person, respectively. These ideas work to help us as humans better understand what causes a model to make the decisions it does. An explanation can take many forms, some of which are more visual - such as creating a model that is based on a decision tree [4] - and naturally interpreted well. While other explanations are purely numerical - such as feature-based explanations which show a graph that assigns a number to each feature based on the weight it has on a given decision [BWM21] - and only vaguely interpretable.

Software tools (OpenXAI, LIME, SHAP[5, 6]) that interpret and explain models have been developing over the last decade or so and are currently available for use by the public. With the expansion of computers into virtually all industries, AI that is highly interpretable and explainable becomes essential.

I will first be reviewing the following list of papers that explore interpretability and explainability.

- Sensible AI: Re-imagining Interpretability and Explainability Using Sensemaking Theory (Paper 1)
- Explainability as a User Requirement for Artificial Intelligence Systems (Paper 2)
- Scope and Sense of Explainability for AI-Systems (Paper 3)

[3] https://www.usatoday.com/story/opinion/2023/05/10/wga-strike-pave-way-ai-generated-tv-movie-scripts/70198801007/

[4] https://www.naturalintelligence.ai/explainable-ai-why-it-matters/

[5] https://www.hitechnectar.com/blogs/explainable-ai-frameworks/

[6] https://open-xai.github.io/

- How to Quantify the Degree of Explainability: Experiments and Practical Implications (Paper 4)
- The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations (Paper 5)
- Explainability and Performance of Anticipatory Learning Classifier Systems in Non-Deterministic Environments (Paper 6)
- Explainability's Gain is Optimality's Loss?: How Explanations Bias Decision-making (Paper 7)
- Materializing Interpretability: Exploring Meaning in Algorithmic Systems (Paper 8)
- Balancing the Tradeoff Between Clustering Value and Interpretability (Paper 9)
- Interpretable Machine Learning in Healthcare (Paper 10)

These papers discuss several key aspects on the topics of explainability and interpretability including the applications of an interpretable system in specific contexts; the effect that an explanation can have on other important aspects of AI design (bias, fairness, optimality); defining the degree to which a model is explainable; designing models with explainability and interpretability in mind; and analyzing the performance of models that incorporate higher explainability.

Following that, I will be demonstrating my work with Flair, an NLP library that works in Python. I use Flair for its sentiment analysis tool and run that on the ten papers I reviewed. In this section, I will be showing what Flair tells me about each paper, and exactly how I used the tool itself.

Finally, I will analyze the context of how these topics are being spoken about, and why that is important for the future of the field of AI.

Literature Review

Interpretability and explainability are two key focuses that have emerged in AI, with the motivation of looking to create a common understanding for machines and humans [Lev+21]. The idea of a middle ground of understanding becomes essential when we recognize the degree to which computers have become indispensable. In fact, it is important not only for those who are designing intelligent

agents but for all people who will eventually interact with them, to have some idea as to what led an agent to make a decision [Kau+22]. It seems to me, the reason this is so significant is that people will always, in general, have more trust in a decision, when they can logically decide the same thing themselves based on the reasoning given. This rings true for both interactions between humans and humans, and between humans and machines. As we use models and agents to help make decisions that are of growing importance, trusting decision made becomes more important. Thus it is critical for explanations to be robust, yet interpretable and concise.

Explainability is not a well-defined - or perhaps more accurate to say strictly-defined - field. It seems fairly intuitive that an explanation would give an idea of what the current decision to make is, and what the current environment looks like. However, an explanation can mean something different depending on who you ask. Keeping that in mind it becomes hard to then take a complex world environment and provide a concise, robust explanation that most people would understand. In addition, when two different people look at a particular explanation, they may come away from it with completely different perspectives as to what the correct decision would be. Some believe that the key to the improvement of explanations is through the advancement of cognitive science to better understand the way the human mind works [JS22]. Following that, it is clear that human bias will tend to leak into the explanation given by AI [WBZ22]. The issue with individual biases leaking in is that automated decision systems are being used right now in contexts that have drastic effects on human lives. This makes the constant monitoring of decisions and those who evaluate the validity of the explanations given especially important [WBZ22]. A general goal when creating a model is to ensure that it behaves fairly. What this means is that the quality (fidelity) of the model remains equally high for protected groups [SV22]. There are critiques regarding explainability, which argue that the way that human biases leak in will cause fidelity to fluctuate too much [SV22]. While there have been discovered to be some fidelity gaps (disparities) when two different groups are compared, upon closer analysis it appears the effect of the gap seems to be mostly minimal (about 1% in the worst case [SV22]) on the overall decision accuracy. Beyond the idea that an explanation could be biased, it seems that explainability and interpretability would be - at least slightly - subjective.

The issue that arises is that an explanation can be presented in a way that leads a user into making selecting a model that is not best suited for their intended use [BRP23]. When I speak of explanations there are two distinct types I am referring to, the first is models that are explainable unto themselves, and the decision-making process is clear (glass box models); the second refers to

post hoc explanations given by a tool for a model whose decision-making process is not made clear to the user (black-box models) [Rai19]. For black-box models, it seems that the usefulness of an explanation would be for determining which model makes decisions in a way that is closest to the desires of the user. As such it becomes clear that if the biases of the designer seep too deeply into explainability tools, it would make it difficult for a user to get an objective explanation. This compromises the usefulness of an explanation, as the bias makes it so that a less accurate model can be selected.

A pair of researchers from the University of Bologna have developed a metric, Degree of Explainability (DoX), that works to evaluate the explainability of a model without needing prior knowledge on it[SV22]. DoX, as they define it, is a ratio of the average amount of information that is given in response to archetype questions compared to the entirety of aspects of the model that must be explained. Answering simple and complex archetypal questions (why-not, how, what-reason) and having a set of information that defines the essential aspects - how the explanation will be used, how clear the explanation is, and how often the explanation is used, etc. - is used to calculate a model's DoX [SV22]. These qualities of the DoX calculation help to objectively determine whether or not a given explanation is sufficient for a user. The researchers conclude by saying that DoX could be applied to any explainable AI model given a textual representation of the output [SV22]. While this metric would be useful in theory, it is relying too heavily on the fact that a written, easily interpretable natural language explanation is directly being provided. It seems to me that with this level of provided explanation, an individual user of an ML model can determine whether or not they would trust the model based on their own subjective feelings and intended usage.

It is crucial to thoughtfully go through the process of designing ML models with explainability in mind. There are many ideas regarding what goes into the design of AI that has proper explainability. One team from the University of Michigan proposes the use of a framework which refer to as *sensible* AI [Kau+22]. The approach of sensible AI is to adapt Weick's sensemaking theory - which uses seven fundamental properties - as the foundation of human-machine interaction in designing an intelligent agent. The seven fundamental properties which affect an explanation include an explanation that affirms a person's identity will be seen more favorably; social contexts of an explanation; providing explanations prior to the user's attempt to reflect on the model; the order of explanations are seen in; completion of reading an explanation with interruption; emphasized parts of an explanation; and the plausibility of explanations being prioritized over accuracy. People tend to have a biased perspective

of the world that most of their beliefs are correct, and so most people tend to seek confirmation for their biases rather than trying to find an accurate truth. The seven properties can therefore be used to design explanations with this in mind and prevent individual biases from influencing how to interpret an explanation [Kau+22].

There are thought to be three levels of how interpretable an algorithm is considered to be. The first is the formal aspect of interpretability which specifies the mathematical or numerical way that a model can be interpreted. The second is achievability, which is specific to how to translate the formal interpretation into something that can be understood. The final level or aspect is thought to be the linearity of interpretability, which defines how an interpretation can best be presented to the appropriate audience [BM19]. With these three levels in mind and the notion that we established earlier regarding how people tend to view the world, it becomes clear that the interpretability of a real-world situation becomes harder to define strictly. This comes from the subjectivity of not just the designers of a model but from those who will need to use and interact with the model as well.

The use of AI in real-world environments will make creating concise and interpretable explanations difficult. This is because these situations tend to be non-deterministic, and the next decision is randomized rather than from a fixed set of possibilities to choose from. This leads to the need for algorithms that work to learn and solve problems in explainable ways. An example of this is the rule-based ML model Anticipatory Learning Classifier Systems (ALCS). ALCS uses classifiers (rules) to anticipate the outcomes of taking an action using the current conditions and the world environment [Orh+21]. These classifiers provide a description of the current state and possible effects of trying to get to the next state - this forms a rough explanation. ALCS has a couple of mechanisms introduced that assist them in working with non-deterministic environments, and the idea is that these mechanisms work to avoid states that are incompletely defined and help provide a more robust image of the environment [Orh+21]. These two concepts work together to provide a larger base of possible features for the ALCS to use in choosing the next action. With more features, ALCS will be able to make decisions in less precisely defined environments that loosely describe all real-world scenarios.

Keeping in mind that we are exploring the significance of interpretability and explainability in AI, it becomes imperative to keep a perspective on why we will need to rely on AI. As the world becomes more connected, there are a growing number of interactions between people that become necessary. With this comes the demand to make processes more efficient to support these increasing

interactions without adding delay. As such, the use of AI has come into the equation to relieve the burden of solving a myriad of complex problems for humans. One method that intelligent agents use to form their decisions is node clustering on a graph, in which clusters of nodes are formed based on the similarity of the collective of their features. As the number of features increases, a conflict arises for the models' interpretability. The higher number of features makes it progressively difficult to interpret the clusters graphically. Three researchers from the University of Massachusetts Amherst propose an alternate clustering algorithm that works to make high interpretability correlate with a greater number of features. The algorithm they propose is β-interpretable clustering, which defines $0 \leq \beta \leq 1$ as the fraction of nodes in each cluster that have the same feature values of interest (these would be defined by the user) [SGZ20]. Typical clusters that are k-center suffer from the issue of having a low level of interpretability when looking at them on a graph, but the algorithm proposed by the researchers mixes k-center with partitioning nodes based on the similarity of the features values of interest to create the β-interpretable clusters. The similarity of these features makes it eminently more interpretable to visualize these clusters graphically, at the same time, because the clusters are still k-center, they represent an optimal solution [SGZ20].

It becomes evident, looking at these works, that the key to designing reliable and effective explainable models is striking a balance between their interpretability and their ability to solve problems efficiently. While we typically are searching for optimal solutions for problems that can easily be represented graphically, many real-world scenarios involve too much information to be graphed in any interpretable manner. This is what makes the design of inherently interpretable models so vital.

Sentiment Analysis

Now that we have discussed why interpretability and explainability are fundamental aspects of proper AI design, we will analyze how these two topics are being talked about in the papers we looked at above.

I began by taking the text from each paper, only the raw text of the paper itself, and putting them into a text file - one for each paper. I excluded abstracts, titles, section headings, and parts of the paper that simply had equations, graphs, and figures from the text I planned to run through Flair. My reasoning for excluding the abstract was that it essentially represented a summary of the

exact text I planned to run for analysis. I, therefore, thought that the score I would get for the abstract would be fairly redundant. In addition, because the abstract is not a part of an actual paper it seemed like it would be wrong to add that score to the score for the rest of the paper. Finally, for the prior reasons and the fact that the abstract is typically far shorter than the actual paper itself, it appeared to me that the score of the abstract would be of much less importance. I also excluded the equation sections, appendixes, and the captions of figures/tables as they had characters that could not be analyzed, included repeated ideas, or in the case of appendixes were typically only the most technical aspects of the papers. I then went through and ensured that each sentence of each paper would be on its own line so that when I would read the text file from Python it would be simple to take that sentence and run it directly.

Paper	Sentiment Score	Number of Lines
1 (Sensible AI)	101.8	355
2 (User Requirement)	28.5	81
3 (Scope and Sense)	43.6	270
4 (Quantify Degree)	41.1	214
5 (Paved with Bias)	21.6	297
6 (Anticipatory)	14.4	36
7 (Optimality Loss)	-47.1	210
8 (Materializing)	31.3	66
9 (Balancing)	18	143
10 (Healthcare)	-2.2	182

Table 1: Aggregate Sentiment Scores - Accumulated by analyzing sentiment ($0.5 \leq$ score ≤ 1 positive or negative) of each individual sentence, and adding the positive scores while subtracting the negative scores

Using the sentiment analysis functionality of the Flair tool, I analyzed each paper line by line and recorded the sentiment score for each line, separating the recordings by paper. Flair's sentiment analyzer simply takes in a sentence and returns a *score*, $0.5 < score < 1.0$ either tagged POSITIVE or NEGATIVE. I then proceeded to combine the scores for a given paper by adding the *score* for each sentence tagged POSITIVE and subtracting the *score* for each sentence tagged NEGATIVE, at the end getting a single sentiment score for each paper. This single sentiment score for a paper essentially tells whether more of the sentences in the paper had a negative sentiment or whether more of them had a positive sentiment, and looking at the number of overall sentences gives an idea of the tone used by the authors when speaking on their subject throughout the entirety of the paper. My claim is that this is a fair way to analyze these papers since it cannot be objectively measured

which sentences in the paper would have more impact on an individual reader. Additionally, in an academic context, it can be safe to assume that a reader will read the whole of a paper and use a combination of all the sections to analyze it.

Looking at these scores, it becomes clear that in the overwhelming majority of papers, most sentences were in a positive tone of voice. Looking at the two papers which gave an overall negative sentiment score, we can see that one of them is barely negative at -2.2, meaning that, in reality, it is closer to being neutral than truly negative. So what we find is that only a single paper, Paper 7, really had a negative tone regarding its topic. This particular paper discusses the way in which the bias of humans has led to the creation of explainable AI that makes sub-optimal decisions. The paper is optimistic in the conclusion, proposing ways to prevent bias by monitoring decisions and confidence levels in those decisions.

Discussion

Finally, we will discuss the sentiment scores and their implications.

The Positivity of Most of the Papers

It seems evident to me that most of these papers spoke positively about explainability and interpretability because the authors are optimistic about the relevance of these fields for the future. Clearly, AI is not something that will be used less as time goes on, and with many models already deployed and in production it is crucial to have high confidence in decisions made by these models. It is thus important to continue to build models that are inherently explainable and to develop accurate tools that help to explain existing abstract models.

Why so Negative Paper 7?

Paper 7 is clearly the most overwhelmingly negative paper, with most of them being positive and slightly neutral. I mainly believe that this is due to the fact that paper 7 is the most technical paper in the entire selection, with most of the papers including no math. This technicality makes it more difficult for Flair to analyze the sentiment which is what I believe led to sentences - which would be considered neutral in a human context - being scored as being more negative.

What About Future Work?

While it is clear that explainability and interpretability are regarded as highly important aspects of proper AI design, it seems to me that generally the current work being done in the field is being spoken of just slightly too positively. It is readily evident that most explainability solutions are fairly specialized for a given model, as there is a push to design models with explainability in mind. What this specialization means is that there would need to be a fair amount of resources dedicated to building models that are explainable. While we have already discussed why it is valuable to have good explanations that are highly interpretable, it becomes clear that this idea is constantly being redefined because AI is a subject that is spreading into different sectors of life constantly. This spread further urges investment into building explanations that are understandable for the widest possible audience. The issue is - I believe - that the requisite monetary contribution necessary to get the appropriate explainable models will be too high for those who have the resources to invest in these technologies. Thus, while the optimism surrounding these aspects is well justified, I am slightly pessimistic that the contributions needed to accomplish them successfully will be secured. I am hopeful that with sufficient discussion and analysis of these topics, attention will be paid to ensure that future AI is as broadly explainable and interpretable as possible.

Conclusion

In this project, I analyzed several papers that overview explainability and interpretability in AI. In my literary analysis, I expounded on the important points of the papers and discussed several key aspects of explainability and interpretability such as how they interact with other AI topics, and what it means to prioritize good explanations over optimality and bias. Then I performed a sentiment analysis with Flair, an NLP tool, which revealed an overwhelmingly positive sentiment in the way these two topics are being spoken about. Finally, I discussed that while this overwhelming positivity is beneficial, it may be difficult to get the monetary investment needed to create the explanations we will need in the future.

Ultimately it is very important that explainability is discussed with greater frequency due to the looming AI boom. Referring to making AI inherently interpretable in a positive way will cause more designers of models to take this into consideration. This should motivate those with the available capital resources to invest more heavily in developing explainable AI.

References

[AET18] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. "Interpretable Machine Learning in Healthcare". In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '18. Washington, DC, USA: Association for Computing Machinery, 2018, pp. 559–560. ISBN: 9781450357944. DOI: 10.1145/3233547.3233667. URL: https://doi.org/10.1145/3233547.3233667.

[BM19] Jesse Josua Benjamin and Claudia Müller-Birn. "Materializing Interpretability: Exploring Meaning in Algorithmic Systems". In: *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. DIS '19 Companion. San Diego, CA, USA: Association for Computing Machinery, 2019, pp. 123–127. ISBN: 9781450362702. DOI: 10.1145/3301019.3323900. URL: https://doi.org/10.1145/3301019.3323900.

[BRP23] Krisztian Balog, Filip Radlinski, and Andrey Petrov. "Measuring the Impact of Explanation Bias: A Study of Natural Language Justifications for Recommender Systems". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394222. DOI: 10.1145/3544549.3585748. URL: https://doi-org.oca.ucsc.edu/10.1145/3544549.3585748.

[BWM21] Umang Bhatt, Adrian Weller, and José M. F. Moura. "Evaluating and Aggregating Feature-Based Model Explanations". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. IJCAI'20. Yokohama, Yokohama, Japan, 2021. ISBN: 9780999241165.

[JS22] Mlaan Jovanović and Mia Schmitz. "Explainability as a User Requirement for Artificial Intelligence Systems". In: *Computer* 55.2 (2022), pp. 90–94. DOI: 10.1109/MC.2021.3127753.

[Kau+22] Harmanpreet Kaur et al. "Sensible AI: Re-Imagining Interpretability and Explainability Using Sensemaking Theory". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing

Machinery, 2022, pp. 702–714. ISBN: 9781450393522. DOI: 10.1145/3531146.3533135. URL: https://doi.org/10.1145/3531146.3533135.

[Lev+21] Anastasia-Maria Leventi-Peetz et al. "Scope and Sense of Explainability for AI-Systems". In: *CoRR* abs/2112.10551 (2021). arXiv: 2112.10551. URL: https://arxiv.org/abs/2112.10551.

[Orh+21] Romain Orhand et al. "Explainability and Performance of Anticipatory Learning Classifier Systems in Non-Deterministic Environments". In: *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. GECCO '21. Lille, France: Association for Computing Machinery, 2021, pp. 163–164. ISBN: 9781450383516. DOI: 10.1145/3449726.3459510. URL: https://doi.org/10.1145/3449726.3459510.

[Rai19] Arun Rai. "Explainable AI: from black box to glass box". In: *Journal of the Academy of Marketing Science* 48 (Dec. 2019). DOI: 10.1007/s11747-019-00710-5.

[SGZ20] Sandhya Saisubramanian, Sainyam Galhotra, and Shlomo Zilberstein. "Balancing the Tradeoff Between Clustering Value and Interpretability". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 351–357. ISBN: 9781450371100. DOI: 10.1145/3375627.3375843. URL: https://doi.org/10.1145/3375627.3375843.

[SV22] Francesco Sovrano and Fabio Vitali. "How to Quantify the Degree of Explainability: Experiments and Practical Implications". In: *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2022, pp. 1–9. DOI: 10.1109/FUZZ-IEEE55066.2022.9882574.

[WBZ22] Charles Wan, Rodrigo Belo, and Leid Zejnilovic. "Explainability's Gain is Optimality's Loss? How Explanations Bias Decision-Making". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22. Oxford, United Kingdom: Association for Computing Machinery, 2022, pp. 778–787. ISBN: 9781450392471. DOI: 10.1145/3514094.3534156. URL: https://doi.org/10.1145/3514094.3534156.