Software Engineering for AI

Amir Mehrpanah (`amirme@kth.se`)

1 Introduction to My Research Area

My research focuses on making AI easier to understand, which is often referred to as explainable AI (XAI). There's still a lot of discussion about what "explainability" really means, but we already have some methods to dig into these models and try to figure out how they operate. So explainability is about understanding how AI makes decisions and why it performs the way it does. As AI grows more complex and is applied in areas where mistakes can have serious consequences, this becomes more and more important.

While much of AI research is focused on improving performance, my work is more about ensuring that AI systems are safe, reliable, and trustworthy. This is more related to things like accountability and transparency, which are indispensable for using AI responsibly in real-world scenarios.

For example, imagine an AI that functions as a black box, and you have no idea why it's making certain decisions. I look at how to uncover the patterns or features in the data that the AI relies on to make those decisions. In my PhD, I'm specifically working on explainability for vision models, such as those used for classification, but my interest extends to explainability in all types of AI systems.

Given my background in mathematics, I approach these explainability problems with a formal perspective. I try to employ mathematical tools to analyze and formulate the challenges rigorously. I hope that it results in a more precise and structured examination of XAI methods, which in turn leads to models that can be trusted in practical, real-world scenarios. I realized that without math it is almost impossible to develop objective, robust, and interpretable explanations.

2 Ideas from Robert's lectures

2.1 Validation and Verification in Software vs AI

Explanations for AI models arise from the fundamental need to verify human-made systems before integrating them into our daily lives. Software systems are explicitly defined by humans, making them inherently explainable and interpretable. In contrast, the behavior of AI systems emerges from the optimization of a continuous objective, making it unclear how they arrive at their conclusions. Hence, I believe the behavior of AI models is harder to investigate and attribute to certain parameters or input patterns.

I believe XAI still lacks robust definitions for validation and verification. In terms of validation, we still do not know if we are truly addressing the right questions. Humans do not yet have a clear definition of what identifies a good explanation. This uncertainty is likely a result of XAI research being in its early stages.

When it comes to verifying explanations, I think we really need better ways to check if a model is making the right decisions for the wrong reasons. It's a lot harder than regular software testing because models can seem super accurate but still rely on totally flawed logic. Unlike software, where we can usually predict how things will behave, AI models are optimized for specific outcomes, which can sometimes lead to surprising or even weird results.

For example, a model might assume dogs always appear on grass and wolves on snow, which doesn't hold up in the real world. This kind of faulty logic could make it label a dog on snow as a wolf. Right now, there's no clear way to measure or explain this kind of behavior for individual examples, and that's a big challenge we still need to figure out.

Maybe one important aspect that is usually missed when comparing software testing and AI explainability, is the fact that variables defined by human are usually human understandable. On the other hand, input variables of an AI model can be pixels, and then the model creates different variables based on a 3x3 block of pixels. Despite the effort to make sense of these variables, the interpretation of those variables is still a mystery.

2.2 AI's Hidden Technical Debt

This phenomenon can also be seen in AI, where researchers often use common approaches simply because of convenience. From the perspective of an AI researcher who is focused on showing what is possible to do with AI, almost everything can be tried. They do not care if their decisions make sense in mathematics or if it can be justified in all cases.

However, from the viewpoint of a XAI researcher, including myself, these decisions have accumulated and created a huge hidden technical debt. Because we do not know what are the long-term implications of such decisions, how they interact with each other, and how they can be explained.

This issue is particularly everywhere in the AI community, where many large-scale models have become black boxes. Despite their capabilities, we lack a clear understanding of how these models make decisions.

3 Ideas from the Invited Speakers

3.1 Adopting Transparent AI Transparent into Software Development

Many companies are fearing that they are falling behind in the AI race. This concern makes them to try to integrate AI into their software by any means or use AI as a software development tool at the very least. This urge is in conflict with the safety concerns of SAAB and many other companies that are active in critical domains, such as military. Therefore, they cannot simply use AI tools provided online as it risks the unwanted leakage of information.

I think this is the direct result of AI models being opaque. As an example, we cannot make sure that these models have memorized parts of the code used in their training phase, neither we know how to remove or recover such information if any.

3.2 Using AI for Writing Tests

As I pointed this out in class during the Dhasarathy's presentation, I think it is safer to write the tests first, then let AI generate the code that passes the tests. I think this is much safer and becomes easier for AI in the long run, since it is getting better and better at filling the gaps every day.

It is important to note that writing tests is where a human judgement is necessary. Assuming that there is a bug in the code, and letting the AI decide to write the test for it, gives the AI the freedom to choose if it should pass the test or not. The AI model then may decide by mistake to incorporate a behavior coming from a bug into the test it generates, hence letting a bug pass as an expected behavior. This can lead to drastic consequences if the bug gets noticed when the car is in the streets.

This problem gets worse when we note that AI models are currently black boxes, so no one can make sure if the decisions that model is producing follows a certain logic, which is writing the correct test for a wrong code.

4 Discussions on Papers

4.1 Novel Contract-based Runtime Explainability Framework for End-to-End Ensemble Machine Learning Serving [2]

This paper talks about making machine learning (ML) services more understandable and trustworthy, with an important constraint that they're used in real-time applications. This constraint that XAI methods should be real-time is not relevant at the current state of my research.

They introduce a new framework that allows AI service providers "explain" (more on this later) how their models are working real-time. This includes showing consumers the quality of their predictions, for example by showing accuracy, a measure of confidence.

They do this by adding explainability rules directly into service contracts (a contract between the service provider and the consumer), so that they know what to expect for. Another constraint in this paper is that it focuses on ensemble of AI models, where multiple models work together to improve predictions. This is a quite common practice in machine learning to use voting of multiple models to reach to a conclusion. Nonetheless, it makes it hard to provide one explanation for multiple models, specially when there are many of them.

The framework proposed in this paper makes it easier to monitor these systems and adjust them in real time based on consumer feedback, and the claim is that it improves service quality and trust. The authors test their approach on two real-world applications: malware detection and CCTV surveillance, to verify their claims that it works in practice.

I found this paper particularly interesting as it shows how the definition for explainability of AI systems can change from context to context. Here, explainability is defined as companies being ”honest” with the consumer and show them some of the classical metrics. In contrast, in my thesis we define it very much differently, hence it cannot be achieved merely by showing a few numbers.

I believe my research can give more depth to the report that a consumer might ask for, both in terms of content and type. So on top of classical metrics, we can pinpoint some of the features that has led to the current outcome. This way, the consumer can make sure that the outcome is fair, and is not based on wrong reasons.

4.2 Towards a roadmap on software engineering for responsible AI [1]

This paper focuses on creating practical ways to responsible and ethical AI systems that can be trusted. It points out the fact that we have many high-level ethical principles, yet there is a gap in using those ideas in actual products. This kind of work is important because it helps us bring AI into the real world responsibly. Also, it is related to my research, because XAI can be one way of making sure that AI systems are actually doing what we think they are doing. Hence, they suggest three ways to fix the issue:

- Having rules or guidelines at different levels. Including industry, organization, and teams. This way, we make sure everyone developing AI is on the same page.
- Having ethics embedded in the development process. This means that fairness, privacy, and safety are checked at every step.
- Designing AI systems while having responsibility in mind. This translates to techniques like accountability tools or privacy-friendly methods like federated learning.

This kind of work is important because it helps us bring AI into the real world responsibly. Finally, this paper is related to my research, because XAI can be one way of making sure that AI systems are actually doing what we think they are doing.

So one way of producing responsible AI systems is designing tests that uncover the inner workings of an AI system but designing tests requires a mental model of how things work, which I believe requires XAI.

References

[1] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, and Zhenchang Xing. Towards a roadmap on software engineering for responsible ai. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, CAIN '22, page 101–112, New York, NY, USA, 2022. Association for Computing Machinery.

[2] Minh-Tri Nguyen, Hong-Linh Truong, and Tram Truong-Huu. Novel contract-based runtime explainability framework for end-to-end ensemble machine learning serving. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, CAIN '24, page 234–244, New York, NY, USA, 2024. Association for Computing Machinery.