Why is the current XAI not meeting the expectations?

Alessio Malizia and Fabio Paternò

Imagine going to space and deciding between Spaceship 1 and Spaceship 2. Although it has never been in flight, Spaceship 1 comes with precise equations outlining how it operates. Even though it is unknown how Spaceship 2 flies, it has undergone considerable testing and years of successful flights, including the one you are about to take. Cassie Kozyrkov, Chief Decision Scientist at Google, posed this dilemma at the World Summit AI in 2018.
We cannot provide a solution to this question because it is philosophical and perhaps generates a more profound inquiry on which better inspires trust - explanation or testing.

For a while, it appeared that one issue with Artificial Intelligence (AI) algorithms, particularly cutting-edge deep learning techniques, was that they were black boxes. It was impossible to pinpoint the precise reason the program predicted a particular outcome in a specific circumstance. Due to this lack of interpretability, businesses and governments were hesitant to use AI in critical sectors such as healthcare, banking, and government. So much so that the EU Commission released its AI package in April 2021, including an AI act, recommending new laws and initiatives to make Europe a relevant hub for reliable AI, for example, in the case of the use of AI in high-risk sectors.

However, in the last few years, researchers, scientists and businesses have increasingly sought ways to provide some understanding of how AI algorithms get to a decision. Their respective AI programs are claimed to be "explainable" due to their post-hoc interpretations of how an AI has made decisions.

One option is to try to provide explanations of black-box machine learning using interpretable models (e.g., decision trees, rule sets, and analytical expressions) (Rudin, 2019). Some have implied a tradeoff between accuracy and explanation, but that is not necessarily the case. Recent research has shown that interpretable machine learning methods can be as accurate as blackbox learning methods on tabular datasets (Caruana and Nori, 2022) and can be especially attractive in situations where interpretability and transparency are important, such as in legal or medical contexts where decisions need to be explained and justified (e.g. (Wang et al., 2020)). On the other hand, deep neural networks have shown remarkable performance in tasks involving image and natural language processing. More

research is necessary to address the limitations of both interpretable models and deep neural networks.

Numerous papers argue for using XAI methods in the literature, as well as multiple suggestions for brand-new XAI family approaches. Nevertheless, finding instances of practical XAI technique implementations that have enhanced the business in industry/societal/real-world applications is more challenging, even if some interesting work in this area has been put forward, for example in the health domain (e.g. Lengerich et al., 2022). Certainly, explanations are not “one size fits all”; an explanation that is understandable to a technical audience might offer little explanataory value for a non-technical audience.

Case studies of the employment of XAI approaches tackling real-world machine learning issues are still lacking. Such case studies would help to clarify what is currently feasible and what is not feasible when employing XAI techniques. This is particularly true in the healthcare industry, where artificial intelligence has significantly progressed in systems that interpret medical imaging automatically. The issue is that most of the XAI techniques the healthcare sector employs are limited (Kahn, 2022).

For example, a saliency map is a well-known explainability technique. It uses the image that was fed to the algorithm to produce a heat map of the areas that the AI program gave the highest weight when generating a forecast. However, as shown in one study published in 2021 by the medical journal Lancet Digital Health (Ghassemi et al., 2021), the heat map that was supposed to explain why the AI system classified the patient as having pneumonia covered a sizable portion of one lung's quadrant, with no further explanation of what precisely it was in that area that the AI system considered to be pneumonia. As mentioned in the study: “The clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factors, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease.”
The authors note that in the absence of such information, people tend to presume the AI focuses on whatever attribute they, as human therapists, would have thought was most crucial. Doctors may not be aware of the mistakes the machine learning system may make due to this cognitive bias. Ghassemi et al. also uncover issues with other well-liked explainability techniques, such as GradCam, LIME, and Shapley Values. Some of these techniques serve as a form of counterfactual by changing the data points that are entered until the algorithm generates a different forecast, at which time it is assumed that those data points must have been the most crucial for the initial prediction. These techniques, however, share the same drawback as

saliency maps. While those techniques might be able to identify elements that the algorithm deems important, they are unable to explain to a doctor why those elements are significant from a medical standpoint, e.g., whether the algorithm relied on significant structural alterations in the shape of an organ or the results were influenced by some noise in the training data.

Furthermore, a recent study (Krishna et al., 2022) found that various state-of-the-art explanation approaches regularly disagreed on the rationale for an algorithm's conclusions. Most people who used the algorithms in real-world situations had no method of resolving such disparities and, therefore, might just choose the explanation that most closely matches their pre-existing notions.

The ability to explain the causality behind an AI-based decision does not mean that the AI system uses this causality and that the actual relationship between inputs and outputs might differ. AI systems are often opaque, black-boxed systems.

Why XAI's success really lies in pushing truly cross-disciplinarity work

Governments consider the potential of XAI to address concerns about the obscurity of algorithmic decision-making with AI. Although XAI is enticing as a solution for automated decisions, using XAI is difficult because of the wickedness of the problems governments face. Wickedness refers to the ambiguity of the facts that characterize a problem and the lack of agreement on the normative standards for resolving it (de Bruijn et al., 2022).
Additionally, these "solutions" to wicked problems frequently evolve over time (Rittel & Webber, 1973). It is difficult to explain something that is unclear especially if explainability is viewed as a strictly technical issue.

The need for explainability and interpretability in AI is a much larger cross-disciplinary problem that requires a more comprehensive solution than XAI alone can offer. Interpretable models for non-technical people suffer from the same problem: technologists are needed to translate technical explanations. Different kinds of explanations are needed for business executives, risk managers, doctors, bankers and officers, i.e. end users.
AI is more likely to serve the interests of the powerful if the aims of explainability from various communities are not clearly stated. Companies adopting AI should be as honest as they can be about how, why, and for what purposes they are using XAI approaches. The National Institute of Standards and Technology (https://www.nist.gov/) is one of the organisations developing XAI standards and regulations. Such organisations should be aware of the current limitations of XAI in practice and seek out diverse expertise on better matching incentives and governance with a comprehensive understanding of XAI objectives. Ethics frameworks might come to the rescue, and the NGOs and business sectors have published a deluge of AI ethics and standards in the past few years. However, these

values are isolated ones that serve corporate goals and are embedded in an industry that often disregards ethical behaviour unless enforced by the justice system. In a recent paper entitled *The uselessness of AI ethics* (2022), Luke Munn gives evidence that such ethics principles are meaningless or isolated at most. The result is a disconnect between high ideals and technological reality. Even if this gap is acknowledged and ideas attempt to be "implemented," it is challenging to translate complicated social concepts into technical rule sets.

We can only achieve the objectives of intelligible, dependable, and controllable AI in practice with the active participation of many stakeholders from the social sciences, computer science, civil society, and industry. Consider, for instance, the disparate requirements of engineers and consumers in explaining an AI system. Developers may use Google's What-If Tool to examine intricate dashboards that visualise a model's performance in many fictitious scenarios, evaluate the significance of various data elements, and assess multiple fairness theories. On the other hand, users could choose something more focused. It might be as easy as telling a user whatever factors, such as a late payment, resulted in a point deduction in a credit score system. Various consumers and scenarios will require different outputs.

Therefore, the explanatory domain must be improved, and its audience must be expanded if we want to reach the grail of trust and confidence in judgments made by black-box AI. In addition to XAI tools for technical teams, what we need is "Understandable AI" (Habayeb, 2022) or AI that serves the needs of non-technical stakeholders. It is helpful to explicitly compare their goals to understand how practitioners in different domains have different expectations for what they hope to achieve by building XAI systems.

Whilst the issue with XAI, as it stands right now, is that many of the current approaches view explainability as a purely technical matter, we think the future success of XAI lies in fostering a genuinely cross-disciplinary approach among AI and other fields of interest, such as ethics, law, psychology, sociology and human-centred design to name but a few. In essence, statistical, mathematical, and scientific analyses are pretty valuable tools. However, it is all too easy to misinterpret their measured certainty as the only "true" method when, in fact, it is just one tool and one tactic—and not one that can be translated or used to explain all qualitative occurrences. We consider situations in which the effect is present and suggest a cause. Still, we overlook all the occasions in which the same cause resulted in no discernible consequence or an entirely different outcome. Model-based storytelling is quite simple. It is often difficult to remember that they are stories, nevertheless.

There is evidence of winning the rigid argument of more technical researchers versus more humanistic approaches thanks to the need for a better encompassing approach required for XAI to provide explanations for the practitioners or the general public. As (Miller, 2019) points out "explanations are not just the presentation of associations and causes, they are contextual. While an event may have many

causes, often the explainee cares only about a small subset (relevant to the context)". In addition, the relevant explanations should be given using the users' language.

The aim is to meet the domain experts' needs and expectations, for instance, by identifying interactive environments that allow people with different backgrounds to communicate and reason on recommendations made by AI. Human-centred design methods can offer tools to foster the design of cross-disciplinary collaborative systems. In particular, meta-design (Fischer et al., 2004) can offer the participatory and cross-disciplinary approach needed to meet non-technical decision-makers expectations. It is a conceptual framework aiming to define and build the social, economic, and technical frameworks necessary for new kinds of collaborative design to function. It comprises several useful design-related tools to help users accomplish this task.

Through tools offered by meta-design, decision-makers can render AI decisions understandable and valuable for their work. For instance, designing interactive systems that can manage the questions and associated answers that domain experts usually pose to AI systems, ranging from medical image interpretation to home automation or conversational agents.

The XAI Question Bank proposed by Vera Liao et al. (2020) is an excellent example of presenting questions that can be used to probe the AI system. For instance, the type of data the system learns from and the related output, how accurate the system is, and how it makes a prediction. There might be questions related to a specific prediction: what would be predicted if an instance changes to a different value, how to change an instance to get another prediction, or what is the scope of change permitted to get precisely the same prediction? Only then would it be possible to create environments that allow the domain experts to configure how decisions are translated in the user language, not that of the AI system, and the desired level of interactivity and multimodality.

In "Why should humans trust AI?" Carroll (2022) proposes to model explanations as inherently pragmatic, conversational, and social. It is always a question of making sense, being aware, and negotiating in a vast sense, as well as of the responsibility people accept and show for one another as they engage in daily encounters. This might be a significant accomplishment for XAI, but we will not get anywhere if we do not acknowledge the problem and the obstacles in our way.

Conclusions

XAI per se shows many limitations in its current form due to an excessively technical approach, often requiring technologists to help end users fathom the explanation provided by a model. Moreover, as demonstrated by Bordt et al. (2022), explanations produced by existing AI approaches depend on a variety of particular characteristics

of the AI system, such as the training data, the precise shape of the decision surface, and the selection of one explanatory algorithm over another. Given that programmers and AI developers are free to select these factors, there is a risk that, despite an explanation's seeming plausible, it occurs only because of some hidden layers computing weights depending on some of the features and does not truly reflect the explanation given at all. The creation of explanations that cast doubt on certain features of AI systems is not something AI developers are interested in doing, at least not right now.

Furthermore, the explanatory approaches mentioned above are just a small set of tactics that can be employed to provide explanations. Indeed, the XAI field will need contributions from different areas to explore more types of explanations.

Nonetheless, we believe that the winning point of XAI truly lies at the intersection of different disciplines like ethics, law, psychology, sociology and human-centred design. XAI could serve as a starting point to develop a common language among scholars from different disciplines, thereby accelerating its real-world impact.

References

Bordt, S., Finck, M., Raidl, E., & von Luxburg, U. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. *arXiv preprint arXiv:2201.10295*.

Carroll, J. M. (2022). Why should humans trust AI?. Interactions, 29(4), 73-77.

Caruana R., Nori H.. Why Data Scientists Prefer Glassbox Machine Learning: Algorithms, Differential Privacy, Editing and Bias Mitigation, SIGKDD 2022, https://doi.org/10.1145/3534678.3542627

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly, 39(2), 101666

Fischer, G., Giaccardi, E., Ye, Y., Sutcliffe, A. G., & Mehandjiev, N. (2004). Meta-design: a manifesto for end-user development. *Communications of the ACM*, *47*(9), 33-37.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745-e750.

Habayeb, A. (2022, February 16). Explainable AI Isn't Enough; We Need Understandable AI. https://www.techopedia.com/explainable-ai-isnt-enough-we-need-understandable-ai/2/34671

Kahn, J. (2022, March 22). What’s wrong with “explainable A.I.” https://fortune.com/2022/03/22/ai-explainable-radiology-medicine-crisis-eye-on-ai/

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01602*.

BJ Lengerich, ME Nunnally, Y Aphinyanaphongs, C Ellington, R Caruana, Automated interpretable discovery of heterogeneous treatment effectiveness: A COVID-19 case study, Journal of biomedical informatics 130, 104086, Jun 2022

Liao, Q. V., Gruen, D., & Miller, S. (2020, April). Questioning the AI: informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-15).

Miller, T. (2019), Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, 267, 1-38.

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics*, 1-9.

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences*, *4*(2), 155-169.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Wang, C, Han, B, Patel, B, Mohideen, F and Rudin, C. 2020 In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction, 11 Mar 2022, https://arxiv.org/abs/2005.04176