Explaining Explainable AI

Swaroop Panda
IIT Kanpur
pandas@iitk.ac.in

Shatarupa Thakurta Roy
IIT Kanpur
stroy@iitk.ac.in

ABSTRACT

An aspect of User friendly AI involves explanation and better transparency of AI. Explainable AI(XAI) is an emerging area of research dedicated to explain and elucidate AI systems. In order to accomplish such an explanation, XAI uses a variety of tools, devices and frameworks. However, some of these tools may prove complex or ambiguous in themselves, requiring explanation. Visualization is such a tool used extensively in XAI. In this paper, we examine how such tools can be complex and ambiguous in itself and thus distort the originally intended AI explanation. We further propose three broad ways to mitigate the risks associated with tools, devices and frameworks used in XAI systems.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Explainable AI, Visualization

1 EXPLAINABLE AI

Explainable AI(XAI) is basically the idea of elucidating AI systems, which are otherwise thought of as opaque black boxes. A DARPA document [4] suggests that though AI provides immense benefits, the effectiveness can be supposedly reduced by the inability to explain and be transparent about decisions and the subsequent decision-making. XAI, then, facilitates trust of these AI systems and enables the user to understand the inner mechanisms. The work by Doran et al. [2] suggests what devices and frameworks the AI community approaches the idea of *explainability* in XAI. This includes dealing with opaque systems, interpretable systems and comprehensible systems. The work by Goebel et al. [3] suggests the need and the scope of XAI across domains for the need of transparency and ethics in the AI systems.

2 VISUALIZATION AS A TOOL FOR XAI

Visualization is an important tool used in XAI. It is used to explain the underlying decision making process, display the algorithm mechanism, visually demonstrate the live-performance of an AI system, clarify the results and so on. Visualization has found its utility across the different stages of a variety of AI systems. For instance, the journal that published the work by Wattenberg et al. [7], Distill, is dedicated for explaining and interpreting AI systems and extensively uses visualizations to explain different concepts. More specifically, the work by Wattenberg et al. [7] visually demonstrates an effective method to work out a t-sne algorithm (a nonlinear dimensionality reduction technique). The text is accompanied by animated and interactive visualizations to demonstrate the use of the algorithm. By this, and many other examples across the XAI literature, visualization is not just a specific framework or a device for *explainablity* in XAI, but rather a tool that is used across all devices and frameworks used in XAI.

3 THE PROBLEM WITH TOOLS SUCH AS VISUALIZATION

Visualization, as an independent subject, has been studied and researched extensively. Visualization research is now a mature and advanced domain of study. One of the major ideas emerging from visualization research is that visualization, in and of itself, contains many complexities, ambiguities and pitfalls that are required to be addressed. The work in Cairo's book How Charts Lie [1] and Munzner's book Visualization Analysis and Design [6] discusses some challenges that are encountered in using visualization as a tool to explain, interpret or make an inference. More specifically, Cairo's book [1] suggests how different visual data charts misrepresents and distorts our reading and inference of data and information.

If visualization, in and of itself, has to be explained, in order to diminish the ambiguities, complexities or any other challenges; then its extensive use as a tool in XAI needs to be reflected upon. This is because the tool used to explain AI systems, when itself needs an explanation, can potentially distort the originally intended explanation of the AI system. For instance, an advanced interactive visualization used to demonstrate an AI algorithm, may require an explanation itself. Such an explanation obfuscates the original intent of the interactive visualization. Moreover, if the explanation of the tool is further complex or ambiguous, then it itself requires explanation leading to recursion (till a satisfactory and an unambiguous explanation is reached). Such a recursion of explanations severely distorts and extends the intended elucidation of the AI system.

We suppose, this aspect can be true for other devices and frameworks (which may be more advanced than simple visualizations) used for XAI. The recursive risk is an inherent property of all tools, devices and frameworks used in XAI; when explanations themselves require explaining. The basic conundrum in XAI, then is, how to use tools and devices to explain AI systems, that do not require (or require minimal) explanation themselves; and how to deal and mitigate the recursive risks present in XAI.

MuC'20 Workshops, Magdeburg, Deutschland

https://doi.org/10.18420/muc2020-ws111-347

4 THREE APPROACHES TO RESOLVE THE CONUNDRUM

The way out of the dilemma is to be mindful of the devices and tools used in XAI. One simple way to deal with recursive risks is to use devices and frameworks which require existing or minimal explanations. For example, using an interactive visualization framework that has been intensively studied (by the visualization research community) can be very effective for XAI. All the complexities and challenges with such a framework would ideally have already been addressed. Or for example, using a device like annotation and labelling to describe an opaque AI system. Such a simple device needs no further explanation and the recursion stops right there. However, as a caveat, annotations do not include advanced mathematical notations or equations from different domains used for describing the AI system. A shortcoming of this approach is that only few rudimentary AI systems can afford such simple explanations. Systems involving heavy domain-knowledge or technical expertise aren't particularly explained by such simple explanations.

Another approach to deal with recursive risks is to inculcate uncertainty to all XAI devices or mechanisms. This uncertainty represents the fact that no matter what tools, devices or frameworks are used for XAI, there will be certain ambiguities and complexities present within the system that cannot be completely eradicated. These ambiguities are a result of tool-explanations distorting the originally intended explanation of the AI system. And such tools inevitably have to be used - maybe because of domain complexity, mathematical rigor or for technical correctness. In these situations, it is optimal to instill the notion of uncertainty to the XAI system. This uncertainty can either be a quantitative one (like Heisenberg's uncertainty Principle in Physics); by using quantitative metrics in XAI Usability tests. Or it may be a qualitative assessment of the XAI systems. The shortcoming of this approach is that calculating and instilling uncertainty in XAI systems is tough. A quantitative uncertainty metric involves performing a lot of usability experiments of the XAI systems with the users and grasping how well (or how much) the users understand the system. Such an approach also includes modelling the user and subsequent creeping of biases into the system. Qualitative approaches have their own unique challenges.

Another approach to deal with such risks is to make the user AI-literate. User literacy is an important and emerging topic of research in the visualization [5] and the broader HCI community. The basic idea is not to entirely burden either the system or the user with the explanation; but rather finding a balance between an AI system and the user. This is especially useful in specialized domains where devices and frameworks require explanations themselves. Some of the burden of explanability then falls on the user. The user has to learn some rudimentary domain knowledge or grasp some concepts. The XAI system then includes the AI system as well as the user. An example of this is the use of AI systems in a high-stakes domain like medicine or aerospace. In such high-stakes and sensitive domains (where risks can trigger into catastrophic consequences), explanations as burdens is best shared by the system and the user. Moreover, user-literacy also brings upon the ideas of transparency and ethics into the XAI system. This is because, a literate user has more agency and control over the AI system. The challenge of this approach involves educating the user. The users emerge from different linguistic and cultural backgrounds. Thus, new variables, that are required to be addressed, now emerge in the XAI system.

The basic premise behind these three approaches is that the *explanations* required to explain AI should not require explanations themselves; and therefore, not distort the AI explanation. However, the three approaches, have their respective shortcomings. Thus, there arises a need to be mindful of the tools used in the XAI system and be context-sensitive towards the deployment of such tools, frameworks or devices. Also, the three approaches have been proposed in context of visualization as a tool in XAI. Visualization is backed by a mature and a strong literature. Uncertainty in visualization is acknowledged and is a growing topic. Visualization literacy among users is also gaining ground in visualization research. We suppose that our proposed approaches extend to other tools and frameworks in XAI as well.

5 CONCLUSION

In this paper, we looked at how eXplainable AI (XAI) uses tools, devices and mechanisms to accomplish explanations of complex and obscure AI systems. These tools sometimes are ambiguous in themselves and require further explanation. We took an example of such a tool, visualization, that is extensively used in XAI systems. Visualization, as we saw, has its own pitfalls and ambiguities; thus requiring an explanation itself. Such a recurring explanation distorts the originally intended explanation of the AI system. To mitigate these risks we proposed three solutions - to use tools that have existing or minimal requisite explanations, to inculcate uncertainty into all the XAI systems and finally to make users AI literate. Further it would be good to explore instituting these three approaches in a variety of XAI systems.

REFERENCES

[1] Alberto Cairo. 2019. *How Charts Lie: Getting Smarter about Visual Information.* WW Norton & Company.
[2] Derek Doran, Sarah Schulz, and Tarek R Besold. 2017. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794* (2017).
[3] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. 2018. Explainable AI: the new 42?. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction.* Springer, 295–303.
[4] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).
[5] Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2016. Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 551–560.
[6] Tamara Munzner. 2014. *Visualization analysis and design.* CRC press.
[7] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. 2016. How to Use t-SNE Effectively. *Distill* (2016). https://doi.org/10.23915/distill.00002