Google Comments on NISTIR 8312: Four Principles of Explainable Artificial Intelligence

October 15, 2020

Google welcomes the opportunity to provide comments in response to the National Institute of Standards and Technology's Four Principles of Explainable Artificial Intelligence. We support the report's focus on enabling trust and confidence in AI systems as a means to promote the adoption of AI to address many of society's most pressing challenges, and welcome NIST's contribution to establishing a common understanding and language around AI explainability.

As the report notes, transparency and accountability are essential to public trust and adoption of AI, and explainability can play an important role in building that trust. We appreciate the nuanced approach of the report to this complex technical, psychological, and social challenge, and the report's emphasis on the need to tailor explanations to different applications, contexts of use, and audiences.

General comments

Frame explainability around the needs of stakeholders: Transparency is not an end in and of itself — it is a means by which to enable accountability, empower users, facilitate appropriate use by third parties, and build trust and confidence. Explainable AI principles should be clearly tied to what they are trying to achieve and how different types and attributes of explanations can help to meet specific needs of key stakeholders in a given context.

In general, NIST's four principles provide a clear articulation of some key considerations that should be taken into account in designing explanations for AI systems, and provide some guidance around these explanations might be implemented. However, while the report emphasizes the importance of tailoring explanations based on context, there appears to be an implicit assumption that all AI systems need to be explainable, and that all explanations should meet all four of the principles.

In practice, not only would it be difficult, if not impossible, to provide explanations that adhere to all four principles for all AI systems in all contexts, but in many cases explanations may not be desirable or appropriate. For example, users may not care to see why an app recommended a movie they do not want to watch when they can simply scroll on to a recommendation that appeals to them.

The report could better articulate the practical limitations of the four principles and how they should be applied in proportion to a system's risks and the needs of stakeholders. While AI

systems should aspire to these principles in proportion to their risks, explanations will not be desirable or appropriate in some contexts, and in others explanations that adhere to some, but not all, of the principles may be the best approach in light of the tradeoffs of adhering to all of the principles (as discussed below). For systems that present greater risks, for example systems that review loan applications, explanations that adhere more literally to the four principles may be necessary, while for systems that present relatively limited risk, like movie recommendations, explanations may not be required at all.

Separate explanations from outputs: The report links explanations directly with system outputs. The first principle calls for AI systems to "deliver accompanying evidence or reason(s) for all outputs," seemingly calling for all systems to be self-explainable. In some cases, for example systems built on Deep Neural Networks (DNNs), it may not be possible to fully explain how a system produced a specific output, but that does not mean that the system is completely opaque, or that key stakeholders cannot get the information they need to make informed decisions.

Rather than focus on providing evidence and reasons for individual outputs, in some contexts it may be more appropriate and meaningful to provide explanations of how the system works in general. Frameworks like Model Cards can provide meaningful and accurate explanations of a system's design, limitations, appropriate uses, and risks that, in some cases, can better enable informed decision-making by users, customers, regulators and the public than explanations of specific outputs.

In other cases, it may be difficult for a system to provide its own explanations, but an external analysis of the system using techniques like TCAV may provide valuable insights into the behavior of the product. Ex-post external analysis of how a system produced a specific output that led to harm could be more useful for investigating incidents and ensuring accountability than attempting to build detailed explanatory functions into all products.

Acknowledge inherent trade-offs of the principles: The four principles each carry significant tradeoffs, both with each other, and with other considerations like accuracy, safety and robustness. Take as an example a model like GPT-3 with 175 billion weights. An explanation that "correctly reflects the system's process for generating an output" based on the complex interaction of 175 billion parameters would be incomprehensible to a human being, and computationally infeasible to generate for each individual output of the model. Often, meaningful explanations will require simplification, generalization or analogies which might not meet the standard suggested by a literal reading of the principles of explanation and explanation accuracy.

Explanations can also require tradeoffs with other equities. Using a much simpler model than GPT-3 to enable a literal interpretation of all four principles, for example, could result in significantly lower accuracy rates for many tasks. In other cases, providing detailed explanations

could expose intellectual property or infringe on privacy rights by directly or indirectly exposing sensitive data. Explanations can also pose risks to safety by exposing the inner workings of AI systems in ways that allow malicious actors to exploit, manipulate, or game systems meant to protect users and the public.

***

Google welcomes NIST's thoughtful and timely contribution to the development of Explainable AI, and supports NIST's ongoing efforts to advance a common understanding of transparent, accountable and trustworthy AI systems.

We appreciate the opportunity to comment on the report and welcome any questions, feedback, or opportunities for further discussion.