Arm, Inc.
comments to
Draft "NISTIR 8312, Four Principles of Explainable Artificial Intelligence"
October 15, 2020

On behalf of Arm, we are pleased to have the opportunity to submit the following comments on the draft National Institute of Standards and Technology Interagency or Internal Report ("NISTIR") 8312, Four Principles of Explainable Artificial Intelligence. More broadly Arm is supportive of the work NIST is performing to examine a framework for trustworthy artificial intelligence ("AI"), of which explainable AI is an important component.

Arm is the leading, and largest supplier of intellectual property in the semiconductor sector. Arm provides central processing units, graphics processing units, and neural processing units, as well as a portfolio of other technologies that enable AI from the cloud to endpoint devices and everywhere between. In fact, the majority of consumer interactions with AI today are likely happening on Arm-based technology offered by Arm's partners. As such, Arm has taken seriously the ethical implications posed by AI, and the need to ensure appropriate protections are taken to address those implications and ensure the public can trust this emerging technology.

As a call to action on this in November 2019 Arm released the AI Trust Manifesto, setting out six principles Arm believes companies must address to demonstrate and establish public trust for AI.[1] As such, Arm supports efforts by NIST to address components of trustworthy AI, and believes the ability to explain AI outputs is essential to trust.[2] Arm included explainability as one of the six principles in Arm's AI Trust Manifesto, explicitly stating:

> *3/ WE BELIEVE AI SHOULD BE CAPABLE OF EXPLAINING ITSELF AS MUCH AS POSSIBLE: WE URGE FURTHER EFFORT TO DEVELOP TECHNOLOGICAL APPROACHES TO HELP AI SYSTEMS RECORD AND EXPLAIN THEIR RESULTS Where appropriate, the way an AI system works should be capable of being transparent, and the decisions that result from it should be explainable to a human interrogator including non-specialist users of AI.*[3]

Comments on NISTIR 8312

In general, Arm agrees with the ideas expressed in the Introduction. Explainability is one of several properties that will build public trust in AI systems.[4] Similar to NISTIR 8312, the Arm AI Trust Manifesto also expresses the need to address accountability and bias. The aim of Arm's inclusion of "security" as a key principle in the Arm AI Trust Manifesto, contributes to two other

---

[1] See Arm AI Trust Manifesto https://www.arm.com/blogs/blueprint/wp-content/uploads/2019/11/Arm-AI-Trust-Manifesto-2019.pdf
[2] NISTIR 8312 suggests this as well, lines 128-132.
[3] Arm AI Trust Manifesto.
[4] NISTIR 8312, lines 133-135.

properties included in NISTIR 8312, "resiliency" and "reliability" which are dependent on security.[5]

2. For Principles of Explainable AI

The four principles of explainable AI set forth in the paper are each well defined, and important components of building trust in AI.[6] In particular, the discussion throughout this section of the different needs for different user groups is very important, as is the insinuation that flexibility may be needed under each principle to make an explanation relevant.

The discussion of the "meaningful" principle under 2.2 is important, particularly that an explanation is not one size fits all.[7] This flexibility is important given the wide variety of applications in which AI will likely be used, and the wide range of technical expertise among users. The ability of different user groups to be able to meaningful understand or need different types of explanations makes sense, as does the discussion about the idea that user groups abilities and understanding may evolve over time and explanations that were sufficient for a user group in the past may no longer be sufficient.[8]

Additionally, the discussion of "explanation accuracy" principle in 2.3, also provides important perspective on the need for flexibility in explanation accuracy metrics. A 100% accurate accounting of an output decision process may be too detailed or technical for a user to understand, and that to achieve the "meaningful" principle the accuracy metrics may need to be adjusted.[9]

The discussion of "knowledge limits" in 2.4 is an important component of establishing trustworthy AI. A system that provides an output simply because it is queried, but without a high level of confidence would certainly diminish trust. That said, based on the discussion in the paper, a "knowledge limit" seems like a distinct capability of an AI system rather than a principle of explainability.

While discussion of different types of explanations and audiences is important, it is also important to note that providing an explanation for every AI output is likely unworkable; during parts of this paper, it could be read to suggest not just an explanation for every output is needed, but different kinds of explanations for every output depending on the intended audience. The processing power needed to achieve either of these scenarios would likely be significant, if not insurmountable. The paper should discuss in greater detail when explanations for outputs are necessary. For example, are explanations necessary when there is a low level of risk from the outputs or tasks being performed? It should also be noted that significant work still needs to be done on the technology side to be able to achieve this. As focus moves from

[5] NISTIR 8312, line 134.
[6] NISTIR 8312, line 151.
[7] NISTIR 8312, lines 184-185.
[8] NISTIR 8312, lines 198-199.
[9] NISTIR 8312, lines 218-227.

principles to frameworks, flexibility will be essential as the capabilities of today's technology are unlikely to be able to meet the ideal aim.

3. Types of Explanations

This is an important section of the paper given the different audiences for an explanation discussed in the four principles. Acknowledging the list is not exhaustive, some types of explanations may be more easily delivered and well-received than others. For instance, "regulatory and compliance" and "system development" explanations could be easier to achieve than "user benefit" or "societal acceptance" as the first two are likely looking for very specific information where as the audience for the latter two is much broader and may be looking for different explanations even within that "type".

The example in "owner benefit" may also not be reflective of how the user ultimately makes a choice to watch a movie; the user may be satisfied with the explanation of *why* a recommendation was delivered, but not satisfied with the output or recommendation.[10] Conversely, a user may not care why a service is recommending specific movies, shows or music so long as the user is satisfied with the output.

The last line of this section is incredibly important however, and points to the statement made earlier that the ability to deliver explanations is very much dependent on the capabilities of the system and the amount of computing power willing to be dedicated to explaining an output: *"The five categories of explanations illustrate the range and types of explanations and points to the need for flexibility in addressing the scope of systems that require explanations."*[11]

Conclusion

Arm appreciates the opportunity to comment on this important product. Arm shares NIST's interest in ensuring there are norms, standards and frameworks that establish trust in AI, and we look forward to continuing to engage as the Explainable AI work moves forward, as well as other work related to trustworthy AI.

Respectfully Submitted,

Vince Jesaitis
Director, Government Affairs
Arm, Inc.
vince.jesaitis@arm.com

[10] NISTIR 8312, lines 271-276.
[11] NISTIR 8312, lines 310-312.