NIST Explainable AI Workshop Summary

Held virtually on January 26-28, 2021

1. Introduction

This report represents a summary of the National Institute of Standards and Technology (NIST) Explainable Artificial Intelligence (AI) Workshop, which NIST held virtually on January 26-28, 2021.

1.1. Disclaimer

This publication is intended to capture external perspectives related to NIST standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and other organizations. This report was prepared as an account of a workshop; it is intended to document external perspectives; and does not represent official NIST positions.

1.2. Background

There is significant interest in explainable AI from the AI community and its stakeholders. As part of NIST's efforts to provide foundational tools, guidance, and best practices for AI-related research, NIST released a draft whitepaper, "Four Principles of Explainable Artificial Intelligence," (draft NIST Internal Report (NISTIR) 8312) for public comment. Inspired by comments received, this workshop further explored the understanding of explainable AI.

2. NIST's Role

As part of an ongoing trustworthy AI initiative, NIST is seeking input from multiple stakeholders on trustworthy AI and its components. Towards achieving this goal, NIST has been holding and sponsoring a series of workshops, releasing a series of white papers (with some selected for public comment), and participating in standards development efforts.

3. Workshop Format

This section outlines the format of the workshop. Sections 4 and 5, we provide an overview of the workshop content and takeaways.

This workshop was held virtually over three days to support a safe exchange of information and provide the opportunity to collaborate for participants from many time zones. The workshop was conducted with the following components: introductory remarks, panel sessions, breakout sessions, and concluding remarks.

There were 920 registrations and 563 attendees across the three days. Attendance varied across sessions, as participants signed up for individual breakout sessions. The breakout sessions allowed for different participants to interact with each other and NIST.

3.1. Introductory Remarks

The workshop began with a welcome statement from the NIST Information Technology Laboratory (ITL) Director, Charles Romine. Following, the chief of staff for ITL, Elham Tabassi, introduced the trustworthiness in AI program at NIST. To cap off the introductory remarks, Project Lead, P. Jonathon Philips provided an overview of the draft whitepaper, "Four Principles of Explainable Artificial Intelligence," (draft NISTIR 8312). These opening remarks served to familiarize participants with the purpose and context of the workshop.

3.2. Panels

The workshop consisted of three panels. Each panel had a moderator and expert panelists discuss topics related to explainability in AI. Panelists were given 10 minutes each to provide an opening statement followed by a period of prepared questions from the moderator. Questions from the audience were taken after the prepared questions. Each panel lasted for 90 minutes.

3.2.1. Panel: What are the roles of explanations throughout the AI life cycle?

Moderator:

- Amy N. Yates – Mathematician, NIST

Panelists:

- Umeshwar Dayal – Corporate Chief Scientist, Senior Fellow: Information Research, Senior Vice President, Hitachi America, Ltd
- Michael Hind – Distinguished Research Staff Member, IBM
- David Leslie – Ethics Team Lead and Ethics Fellow, the Alan Turing Institute

3.2.2. Panel: How does explainable AI fit into the trustworthy AI ecosystem?

Moderator:

- Carina Hahn – Social Scientist, NIST

Panelists:

- Will Carter – Global Policy Lead for Responsible AI, Google
- Sumeet Chabria – Head of Global Business Services, Bank of America
- Frank Torres – Director of Public Policy, Office of Responsible AI, Microsoft

3.2.3. Panel: What does a risk management strategy look like for explainable AI?

Moderator:

- Peter Fontana – Computer Scientist, NIST

Panelists:

- Patrick Hall – Principal Scientist, BNH.AI and George Washington University
- Nahla Ivy – Enterprise Risk Management Officer, NIST
- Courtney Lang – Director of Policy, Information Technology Industry Council (ITI)

3.3. Breakout Sessions

The workshop featured a total 42 facilitated breakout rooms held during 9 time slots, each discussing one of the three panel topics: "What are the roles of explanations throughout the AI

life cycle?", "How does explainable AI fit into the trustworthy AI ecosystem?", and "What does a risk management strategy look like for explainable AI?" Participants signed up in advance for these sessions. The sessions were an hour long to promote interaction. We made multiple breakout sessions available for each topic to provide variety of options for participants in different time zones. Themes and highlights from these sessions are in Section 5.

3.4. Report out

Following all the breakout sessions, a debrief was conducted. The five facilitators who helped guide the discussion presented common themes from the previous days' sessions. The session lasted 50 minutes. A summary of these themes and highlights can be found in Section 5.

4. What we heard: Panel Sessions

NIST operated in the convener role for bringing together perspectives on explainable AI in this workshop. In that capacity, the sections below represent a summary of what was discussed at the workshop. These summaries are not meant to be exhaustive or express NIST's point of view. Rather, our goal is to inform the themes and highlights expressed in each session.

4.1. Panel Session Summary: What are the roles of explanations throughout the AI life cycle?

This panel session focused on whether there was a need for explanations at all stages of the AI life cycle, what those explanations looked like, and the evaluation of those explanations. The panelists and moderator discussed the need for explanations at all life cycle stages. Panelists confirmed a need for explanations at all life cycle stages.

The panelists discussed the intricacies of explanations within the life cycle. This included the need for a life cycle description that adequately describes the AI process. The context surrounding explanations is a challenge. Panelists highlighted the need to understand who is explaining what to whom. Additionally, given that different roles require different explanations, the flow of the explanation is not in a single direction.

Therefore, each context may be different and require different explanations. To meet this challenge, a framework for explanations was suggested. Additionally, panelists suggested that the impact of the AI system was critical to consider when providing explanations.

4.2. Panel Session Summary: How does explainable AI fit into the trustworthy AI ecosystem?

This panel compared and contrasted the elements of trustworthy AI systems and how explainability was integrated. Panelists noted that building trustworthy AI requires diversity, clarity, and accountability at all parts of the life cycle. AI is not completely new in terms of creating trustworthy and responsible technology systems; however, it will require some innovation to meet the complexity inherent in AI systems.

Building trustworthy AI means being aware of unintended consequences of AI, including exposing potential bias. Panelists noted explainability makes it easier to identify issues for developers, help users trust AI, and expose how decisions are made at all stages. It can inform a determination of when a human needs to be the key decision maker.

Panelists suggested transparency involves providing the right kind of explanation according to the audience, and there are many different audiences – developers, consumers, policymakers, etc. Explanations could spell out the capabilities and limitations of a system, so it is used for its intended purpose. Trustworthy AI requires thinking through the entire process holistically, from acquiring data, to measuring performance and knowing how a system works, and its limitations. Panelists suggested tools to help create trustworthy AI and mentioned efforts including industry groups from across sectors.

4.3. Panel Session Summary: What does a risk management strategy look like for explainable AI?

This panel focused on the intersection of risk management and explainability. The key questions surrounded the concepts of need, frameworks, and best practices in risk management.

Panelists discussed the concept of context in risk management. The panelists suggested that all organizations have a unique risk context and stakeholders which affect risk disposition. Therefore, the tailoring of explanations to the specific risk scenario is critical.

The panelists also discussed the intersection of risk management and the AI system's life cycle. Panelists suggested that applying risk management techniques earlier in the life cycle produces better outcomes. Additionally, the panelists discussed the ability to re-use other risk management frameworks. Some suggested the tailoring of previous frameworks, while others suggested creating more specific risk management tactics.

5. What we heard: Breakout Sessions

5.1. Breakout Session Summary: Explanations in AI Life Cycle

While there was significant discussion over what constituted the beginning, middle, and end of an AI system life cycle, many participants agreed there was need for explanations in some form throughout the process. The context of each organization and system is critical to understanding the impact explanations have throughout the life cycle.

The focus on internal vs. external communication of explanations was a key theme suggested by participants. Since explanations serve different purposes throughout the life cycle, participants commented on the importance of audience of the explanation. For example, some explanations are designed for internal communication between managers and developers, whereas other explanations are designed for end users.

Along with the context, participants suggested the validation and verification of explanations would depend on factors such as the stage in the life cycle and audience of an explanation. Participants commented on the importance of designing explanations with end users of the explanation in mind. Through this goal-oriented process, explanations could have traceability through the life cycle. Otherwise, participants warned of "baseline drift" of validation and verification measures.

Additionally, some participants commented on the intersection of bias and explainability within AI systems when creating validation and verification processes. Participants suggested involving many stakeholders at all stages within a life cycle and especially during the validation and verification process.

5.2. Breakout Session Summary: Trustworthy Ecosystem and Explainable AI

Participants suggested trustworthiness is broader than explainability and that trustworthiness and transparency includes other facets in addition to explainability. Participants confirmed explanations are essential for trustworthiness, especially for human consumers. Explanations enable verification and request for more information. Participants gave examples in autonomous driving, credit scores, and privacy where transparency is a top goal.

Participants suggested explanations and trustworthiness are inextricably linked. The ability to measure against well-accepted standards can affect trustworthiness of explainable AI. Therefore, participants said that creating these standards can increase trustworthiness of explainable AI.

Participants suggested a simultaneous bottom-up and top-down approach, with the following considerations. A top-down approach indicates you know the whole system, which may or may not be possible. Bottom-up approaches precludes a global vision for the system. Top-down approaches may be too theoretical in most cases. Bottom-up approaches may help to understand the components of the system and lead to higher level interpretability. A top-down approach may be more suitable for accountability.

5.3. Breakout Session Summary: Risk Management and Explainable AI

In these breakout sessions, participants discussed both the concept of risk management of AI systems as well as the role of explanations within risk management. Participants said risks related to explainable AI are throughout the lifecycle: at the decision to start a system, data acquisition, data modeling or training, through deployment. Participants suggested explainable AI implies risks to privacy and security.

Participants highlighted that there are risks inherent in using explanations. There is risk in explaining the right amount and the right way to the right audience. Depending on the audience, there is a risk of criticism from different ways of thinking, especially with different cultures. Special concerns for internationalization in explanations are important to mitigate

risks. Explanations can also mitigate risk by providing traceability augmenting solutions for risk management

Existing risk management frameworks may apply or be adapted to AI risk management scenarios. Participants stressed the importance of harmonization with respect to terms of reference and a common taxonomy in explanations.

6. What's Next

NIST anticipates using the feedback provided in the request for comments for the first draft of "Four Principles of Explainable Artificial Intelligence," (draft NISTIR 8312) and the feedback received in the workshop to revise the draft. Keep up to date with progress on the Explainable AI web page.

NIST continues to welcome feedback and questions throughout this process. Please contact explainable-AI@nist.gov with any questions.

Appendix: Workshop Agenda

This workshop consisted of plenary speakers, panels, and breakout sessions. Plenary speakers and panels provided expert insight and a starting point to the discussions. Breakout sessions on Wednesday and Thursday enabled further facilitated discussions among attendees on the panel topics. All times are Eastern Standard Time. The workshop agenda can also be found on the event webpage with additional information regarding the workshop.

Day 1: Tuesday January 26, 2021

Time Start	Time End	Topic
11:00 AM	11:30 AM	Welcome to NIST: Charles Romine – ITL Director, NIST Opening Remarks: Elham Tabassi – ITL Chief of Staff, NIST
11:30 AM	12:00 PM	Overview of *Four Principles of Explainable Artificial Intelligence, Draft NISTIR* 8312 – P. Jonathon Phillips – Electronic Engineer, NIST
12:00 PM	1:30 PM	Panel: What are the roles of explanations throughout the AI life cycle? Moderator: Amy N. Yates – Mathematician, NIST Panelists: Umeshwar Dayal – Corporate Chief Scientist, Senior Fellow: Information Research, Senior Vice President, Hitachi America, Ltd Michael Hind – Distinguished Research Staff Member, IBM

		David Leslie – Ethics Team Lead and Ethics Fellow, the Alan Turing Institute
1:30 PM	2:30 PM	Lunch Break
2:30 PM	4:00 PM	Panel: How does explainable AI fit into the trustworthy AI ecosystem? Moderator: Carina Hahn – Social Scientist, NIST Panelists: Will Carter – Global Policy Lead for Responsible AI, Google Sumeet Chabria – Head of Global Business Services, Bank of America Frank Torres – Director of Public Policy, Office of Responsible AI, Microsoft
4:00 PM	4:15 PM	Break
4:15 PM	5:45 PM	Panel: What does a risk management strategy look like for explainable AI? Moderator: Peter Fontana – Computer Scientist, NIST Panelists: Patrick Hall – Principal Scientist, BNH.AI and George Washington University Nahla Ivy – Enterprise Risk Management Officer, NIST Courtney Lang – Director of Policy, Information Technology Industry Council (ITI)
5:45 PM	6:00 PM	Closing Remarks – Carina Hahn, NIST

Day 2 Breakout Sessions: Wednesday, January 27, 2021

Time Start	Time End	Topic
9:00 AM	10:00 AM	Trustworthy Ecosystem and Explainable AI
10:00 AM	11:00 AM	Break
11:00 AM	12:00 PM	Risk Management and Explainable AI
12:00 PM	1:00 PM	Lunch Break
1:00 PM	2:00 PM	Explanations in AI Life Cycle
2:00 PM	3:00 PM	Break
3:00 PM	4:00 PM	Trustworthy Ecosystem and Explainable AI
4:00 PM	5:00 PM	Break
5:00 PM	6:00 PM	Risk Management and Explainable AI

Day 3 Breakout Sessions and Closing Remarks: Thursday, January 28, 2021

Time Start	Time End	Topic
9:00 AM	10:00 AM	Explanations in AI Life Cycle
10:00 AM	11:00 AM	Break
11:00 AM	12:00 PM	Trustworthy Ecosystem and Explainable AI
12:00 PM	1:00 PM	Lunch Break
1:00 PM	2:00 PM	Risk Management and Explainable AI
2:00 PM	3:00 PM	Break
3:00 PM	4:00 PM	Explanations in AI Life Cycle
4:00 PM	5:00 PM	Break
5:00 PM	6:00 PM	Breakout Debrief Closing Remarks – Amy N. Yates, NIST