What Do People Really Want When They Say They Want "Explainable AI"? We Asked 60 Stakeholders.

Andrea Brennen
In-Q-Tel Labs
Waltham, MA 02451, USA
abrennen@iqt.org

CHI'20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA
© 2020 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.3383047

Abstract

This paper summarizes findings from a qualitative research effort aimed at understanding how various stakeholders characterize the problem of Explainable Artificial Intelligence (Explainable AI or XAI). During a nine-month period, the author conducted 40 interviews and 2 focus groups. An analysis of data gathered led to two significant initial findings: (1) current discourse on Explainable AI is hindered by a lack of consistent terminology; and (2) there are multiple distinct use cases for Explainable AI, including: debugging models, understanding bias, and building trust. These uses cases assume different user personas, will likely require different explanation strategies, and are not evenly addressed by current XAI tools. This stakeholder research supports a broad characterization of the problem of Explainable AI and can provide important context to inform the design of future capabilities.

Author Keywords

Explainable AI; User Research; Machine Learning; UI/UX design; Data Science; Interface design.

CSS Concepts

• **Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI**

Introduction & Approach

In 2019, In-Q-Tel Labs launched an effort to understand emerging tools and technologies that support Explainable AI (XAI). During that exploration, we familiarized ourselves with frequently-referenced academic publications on techniques such as LIME [1], SHAP [2] and TCAV [3] and investigated the research and evaluation efforts associated with DARPA's XAI program [4]. We learned that several software start-ups are building Explainable AI products (for example, Fiddler Labs [5], Untangle [6]) while a number of established technology companies are adding Explainability features into existing platforms (see Microsoft's Interpretability packages for Azure [7] or Oracle's Skater project [8]). We also saw that for some stakeholders, Explainable AI is viewed as a challenge best addressed not through tools and technologies, but via policies or standards. The Global Data Protection Regulation [9, 10] and the OECD Principles on AI [11] both incorporate explicit references to explainability and transparency. For example, Section 1.3 of the OECD Principles on AI is titled "Transparency and explainability" and includes the stipulation that "AI Actors should commit to transparency … to enable those affected by an AI system to understand the outcome … based on plain and easy-to-understand information" [12]. Neither policy document, however, includes specifications or guidance as to how this transparency ought to be provided or what constitutes a sufficient explanation [13]. When we shared the findings from this technology study with others in government and industry, we saw that today's XAI tools do not fully capture the types of explanations that many people want. In response, we undertook a qualitative research effort to interview a diverse group of stakeholders about this topic. Our objectives were to develop a better understanding of how different stakeholders across government and industry characterize the problem of Explainable AI and to identify needs not met by current tools. This short paper summarizes two initial findings: (1) there is a lack of consistent terminology to discuss Explainable AI; and (2) Explainable AI incorporates multiple distinct use cases which are important to different people for different reasons. We view these findings as preliminary user research, that can provide context to inform the design of future XAI tools and capabilities. From March to November 2019, I conducted semi-structured interviews with 40 stakeholders whose work involves Explainable AI. These stakeholders included technologists, start-up company founders, academics, venture capital investors, people in oversight or policy roles in organizations currently using machine learning (ML) or artificial intelligence (AI) technologies and individuals who see themselves as current or future end-users of ML and AI systems. I also collected input from an additional 24 people through two focus groups, a panel discussion and several informal meetings. Both focus groups were exclusively with end users of ML and AI systems. When possible, the interviews were recorded and transcribed. However, in some cases, discussion of sensitive information precluded the use of recording equipment.

Lack of Consistent Terminology

Researchers, technologists, lawyers, policymakers, domain experts and business leaders all use the phrase Explainable AI, but there is no consensus on what specific capabilities this term refers to, or what objectives it is meant to satisfy. During the interviews I conducted, all of the terms listed in the sidebar were used by at least one interviewee as a synonym for

Explainable. The use of such a wide range of synonyms indicates a breadth of interpretations of what Explainable AI means, as well as a lack of agreed-upon terminology related to this emerging technology. For each interviewee who used two of the above terms interchangeably, there was another who was adamant about the need to distinguish between them. For example, one employee who vets AI companies at a strategic investment firm responded to the question "How would you summarize what Explainable AI means?" by saying, "Interpretable AI is about enabling analysts to use modern probabilistic tools to find conclusions in massive amounts of data, and to understand the mechanisms that help form those conclusions."[1] The interviewee literally substituted the word *Interpretable* for *Explainable* in his response. A data scientist who works at the same firm, however, provided a nuanced distinction between *Explainability* and *Interpretability,* suggesting: "With explainability, you have a black box and you try to explain what [a model] did; with interpretability you are actually doing something within the black box that lets you understand how it functions. Interpretability requires some sort of manipulation of the actual model to test if your explanation is valid or not." In addition to varying interpretations of the word *Explainable*, interviewees also referenced surprisingly divergent definitions of *AI*. Several interviewees viewed the problem of Explainable AI as something specific to neural networks, a class of notoriously complex models used in deep learning. And for some, the impenetrability of neural networks was reason to dismiss the entire project of Explainable AI as a "red herring" or misleading fallacy. Multiple people explained that there is a limit to how much or what sort of transparency is achievable for phenomena that exceed a certain level of complexity and that neural networks exceed this threshold. A second group saw Explainable AI (or *Interpretable ML*, to use the term many of them preferred) as applicable to machine learning more broadly. For this group, AI was inclusive of, but not limited to deep learning's neural networks. A third group of interviewees defined AI much more broadly, incorporating nearly any automated manipulation of data, including traditional statistical modeling techniques and in some cases, even simple computational functions available in Microsoft Excel. Of note, only two stakeholders mentioned Artificial General Intelligence (AGI) and neither did so in a way that referenced specific implications for Explainability. At least among these stakeholders, discussion of XAI was almost exclusively focused on specific applications of Narrow AI — for example, SPAM filters, the use of models to predict debt default or recidivism rates, or the operation of autonomous vehicles.

Terms used by stakeholders as synonyms for *Explainable:*

Accountable,
Auditable,
Certifiable,
Fair,
Inspectable,
Interpretable,
Justifiable,
Operational,
Ready-to-Use,
Reliable,
Repeatable,
Reproducible,
Responsible,
Self-service,
Tested,
Transparent,
Trusted,
Unbiased,
Understandable,
Verifiable.

The lack of consistent terminology about Explainable AI hinders discussion and makes it easy for people to talk past one another, sometimes without even realizing they are doing so. Miscommunication is exacerbated by the frequent use of words that have both a technical definition (or multiple technical definitions) and a colloquial usage (that may mean something else entirely). For example, while the academic research community is currently debating at least 20 technical (and in some cases mutually exclusive) definitions of *fairness* [14], a lay person affected by an automated decision likely has his or her own intuitive notion of what outcomes seem *fair* or *just*. It is easy to imagine an affected individual dismissing an outcome that

[1] Interviewee's names have been omitted to protect their privacy and interviewees have approved all demographic information.

deviates from his or her notion of justice as *unfair*, even if a data scientist can demonstrate that the underlying model satisfies a technical definition of fairness. Despite using the same term, these are very different concepts of fairness. While the data scientist is talking about fairness in how the model performs, the affected person is more concerned with fairness in how the model is used in some social application.

Multiple motivations for Explainable AI

Despite the lack of consistent terminology, one image that interviewees invoked again and again was that of AI as a *black box.* There was widespread agreement about a lack of transparency in current AI/ML models and systems. For those who wanted Explainable AI, their desire was generally motivated by a mismatch between what they wanted to understand about AI and what they currently understood. There was considerable variation, however, in how people described what they wanted to understand, why they wanted to understand it, and what they deemed *possible* (for anyone) to understand*.* If AI is a black box, everyone may want to look inside, but they want to do so for very different reasons. The variations are illustrated in the following high level use cases: debugging models, detecting bias, and building trust.

Debugging Models. Several interviewees emphasized the need for greater transparency into the mechanics of complex machine learning models. As with any software product, some transparency is important for debugging code and improving performance. Many interviewees whose work involved producing or deploying models said they wanted better tools to determine if their code was executing properly, to confirm that code was doing what they thought it was doing, and to help them build intuition about how models worked. As one data scientist put it: "as a machine learning practitioner, one of my goals is improving model performance. So, I would basically use any [interpretability] tool to make sure that my model is working correctly.” Others discussed how Explainable AI could help data scientists address a problem known as “model drift,” which one interviewee described in this way: “software that has a probabilistic component is different from traditional software in that you can’t just ‘fire and forget;’ ongoing maintenance is needed because the data that flows through your system might change.” The implication was that XAI capabilities should not be static, but rather, should enable continuous monitoring and performance testing. The Chief Technology Officer of one company emphasized that the current lack of transparency into, and intuition about, neural networks is a key difference between today’s deep learning systems and previous types of engineered systems: “In engineering, we have worked really hard to make mission critical systems fail in predictable ways, but for AI, that arc of predictability is missing…debugging a large, complex computer program…starts with an expectation of what will happen…for deep learning…the math is completely understood, but we still don’t know what to expect.” Many of the Explainable AI tools available today were designed to support aspects of this use case, which is sometimes broadly referred to as ML Ops (or Machine Learning Operations). These tools employ a variety of different explanation strategies. For example, tools like LIME [1], use a simpler model to approximate the behavior of a complex model [15]. Other tools, such as TCAV [3] and SHAP [2], help users build intuition about how models work by allowing them to test and explore how different inputs relate to different outputs.

Regardless of the specific technique, most existing tools are intended for use by data scientists, engineers, or AI researchers and assume substantial expertise in ML. As one interviewee noted: "[Most of] the people making these tools are machine learning people, so they're kind of tailoring the tools for themselves."

Identifying Bias. Other interviewees emphasized the need for more transparency into biases that might exist within training data, models, or deployed systems, i.e. systematic errors or inconsistencies that could lead to unsupported decisions [16]. Several stakeholders stressed the need for Explainable AI tools that would provide additional context about machine-generated results and tell them *why* an ML/AI system produced a particular output or prediction. They felt additional context would help them judge the reliability of model output, understand the stability of predictions, and determine an appropriate level of confidence for conclusions derived from machine-generated output. Some individuals worried that without sufficient context, machine-generated insights could exacerbate confirmation bias, as results might only be accepted by those who already agreed with the conclusion. Several people spoke about Explainable AI as an important means of auditing models, for example, to identify biases in training data that might lead to unjust outcomes or unlawful discrimination. A philosopher of science summarized his perspective this way: "we build the algorithms, so we can build our biases into them…we're constructing…a system that we don't totally understand and we're going to need to explain how that system works." In comments underscoring both the difficulty and the importance of this undertaking, multiple interviewees described how including seemingly benign data types in training data, such as individuals' zip codes, could lead to unjust outcomes. For example, an individual could remove data about race, religion or age from training data, but still use other factors as a proxy for that data. One interviewee noted, "in this way, they could still create a biased algorithm that would hurt certain classes of people. A model that meets the letter of the law can produce outcomes against the spirit of the law." Whether stakeholders emphasized the potential for confirmation bias or unlawful discrimination, many expressed concern that undetected biases would lead to unintended consequences, and a sense of urgency that today's tools do not provide sufficient transparency — at least, not in the way that they seemed to want it.

Building Trust. Several interviewees spoke about a third motivation for Explainable AI: building trust in unfamiliar technologies [17]. For them, many people are not ready to trust AI systems, in part because they don't understand how these systems work. The CEO of one tech start-up used the analogy of electronic plane tickets: "If I go online now and I buy a plane ticket and they give me an e-ticket, I'm fairly certain that if I show up at the gate I can get on the plane. But years ago, back when people were first doing this, without a [paper] ticket in hand, a lot of people were anxious. If they didn't talk to a person, to confirm they had a seat on the plane, they were anxious…we're at that stage with A.I. today. People aren't yet ready to trust that they understand how it works. And, of course these systems are too complicated to explain how they work." This example is interesting because it suggests that users will build trust in AI technologies through familiarity, exposure, and personal experience, as opposed to detailed explanations of how the technology works. This theme was echoed by other technologists,

Acknowledgments*:*
The author wishes to thank Marissa Gerchick for her thoughtful research on XAI tools and technologies; Bonnie Brennen for the invaluable input on qualitative research methods and her help framing this study; Glenn Gaffney and IQT for supporting this work; Remco Chang for his encouragement to write this paper; and all of the people who graciously offered their time and insight during interviews.

several of whom suggested that addressing the challenges of Explainable AI might be more about helping users see value in new tools and feel comfortable using them, as opposed to explaining how technologies work. To do this, one interviewee explained, “you need to talk to people in a language that they understand.”

Analysis

Future XAI tools will need to employ a variety of explanation strategies to address these different use cases, as they require explaining different things to different audiences. However, much of the current work on XAI is focused on the first use case — providing transparency into model mechanics via explanations that are accessible to people with substantial expertise in ML. Stakeholders outside of academia and research labs want explanations that will help them use model output more effectively and more responsibly. Many of these stakeholders lack the technical expertise required to use current XAI tools and would benefit from capabilities designed with them in mind — users who are experts not in ML techniques, but in the *content* of their data. Some organizations and researchers are working towards more accessible interfaces for Explainable AI. Researchers affiliated with Google’s PAIR initiative [18] have created several prototypes that use interactive data visualizations and Outlier.ai [19] has developed a capability to auto-generate narrative “data stories” to explain insights to business analysts via short declarative sentences. [2] However, there is considerable opportunity for UI/UX designers to contribute through new design patterns that make aspects of AI more accessible to diverse audiences. In some ways, today’s debates about Explainable AI can be seen as debates about who AI should be explainable to. Without a greater focus on accessibility, many communities could inadvertently be excluded from debates about the future use, oversight, and regulation of AI/ML systems. Despite the range of topics that stakeholders discussed during this study, one topic was conspicuously absent — the need for explanations of the errors machines can make. For Explainable AI to provide insight into bias, foster trust, or help those affected by an AI system understand outcomes, end users, policy makers, and the general public need more transparency into how ML/AI systems can fail and what is at stake when they do. Ideally, future XAI tools will include interfaces that clearly communicate the likelihood of false positives and false negatives (type 1 and type 2 errors), foreground the costs — or disutilities — associated with various outcomes, and help people reason about the tradeoffs and risks of different decisions. Explainable AI may offer the most value if it can provide a framework to help people reason about available information in a way that helps them make better decisions.

Conclusion

Interest in Explainable AI is growing quickly. Yet, the lack of consistent terminology continues to hinder discussion, particularly between people with different disciplinary backgrounds. Given the increased attention on Explainable AI, we wanted to share our initial findings with the broader CHI community. To make progress in this area we need clear definitions of key terms in order to help stakeholders communicate more effectively. We believe identifying the current *lack* of consistent terminology is an important first step.

[2] IQT Labs is a wholly owned subsidiary of IQT, which is an investor in Outlier.

References

[1] (Local Interpretable Model-Agnostic Explanation) Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* ACM.

[2] (SHapley Additive exPlanations) Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems.*

[3] (Testing with Concept Activation Vectors) Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2017). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*.

[4] Gunning, D. (2016). Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency.* Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence. And Gunning, D. (2017). Explainable Artificial Intelligence (XAI). Program update. *Defense Advanced Research Projects Agency.* Retrieved from https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

[5] https://www.fiddler.ai/

[6] https://www.untangle.ai/

[7] https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability

[8] https://github.com/oracle/Skater

[9] *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

[10] Vogl, Roland, Askhon Farhangi, Bryan Casey. "Rethinking Explainable Machines:The Next Chapter in the GDPR's 'Right to Explanation' Debate." University of Oxford Faculty of Law Blog. 15 May 2018.

[11] https://www.oecd.org/going-digital/ai/principles/

[12] See Section 1.3 on "Transparency and explainability" https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[13] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76-99.

[14] Verma, Sahil and Julia Rubin. 2018. Fairness Definitions Explained. In FairWare'18: IEEE/ACM International Workshop on Software Fairness, May 29, 2018, Gothenburg, Sweden. ACM, New York, NY, USA. https://doi.org/10.1145/3194770.3194776

[15] Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, *87*, 1085.

[16] The notion of bias as related to systematic errors in judgment is from: Amos Tversky and Daniel Kahneman's "Judgment Under Uncertainty: Heuristics and Biases." *Science*, vol. 185, 1974.

[17] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM.

[18] (People and AI Research) https://research.google/teams/brain/pair/

[19] https://outlier.ai/