Assessing Explainability in XAI

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Austin Huang

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Joshua Earle, Department of Engineering and Society

Introduction

In 2012, AlexNet became the first deep learning model to win the annual ImageNet Contest, which is a historic benchmark event testing a computer's capability to identify objects in images from more than 20,000 categories, ushering in a new age of Artificial Intelligence (AI) (Krizhevsky et al., 2017). In their official report, Alex Krizhevsky and his team claimed that the recent ability to access greater processing power allowed for their model's success (2017). This advent of technology spurred on an excitement for AI and acceleration of its development towards more complex and robust models, now capable of being used for more advanced purposes (Gunning, 2021). However, one repercussion of the growing capability of AI is the public's lessening ability to understand and trust AI (Xu et al., 2019). When AI now is able to assist in major financial decisions, aid in medical diagnosis, and other activities that change the trajectory of someone's life (Xu et al., 2019), is there a method to establish trust and keep AI accountable?

The field of explainable AI (XAI) offers one way to bridge the gap between an AI model and the user's understanding to increase trust and usability (Gunning, 2021). Defining XAI as "systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future," the Defense Advanced Research Projects Agency (DARPA) took an interest in XAI, recognizing its growing importance in the more complex, harder-to-understand field of AI today (Information Innovation, 2016). As a result, DARPA began its official XAI program in 2015, lasting until 2021, as a way to develop and further the understanding of XAI and provide users with a tool to build trust in its models (Gunning, 2019).

In this paper, I research the question of whether DARPA's XAI program effectively demonstrated XAI as a way to increase the trust of users and the usability of AI. In the following section, I begin by detailing the specific methodology and framework of analysis for my research. Then, in my background and overview section, I provide context for the history and development of XAI so far. This leads into an exploration of DARPA and its XAI program, and the program's underlying specific goals and motivations. Following that, I provide insight into the complexities of XAI explainability and how it might be defined and measured. In my analysis section, I utilize Actor-Network Theory to examine DARPA's research results to answer how the explainability of different models, presentations, and frameworks impacted user trust. I find that the DARPA XAI program was effective in finding at least one strategy to increase user trust by funding research that explored a wide variety of techniques of XAI, and even paved a way for the future development of XAI going forward to maximize user trust.

Methods

In my initial research gathering phase, I gathered broad overview articles that were effective in introducing XAI as a whole and explaining about specific parts of XAI–mainly the history of Local Interpretable Model-Agnostic Explanations (LIME), a recent prominent idea for a tool able to create explanations regardless of the AI model used to make a decision. I then gathered primary resource retrospectives, evaluation reports, and research reports directly from DARPA to gather firsthand information about the program to use for analysis later. These primary reports also led me to additional resources regarding similar concurrent research outside of DARPA, which I used as comparison studies. Because the DARPA project happened in 2016, I looked at sources that were all published within the last decade in my analysis, emphasizing the quickly evolving understanding and exploration of XAI.

The framework I used to analyze my gathered data was Actor Network Theory (ANT). In his textbook of STS Frameworks, Sergio Sismondo explains that ANT is a famous STS framework that views all entities as actors, human or non-human, and connects them to a network. The reason I used ANT is that it has generally been effective in analyzing new emerging technologies (Sismondo, 2009). To think in ANT terms will mean identifying main actors, establishing connections between them and examining the quality of those connections (Sismondo, 2009). In my report, I have three classes of actors: the XAI developers, the XAI models, and the users of these models. In my network, the connections are the user's desire to utilize XAI as an efficient tool to help with their tasks, the developer's desire to establish trust between XAI and the user, how the model can be influenced by the user, and how the user can influence the model. With this paper, I explore these connections, specifically between model and user to answer my overarching question of how explainability in XAI is perceived by users and how that will affect the future use and development of XAI.

Background and Overview

In this section, I provide context about the subject matter of XAI and important details pertaining to the main analytical content of the paper. In the following paragraphs, I include brief research and background on three topics: an overview of XAI, an overview of DARPA and their XAI program, and a brief exploration of the effort to quantify XAI explainability.

XAI Overview

Even though DARPA coined the term "explainable AI" in 2015, research had already begun regarding a type of AI which could explain itself for some decades. In their overview of XAI, Feiyu Xu and colleagues date the earliest work some 40 years ago. According to Xu and her team, "the rationale for XAI initially was that for any intelligent system, it should be able to

explain its decision" (2019, pg. 2). While the technology was likely not advanced enough 40 years ago to build complex models, researchers were still able to demonstrate XAI with models that were inherently structured to explain their decisions, such as with decision trees (Xu et al., 2019). A decision tree is a model with nodes representing questions and pathways connecting nodes between its levels depending on the answer to each question. Using backtracking, a person can trace the line of thinking a decision tree followed in making its final decision in a given problem. Other models that are explainable by structure include linear regression and logistic regression.

With the recent cultural AI revolution introducing deep learning as more complicated models, the demand for XAI grew, and the capability of XAI also grew with corresponding improvements to technology (Gunning, 2019). In his book, *Interpretable Machine Learning*, Chris Molnar explains that "machines now surpass humans in many tasks, such as playing chess…or predicting the weather" (2022, pg. 10). Molnar argues that if machines and humans were equally able to complete a task, machines would still be advantageous to humans because they can also scale, reproduce, and develop (2022). However, the major disadvantage that these machines lose are insights on data and the ability for human comprehension (Xu et al., 2019). Defined as Black-Box models, deep learning models may be accurate to perform tasks, but these complex models are held together by layers of nodes, "needing millions of numbers" to describe one such deep neural network (Molnar, 2022). Therefore, the need for XAI in most recent years stems from the human inability to interpret a model's decisions.

DARPA and their XAI Program

DARPA is a US Department of Defense research and development agency overseeing the development of technology that the military could use. The agency was first created by President

Eisenhower in 1958 as a response to the Soviet Union's Sputnik launch. Their mission statement is "to make pivotal investments in breakthrough technologies for national security" and have over the years partnered with academia, research, corporations, and other entities to support emerging technologies. *The Economist* has stated that DARPA is an agency "that shaped the modern world," having been involved in Moderna's COVID vaccine, GPS, personal computer, and other prominently used technologies (Economist, 2021).

Program manager, David Gunning, spearheaded the XAI program in 2016 until its midpoint in 2019 at which point, Matt Turek, a new manager, later completed in 2021. Gunning began the program due to the demand of the Department of Defense for more intelligent, autonomous, and symbiotic systems (Gunning, 2021). Envisioning a future likely partnering with AI tools, Gunning reasoned that working to understand and establish trust with these AI tools and its users was important. The three challenges Gunning sought to explore were how to produce more explainable models, how to design explanation interfaces, and how to understand psychological requirements for effective explanations (Gunning, 2019). To solve these challenges, the program selected 11 technical teams to focus on the first two challenges, and one team to address the third challenge. Most of these teams were from universities, such as Carnegie Mellon and Texas A&M, and some were from bigger corporations, such as Raytheon. These selected teams all proposed a specific area of research that DARPA thought would adequately provide detail relevant to the XAI program (Gunning, 2021). In this paper, I select the work of three technical teams to explore.

Background of Measuring XAI Explainability

Defining and quantifying explainability remains a challenging task before analyzing the effects of explainability on a system (Silva et al., 2022). Unlike accuracy, which can be

definitively measured with test data, explainability is subjectively defined. Different users value different information and will rate XAI explainability differently. Other questions emerge from the general phrase of explainability, such as when it is to be measured, and also if there is a way for XAI model explainability to be standardized (Silva et al., 2022). In Silva's and his team's research on explainability, they say that "explainability research lacks consistent definitions and evaluations making it difficult to draw sound conclusions about the efficacy of explainability techniques" and as a result "the enthusiastic pace of progress in XAI is outpacing the ability of the community to settle these debates with rigorous empirical or analytical study" (2022, pg. 2).

Authors Geoffrey Bowker and Susan Starr support Silva's claims in their book, *Sorting Things Out,* where they explain the problems with trying to quantifiably explain a complex ethical problem: "When a seemingly neutral data collection mechanism is substituted for ethical conflict about the contents of the forms, the moral debate is partially erased. One may get ever more precise knowledge, without having resolved deeper questions, and indeed, by burying those questions" (1999, pg. 24). According to Bowker and Star, throwing numbers and algorithms into a model to make it more robust may actually cause a model to overlook other aspects of a nuanced problem, and the task to evaluate the model's holistic effectiveness would become even more challenging. Furthermore, explanations, as a form of classification system to which Bowker and Star refer, are intended to act as neutral tools to judge biased decisions, but actually are socio-technical constructs themselves influenced by historical and cultural contexts. Therefore, an attempt to account for all these factors to quantify explainability continues to be a challenge with no clear right direction forward.

Quantifying explainability as a metric to evaluate the usefulness of a model's accuracy of truth is a dangerous task. In *Trust in Numbers* (Porter, 1995), Theodore Porter writes that

quantification has been a way to establish objectivity in our world and a way to universally communicate truth between people, "setting free the critical powers of man" (pg. 73). However, the danger is that "objectivity as impersonality is often conflated as objectivity with truth" (pg. 74). Porter argues that the effort to establish an objectivity over anything may appear elitist, "suppressing moral feeling in favor of rigor and impartiality" (ibid, pg. 77). Most threateningly, when quantification is applied to people, there is a power imbalance created, "turning people to objects to be manipulated…not exercised blatantly, it acts instead secretly, insidiously" (ibid, pg. 77). A society that accepts quantification of explanations as truth becomes a "pervasive bureaucracy of experts and calculators" and epitomizes why the work in quantifying explainability must carefully tread forward, considering the politics and the reality shifts of that before implementation (ibid, pg. 77). For any effort to judge the effectiveness of explainability, one must first consider the broader politics and social ramifications that extend beyond mere numbers and data to avoid the invisible danger of negligence.

More recently, some scientific research has pivoted to consider psychological and social consideration. For example, Tim Miller rejects trying to define explainability in terms of transparency or some other quantitative surveyable measure, insisting that explainability defined in that way will inevitably contain "certain cognitive biases and social expectations in the explanation process" (2019, pg. 1). He argues that explanations are social by nature, "presented as part of a conversation or interaction, relative to the explainer's beliefs about the explained beliefs" (ibid, pg. 3). In other words, Miller's research shifts explainability from one in which testers seek to evaluate the validity of a single all-encapsulating answer, but rather holistically evaluate a series of back and forth responses between user and machine, similar to how a

human-human interaction might need multiple interactions to clarify and hone in on a contextual response.

The DARPA XAI program also developed a way to evaluate explanations, via an explanation scoring system. The system was based on a scale with five categories: explanation satisfaction, explanation goodness, mental model understanding, user-machine task performance, and trust assessment. The way that this scale was intended to be used was that participants in user studies would rate these traits of different models. Trying to quantify explainability using a subjective judgment system is tricky, and Gunning claims that cross-disciplinary expertise, in particular with experimental psychology, was helpful (Gunning, 2021).

Research Review

I have selected three of DARPA's funded XAI studies to review: one that examines saliency maps, an existing XAI technology, one that investigates XAI as a user interface tool, and one that develops a framework for XAI. For the first two studies, I also found concurrent studies that researched similar respective topics and I included a brief review of those. This review will lead into the analysis section, where I apply Actor Network Theory to my findings from a review of these studies.

Salience Maps

One of the projects that DARPA funded was UC Berkeley's Salience Map research, done by Bhahan Vasu and coauthors. Salience Maps, otherwise known as heat maps, are a type of local method XAI tool, typically used in image-related tasks, such as classification, that highlight regions in an image that the AI thinks is most important relative to its task. According to the research team, the work they did was the first large-scale visual saliency map improvement, motivated to study how the task a user is performing can affect and consequently

be used to affect explainability. Because of the emphasis on task-driven explainability, this team highlighted the importance of XAI to be included in what's called Human-in-the-Loop, where users can provide feedback to a machine, which will affect its ongoing, adapting explanations (Vasu et al., 2021).

The team found that the usage of Saliency Maps benefitted the most in image identification tasks when the object to be identified was contained in a cluttered image, and when the object was small to a lesser extent. With objects contained in images with little clutter, the team found that Salience Maps had no benefit or, in some cases, hurt the performance of tasks. One other major finding from the research done was through a set of survey questions collected at the end of their research. While 60% of users agreed that XAI helps users give feedback, 83% of users agreed that XAI helps users understand, and 62% of users agreed that XAI improves ease-of-use, there was 60%, which is a majority of users, that indicated an ambiguity on their preference for salience maps, not strongly preferring to do the task with or without the salience maps (Vasu et al., 2021).

To give brief context regarding salience maps independent of DARPA, a team of researchers wrote a critical report that was published one year prior. In this report, Ahmed Alqaraawi and coauthors found that even though saliency maps significantly increased prediction of classifiers accurately, success rates were still low at 60.7%. Also, users self-reported confidence on average was low regardless of the presence of saliency maps. Salience maps were effective in their ability for a user to learn more about specific image features; however, this would have the risk of averting their attention to other important properties of the image. As a result of their findings, Alqaraawi and coauthors argue that deep learning models remained

unpredictable and would need something more besides local based XAI methods, suggesting a combination of global and local methods for future research (Alqaraawi et al., 2020).

Interpretable Intelligent Systems: XAI as interpretable tools

Besides image processing, DARPA also funded research at Texas A&M to look into fact-checking. In this particular research project, Rhema Linder and colleagues focused on different methods of presenting explanations of AI to assist users in determining the validity of facts. The particular ways of presentation are called explainable interfaces, which have already been shown to significantly increase user acceptance and reliance on recommendations made by machines. What this study did want to investigate was the tradeoff between quality of explanation and how the explanation was received by a user, specifically how the user's performance and the user's understanding was affected. The researchers hypothesized that a certain amount of explanation is necessary to enable understanding, but too much might be confusing and overwhelming for a user (Linder et al., 2021).

From their study, Linder and colleagues found that added explanation did not improve significantly to human task performance; however, they attributed this fact to news checking to be an inherently difficult task, boiling down to guessing if the user had no prior information. What the researchers did find was that those with access to the AI meter had significantly more similar responses with each other than to others, indicating that participants were utilizing the AI meter in their decisions. From surveys, a common strategy was participants would first attempt to use their own intellect or hunches, then would use the AI meter as a tiebreaker factor. Another finding was that users with more explanation took more time on average to complete tasks compared to those with less information (Linder et al., 2021).

In another research article outside of DARPA by Harmanpreet Kaur and colleagues also examining the effect of interpretability tools, these researchers found a divide between the intended use of a tool and the way that the user used the tool. These researchers defined explanations with regard to a social framework, bringing the idea that the main criteria for explanations should be simplicity, generality, and coherence, thinking of conversation as a conversation. Explanations, according to Kaur and colleagues, are inherently biased and when merely used to refer to "probabilities or statistical generalization, usually unhelpful" (2020, pg. 2). The team found that participants of their experiment relied too heavily on their interpretability tools and were more likely to accept it at face value rather than to use it to pursue a deeper understanding, and for example point out flaws with the explanation or data (Kaur et al., 2020).

Integrating Theory of Mind with XAI

For the last research I reviewed, I looked at a study from UCLA that introduced a novel framework, which redefines XAI explanation as viewed traditionally and instead integrates a Theory of Mind approach. Arjun Akula and colleagues formed a research team from UCLA, and were motivated by their criticism of current explanation methodology, that solely generating explanations, regardless of type and utility is not sufficient for increasing understandability and predictability, and that explanation is an interactive communication process. With this framework, Theory of Mind does not aim to leave XAI's explanation as is, but instead utilize an ongoing dialogue encapsulating a human's curiosity, the human's understanding of the machine, and the machine's understanding of the user. Theory of Mind places emphasis on communication between the user and the machine from a perspective where both initially do not know what the other party knows and a gradual progression to fill the gap of knowledge (Akula et al., 2022).

Akula’s and colleagues' study also introduced a new technique in explanation optimization called fault lines, identifying which aspect of a situation would be the most important to highlight along these fault lines. Like Miller’s account of explainability, Akula and colleagues agreed that good explanations require contextuality. Fault lines addressed Akula’s and colleagues’ research showing that attention-based maps, otherwise known as heatmaps, were not human friendly and thus did not increase user trust. Rather than adopting a causal model by explaining why the model chose a certain decision, their counterfactual model explained a decision by comparing variations of similar input that could lead it to choose an alternative decision in order to show why a model did not choose other decisions. Akula’s team found this counterfactual approach to be more effective to garner user trust (Akula et al., 2022).

With this study, participants found that the theory of mind explanations were more useful, sufficient, understandable, detailed and more confident in answering the questions, and timing was not significantly different from other groups either. In addition, using the theory of mind approach, users' trust metric was consistently higher than all other methods, including LIME and other widely used local interpretable XAI methods, and grew with each session. This trust metric was measured by looking at how well users could predict what decision a model would choose. For example, when comparing the fault-line Theory of Mind framework and Saliency Maps, user trust and reliance was higher in the Theory of Mind framework than Saliency Maps and the control group without any XAI tools. One other finding was that for attribution techniques, Saliency Maps were not found to perform better than the control group, indicating that Saliency Maps may indicate the region of importance, but not do well in revealing the exact inference mechanism for a model (Akula et al., 2022).

Analysis

Through analysis of these particular DARPA case studies and the particular research articles pertaining to the respective topic, I was able to construct a network consisting of these three main actors: the user, the XAI developers, and XAI itself. In the following paragraphs, I examined each of the three groups and their connections to establish the obligatory passage point in the network.

The User: The user can be vulnerable to AI through abusive data manipulation and bias, so users should know why particular AI makes its decisions to detect any systematic bias (Xu et al., 2019). However, in the studies that I examined, users currently tend to overtrust AI and also undersell the value of XAI. For example, with interpretable tools, Kaur and colleagues observed that users tended to accept what was presented to them visually as a quick solution rather than using it to research deeper (Kaur et al., 2020). In addition, if the task is difficult and the user has no prior information regarding it, research shows that users were more susceptible to accept an answer shrouded behind statistics and probability (Linder et al., 2021). As Silva and colleagues. put it: "surprisingly, prior research on explainability and compliance has found that users were more likely to agree with a decision-making tool if the tool provided an explanation – even if the tool was incorrect" (Silva et al., 2022).

Furthermore, users do not prefer the use of XAI, specifically with the salience map tool study example even though it did improve results (Vasu et al. 2021). Other XAI tools were used only as a tiebreaker, last resort option (Alqaraawi, 2020). With such variability to the needs of a user and how much a user values XAI, it is no surprise that developers have been struggling to find a consistent satisfactory method to build XAI.

XAI developers: This group represents all subject matter experts who research, create, or are aware of XAI. This group creates XAI for users and are motivated with finding a way to

bring accountability to AI and make AI more accessible to common users as AI has become more complex (Gunning, 2021). The traditional approach of XAI has been to develop explanations using algorithms, data, and statistics, paralleling the main developments of AI (Xu et al., 2019). When AI became too complex to integrate explainability within the model, XAI developers looked to find a way to create a tool that explained AI externally. This classic approach aimed to simplify the black-box XAI to return to simple comprehendible heuristics. However, many developers observed the tension between simplifying heuristics and satisfying users, ultimately impacting user trust. Such developers mention further research needed with a more multidisciplinary approach on human computer interaction and psychology.

XAI: XAI is perhaps a non-obvious actor because it is not human or tangible. However, XAI has influence and has influenced the other two actor groups in an important way. Researchers and XAI creators have always struggled to define explainability (Silva et al., 2022). There is a growing concern that "explainability research lacks consistent definitions and evaluations making it difficult to draw sound conclusions about the efficacy of explainability techniques" (Silva et al., 2022). DARPA also recognized that there was a need to develop a psychological understanding of explanations, beyond algorithms and probabilistic thinking. (Gunning, 2021). Interestingly, the research article from DARPA that yielded the best results defined explainability is not related to interpretability or transparency of the model, but rather the iterative nature of explanations and the iterative ability to adapt to the user (Akula et al., 2022). Furthermore, this research considered social definitions of explanations, regarding them not as causal, but as contextual and counterfactual; a good explanation should not attempt to explain merely why or how a decision was made, but rather consider how it could have made other decisions and explain why it did not choose them. This result was from the UCLA study

proposing Theory of Mind in XAI, in which the user not only evaluates what the XAI knows, but the XAI model also evaluates what the user knows as well. When a counterfactual theory of mind model of XAI was adopted, the level of trust from its users increased the most (Akula et al., 2022).

Therefore, the most important connection to focus on for the continued development of XAI should not be XAI's relationship to the user, which would imply that we prioritize an XAI model being able to cater to varying needs of users. It also should not be the user's relationship with the XAI creator, which would imply that we prioritize making accountable an XAI creator's method of finding data, training XAI to best fit the needs of the user. The most important connection is the user's relationship to XAI, or how much an XAI can understand its users. We should place XAI as an active character, similar to a teacher, who would explain the catered to the needs of their students. With an ongoing dialogue between the students and the teacher, seeing what each other knows, a teacher would be better able to explain to a student as they learn what the student knows and does not know. A satisfactory answer based on a user's individual needs rather than a general response, will ultimately lead to an increase in trust and reliability in XAI's answers. When presented with a Salience Map tool that highlights the exact regions the model uses to compute, users are confused and do not understand the reasoning behind the selected region of focus (Akula et al., 2022). When presented with an interface tool, if the interface tool presents an answer and explanation, users are susceptible to accept what is presented rather than question it. If developers focus on that connection from XAI to the user, then the danger of exploitation and data abuse can follow. However, when an XAI was developed specifically to be contextual and iterative, explanations are clearer and easier to understand for a human (Akula et al., 2022). Therefore, developers will benefit user trust more if

they instead focus on the connection from the user to XAI, exploring how to get XAI to understand their user.

Conclusion

DARPA funded 11 teams and allowed research for a wide variety of approaches and studies. Within their studies, some focused on branching out on already existing XAI models, such as with saliency maps, which did not lead to significant user trust. Others focused on analyzing the method to which explanations were given to users, which also did not improve significant user trust. However, one team focused on creating an entirely new approach for XAI models, incorporating an originally psychological framework, Theory of Mind, and this was one study that did increase the level of trust of users. These researchers accomplished this by focusing on bridging the gap not only between a user to XAI but XAI to a user, looking at how to iteratively improve explanations by learning what a user knows. Because the DARPA XAI program covered a wide range, they were able to hit upon a new method of XAI that has promise to bring validity to XAI and allow for a modern wave of XAI to continue to establish trust with its users.

Modern XAI, in order to be effective in its purpose, should not lose its goal of establishing trust with its users. Incorporating a social lens into explanation, researchers are able to expand the definition of explanation beyond just satisfaction, transparency, and interpretability. This new scoping of explainability calls for XAI to be an active agent, capable and responsible for engaging with the user. Going forward, it is my suggestion that further XAI development research continues to explore this XAI-to-user connection, transforming the traditional view of explanations from those which are one-shot explanations based on

information best available into an explanatory process which gets better with more accurate answers each iteration.

References

Akula, A. R., Wang, K., Liu, C., Saba-Sadiya, S., Lu, H., Todorovic, S., Chai, J., & Zhu, S.-C. (2022). CX-Tom: Counterfactual explanations with theory-of-mind for Enhancing Human Trust in image recognition models. *iScience*, *25*(1), 103581. https://doi.org/10.1016/j.isci.2021.103581

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks. *Proceedings of the 25th International Conference on Intelligent User Interfaces*. https://doi.org/10.1145/3377325.3377519

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

Gunning, D. and Aha, D.W. (2019), DARPA's Explainable Artificial Intelligence Program. AI Magazine, 40: 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Gunning, D., Vorm, E., Wang, J.Y. and Turek, M. (2021), DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters, 2: e61. https://doi.org/10.1002/ail2.61

Information Innovation, & Gunning, D., Broad Agency Announcement Explainable Artificial Intelligence (XAI) (2016). Arlington, VA.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for

machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376219

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional Neural Networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Linder, R., Mohseni, S., Yang, F., Pentyala, S. K., Ragan, E. D., & Hu, X. B. (2021). How level of explanation detail affects human performance in Interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters*, *2*(4). https://doi.org/10.1002/ail2.49

Miller, T. (2019). Explanation in artificial intelligence: Insights from the Social Sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2022). *Interpretable machine learning: A guide for making Black Box models explainable*. Munich.

Porter, T. M. (1995). *Trust in numbers the pursuit of objectivity in science and public life*. Princeton University Press.

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2022). Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, *39*(7), 1390–1404. https://doi.org/10.1080/10447318.2022.2101698

Sismondo, S. (2009). Chapter 8 Actor-Network Theory. In *An introduction to science and*

technology studies (2nd ed.). Wiley.

The Economist Newspaper. (2021). *A growing number of governments hope to Clone America's DARPA*. The Economist. https://www.economist.com/science-and-technology/2021/06/03/a-growing-number-of-governments-hope-to-clone-americas-darpa

Vasu, B., Hu, B., Dong, B., Collins, R., & Hoogs, A. (2021). Explainable, interactive Content‐based image retrieval. *Applied AI Letters*, *2*(4). https://doi.org/10.1002/ail2.41

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. *Natural Language Processing and Chinese Computing*, 563–574. https://doi.org/10.1007/978-3-030-32236-6_51