Explaining Explainable Artificial Intelligence: An integrative model of objective and subjective influences on XAI

Gene M. Alarcon
Air Force Research Laboratory
Wright Patterson AFB, OH
gene.alarcon.1@us.af.mil

Sasha M. Willis
General Dynamics Information
Technology, Dayton, OH
sasha.willis@gdit.com

Abstract

Explainable artificial intelligence (XAI) is a new field within artificial intelligence (AI) and machine learning (ML). XAI offers transparency of AI and ML that can bridge the gap in information that has been absent from "black-box" models. Given its nascency, there are several taxonomies of XAI in the literature. The current paper incorporates the taxonomies in the literature into one unifying framework, which defines the types of explanations, types of transparency, and model methods that together inform the user's processes towards developing trust in AI and ML systems.

Keywords: Artificial intelligence, machine learning, XAI, trust, trustworthiness.

1. Introduction

The field of machine learning (ML) has grown exponentially in the last few decades as a subdiscipline of artificial intelligence (AI). The fast growth of both fields is due, in part, to increases in the availability of extremely large datasets, otherwise known as "big data" (Medeiros et al., 2021). ML has demonstrated its usefulness in domains including healthcare, autonomous driving, finance, and the criminal justice system (see Arrieta, 2022), to name a few. The implementation of AI/ML in various contexts has proven to be beneficial in many ways including increasing productivity (Sharma et al., 2020), improving accuracy and performance in certain types of tasks (e.g., facial recognition; Medeiros et al., 2021), and providing the capability of completing tasks that humans perform poorly or can't perform at all (e.g., text mining from big data, language translation, etc.; Suman et al., 2020). However, many of these systems have historically lacked transparency in their decision-making process, which has limited the user's acceptance and willingness to trust and use them (Miller, 2019).

A particular subset of the AI/ML fields is explainable artificial intelligence (XAI), which seeks to increase the understandability of models that are traditionally opaque (Zhou et al., 2021). The rapid development of XAI in the last decade has been driven by demand from the industries mentioned above. Ultimately, XAI seeks to provide more information about how the AI/ML system made its decision (e.g., prediction, classification) so that operators will use the algorithms appropriately. However, due to the nascency of XAI methods, there are several explanations and theoretical frameworks that attempt to demarcate the XAI field, but few share a common language or focus, and most importantly, there lacks a comprehensive architecture to unite them. This has led to many different delineations of factors that XAI encompasses, which fail to contribute to building theories that effectively advance the field. Importantly, much of the previous research has overlooked the psychological processes inherent in user perceptions of AI/ML models and systems (Miller, 2019). The current paper seeks to clarify the current state of the literature by combining the theoretical taxonomies into one cohesive framework and highlight how XAI relates to the psychological perceptions of system trustworthiness.

2. Artificial intelligence/machine learning

Artificial intelligence (AI) and machine learning (ML) algorithms have been utilized for handling many different types of data, including images, text, audio, and video. AI utilizes different learning algorithms (e.g., logistic regression, k-nearest neighbor, neural networks, etc.) to help develop the underlying model. ML grew out of the AI domain when the field began changing its goal from attaining AI to the practical use of applying algorithms to sort through large data sets. Although there is still debate over whether AI and ML are separate fields, as some advocate ML is a subfield of AI (Gareth et al., 2013), and others have advocated only intelligent subsets of ML (e.g., neural networks) are a subfield of AI (Alpaydin, 2010); we refer to AI and ML as a joint construct in the current paper as AI/ML, since the principles we are discussing are relevant to both.

Given the vast application of AI/ML to so many aspects of human life, there has been an increase in the desire to understand how AI/ML models make their

URI: https://hdl.handle.net/10125/102764
978-0-9981331-6-4

decisions. Increasing the transparency of the models can ensure impartiality, increase the robustness of the system by highlighting potential adversarial vulnerabilities, and can help to guarantee that only relevant variables influence the output of the model (Arrieta et al., 2020). For example, traditional AI/ML models have led to biased outcomes such as racial (Rudin, 2019) or gender discrimination (Dastin, 2018), which can be noncompliant with current legal requirements (Bibal et al., 2021). Additionally, AI/ML systems can seriously impact safety, particularly for those affected by these systems' decisions (Rudin, 2019), such as when the AI/ML is applied in high-risk contexts (e.g., autonomous vehicles). Developing a model of the factors and external variables that affect transparency of AI/ML systems will provide developers with the necessary principles to be incorporated into the AI/ML's design. The objective of developing such a model would be a resulting increase in the information that is pertinent to the user's understanding of the system's capabilities and limitations. Further, the implications of these decisions need to be considered, particularly in high-risk contexts.

One issue that has been problematic for utilizing AI/ML models is that the decision-making process of the most reliable models are opaque (i.e., "black-box" models). Researchers have delineated ML models into black-box models, in which the decision-making process is intrinsically obscure, and white-box models, in which a human can readily understand what variables were considered in the system's decision-making process. Many AI/ML systems rely on statistical models for their decision processes (e.g., logistic regression), which are inherently understandable, as the variables and their associated weights in the formula are known or available to the user. In contrast, with black-box models (e.g., neural networks), the underlying variables and processes used by the algorithm to make the decision are unknown to the user or too complex to interpret (Marcus, 2018; Medeiros et al., 2021).

Despite the limitations in their interpretability, researchers often opt to use black-box models over white-box models for a variety of reasons. For instance, neural network models can be used for processing extremely complex data (e.g., natural language processors or models with hundreds or thousands of predictors; Sheu, 2020). Proprietary models, as another example, can provide added security to the system at the expense of interpretability (Papernot et al., 2017). Importantly, the performance of black-box models are contended to be unmatched (Zhou et al., 2021).

3. Explainable artificial intelligence

The past decade has seen an advancement in AI/ML that has focused on understanding the decision-making processes of black-box models, particularly through the developing field known as explainable artificial intelligence (XAI; Zhou et al., 2021). As noted above, researchers have called for the development of more explainable models because of the importance of the decisions being made by the models without compromising the performance capabilities of these systems. High-risk scenarios, such as medical diagnosis, self-driving cars, and military operations, have particularly increased the demand for AI/ML models that can be easily interpreted (Adadi & Berrada, 2018).

The advent of XAI has increased the transparency of previously obfuscated models, and several theoretical frameworks have been developed for organizing the variety of XAI methods and creating taxonomies of such methods (e.g., Arrieta et al., 2020). Although these frameworks have several different monikers, we classify the primary dimensions into types of explanation methods, types of transparency, and model methods based on previous taxonomies. We discuss each in turn.

3.1 Model methods

Model methods are the application of the transparency principles described above to actual XAI. The XAI models in the current literature can be classified into one of three methods: model-agnostic, model-specific, and example-based methods (Molnar, 2019).

3.1.1 Model-agnostic methods. Model-agnostic methods of XAI are separate models or algorithms that can be applied to any existing black-box model. In other words, these methods are not built into the AI/ML algorithm but rather applied post-hoc (or after the model has been developed) to uncover meaningful information about how the outcome decision was reached (Arrieta et al., 2020). For example, if a researcher wanted to understand how a neural network weighs variables in a model, a Shapley Additive Explanation (SHAP; Lundberg & Lee, 2017) could be provided. SHAP removes one variable at a time to determine its impact on the model, both directly and through other variables. The top portion of Figure 1 illustrates a scenario where four variables are included in a model prediction; the bottom portion illustrates one iteration of the SHAP algorithm, in which one variable (Cholesterol) was removed from the ML model. The results help researchers to understand the variable in the context of the model through main effects, interactions, and

possible suppression effects by examining how changes in the model affect the outcome variable.

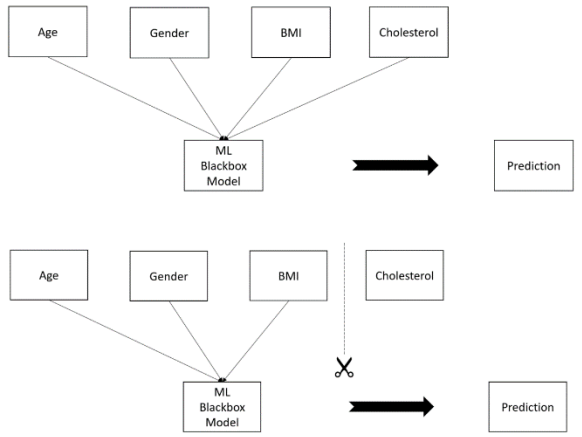

Figure 1. SHAP illustration. Top picture is original model, bottom picture is model without cholesterol.

3.1.2 Model-specific methods. Model-specific methods are XAI methods built into the AI/ML algorithms themselves to help understand the processes driving decisions and can only be applied to certain types of AI/ML models to aid in their interpretation (Molnar, 2019). That is, there is not another algorithm to apply when running the model (although there are a few exceptions), and one type of model-specific method cannot be applied to another type of model without some type of adaptation (Carvalho et al., 2019). For example, gradient saliency algorithms use highlighting in a neural network to illustrate what features of the image are compelling the decision processes. Figure 2 illustrates the results of a neural network that has been trained to classify cats and dogs. On the left-hand side of the figure, we see the image of the dog being classified. On the right-hand side of the figure, we see the pixels the algorithm used to classify the image. These methods are considered model-specific because they are built into the algorithm and cannot be applied to others (e.g., regression-based models).

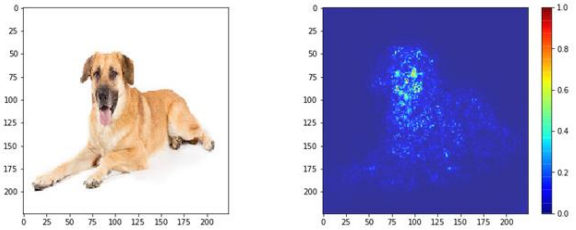

Figure 2. Gradient saliency example, data is on left and gradient saliency is on right.

3.1.3 Example-based methods. Lastly, example-based methods use individual dataset examples to illustrate how the model can be deceived (Szegedy et al., 2013). Example-based methods are different from the two previous methods in that they explain a model by selecting specific instances of the dataset, whereas the former two create summaries of the data, based on either data or rationale explanations (Molnar, 2019). Figure 3 illustrates a common adversarial attack for binning traffic signs. On the left-hand side of Figure 3 is a picture of a stop sign that is correctly classified. However, adding a few rectangular white and black boxes to the image, as illustrated on the right-hand side of Figure 3, can result in the model incorrectly classifying the image as 45-mile-per-hour sign.

Figure 3. Example of adversarial attack on stop sign.

3.2 Explanation Types

Zhou and colleagues (2021) describe six main AI/ML explanation types, each with a different objective, that exist for XAI methods, outlined as follows: rationale explanations, data explanations, responsibility explanations, impact explanations, fairness explanations, and safety and performance explanations (ICO & Turing, 2019, as cited by Zhou et al., 2021). *Rationale explanations* focus on increasing the understanding of why a model made a certain decision. In other words, they describe what the underlying processes were that led to the decision or output of the model. *Data explanations* focus on what data were used and how the data were used to influence the decision. This type of explanation helps users understand the influence different data types have on the decision or output. *Responsibility explanations* are explanations that focus on who developed, managed, and implemented the AI/ML model. That is, this type of explanation is concerned with who is accountable for the decision of the AI/ML system and what their reason was for designing the system. *Impact explanations* are based on the broader implications of the use of AI/ML decisions on society in general. This type of explanation aids the user in understanding the consequences of the AI/ML model's output and helps the user decide if they

want to use the system (e.g., to mitigate negative outcomes or increase positive outcomes). *Fairness explanations* are a subset of impact explanations, in that they are specifically concerned with AI/ML decisions being unbiased. For example, the use of AI/ML in the legal system or for hiring practices can have adverse impact on minorities (Guidotti et al., 2018). *Safety and performance explanations* focus on what steps were taken to strengthen the accuracy, robustness, security, and reliability of the model's decision or output. Importantly, the model methods provide the different types of explanations through the output of the model.

3.3 Types of transparency

Transparency, or the ability to understand AI/ML models and algorithms, is key to XAI models. Many different theoretical models of the design of XAI advocate transparency as a factor in their models (Arrieta et al., 2020; Miller, 2019; Muir, 1994). The term transparency has been used to express interpretability (specifically, intrinsic understandability of models, or white-box models) as well as in a more common usage manner to refer to the access of information related to explainability of AI/ML models (Meske & Bunde, 2020). We take the latter approach in this paper in order to bridge the concept of explainability within the XAI literature with the broader use of the transparency term in automation and human-machine interaction fields (e.g., Lyons, 2013). As such, we refer to transparency as a larger factor with many different sub-types that facilitate explanations.

Arrieta (2022) notes several different levels of transparency in their theoretical review of XAI. However, rather than viewing transparency as levels, we view them as factors that can influence the larger order constructs of explanations, which we reviewed above. Specifically, Arrieta observes the constructs of understandability, comprehensibility, interpretability, and explainability. Muir (1994) explains the key aspects of understanding machines are the dependability, reliability, competence, and helpfulness of the system. We note that Arrieta's terminology focuses on the user's perceptions of the procedures of the XAI (i.e., how the XAI arrives at its decision), with understandability, comprehensibility, and interpretability relaying rationale and data explanation types to the user. In contrast, Muir's factors of dependability, reliability, and competence relay safety and performance explanations. Lastly, Muir's helpfulness factor relays reasonability and impact explanations. Thus, the transparency factors mentioned above are conduits for the type of explanation posited by Zhou and colleagues (2021).

4. Trust in machines

In their influential article on human-machine interactions, Lee and See (2004) demarcate the trust process into individual differences, trustworthiness perceptions, trust, and risk-taking behaviors. Individual difference variables are any variable subsumed in the human operator that influences trust perceptions. These can be variables such as a general trust in automation, experience with a type of system, or a schema to view automation as perfectly reliable. Trustworthiness perceptions are the state perceptions of the system, which are comprised of performance, purpose, and process factors. The *performance* construct is the operator's understanding of how well the system achieves the goals of its task. The *purpose* construct is the operator's understanding of why the system was developed and for what reason. The *process* construct is the operator's understanding of how the system operates.

Transparency of the system is another key aspect of trust in machines in Lee and See's (2004) theory (per Lyons, 2013). Transparency is how the aspects of performance, purpose, and process are conveyed to the user (Lyons, 2013). Thus far, research in the area of automation may have conflicting findings with those in AI/ML. For example, Lee and See note that illustrating the process of the system, possibly by intermediate results, can increase trust in machines. However, in an AI/ML context, if the system's underlying process is explained to the user (e.g., XAI systems), but the reasoning lacks face validity or otherwise indicates a faulty process to the operator, this could potentially decrease the user's trust in the model (Chen & Barnes, 2014). Indeed, one issue that can arise in AI/ML systems analyzing data is that they could overfit the data by including spurious correlations between variables (Obermeyer & Emanuel, 2016). Ultimately, increasing transparency may not increase trust in the system, but it should help calibrate trust more appropriately to the system's true reliability (Chen & Barnes, 2014).

The trustworthiness constructs influence the operator's level of trust in the system, described as "the attitude that an agent [automation] will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability," and thus comprises a willingness to rely on the system (Lee & See, 2004, p. 54). Importantly, trust is an attitude not the actual behaviors of the operator. Risk-taking behaviors are the actual behaviors the operator takes to rely on the system, such as using an autopilot mode or accepting the outcomes of an AI/ML algorithm for decision making. Importantly, this creates a feedback loop, such that the operator continuously updates their perceptions of the system, and possibly their overall automation schema.

A defining feature of AI/ML systems that differs from other types of automation is that their decision-making process can iteratively update as new data is acquired (El Naqa & Murphy, 2015). Such changes can affect the feedback loop and the operator's trust in the system over time. Further, the operator may or may not have an influence on how the algorithms are changed over time (e.g., reinforcement learning versus unsupervised learning), which may affect trustworthiness perceptions in different ways (Dietvorst et al., 2018), thus necessitating transparency.

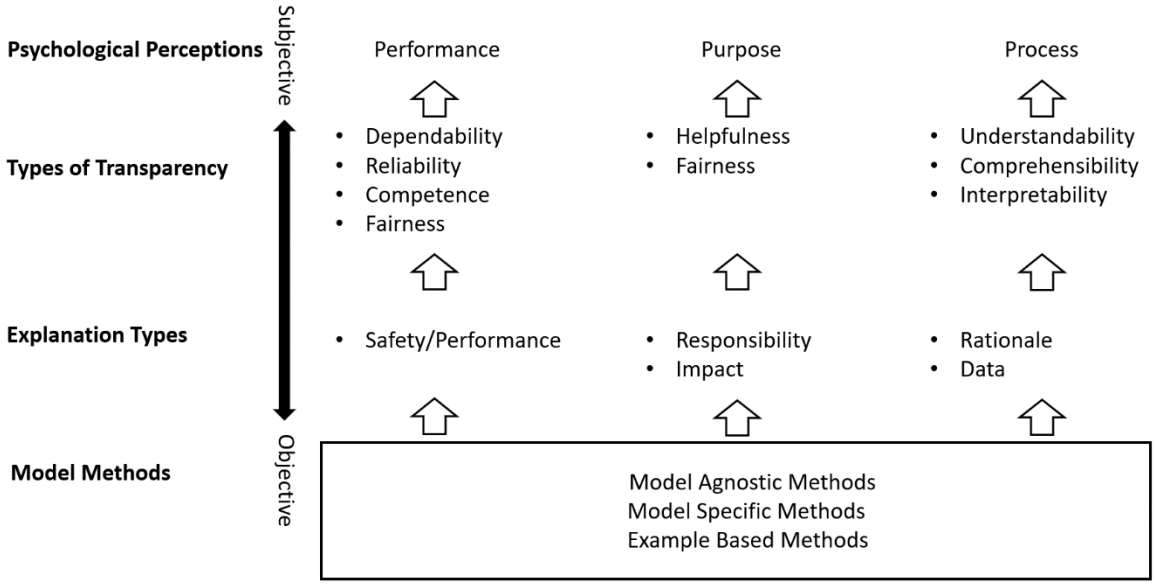

Figure 4. Proposed hierarchical framework of Explainable Artificial Intelligence methods and their relation to psychological perceptions of system trustworthiness.

5. Theoretical Integration of XAI Models

In this section, we integrate the theoretical taxonomies described above into one model. We postulate the theoretical structure outlined in Figure 4. The theoretical framework in Figure 4 illustrates the three subjective factors theorized by Lee and See (2004), with the various types of explanation and types of transparency subsumed under the factors. We created this model to reflect the spectrum of psychological perspectives from objective factors (model methods) to subjective factors (psychological perceptions). We note that the bottom of the figure represents the most objective taxonomies, namely the actual methods of XAI. The top of the figure represents the most subjective aspects of XAI, which are psychological perceptions. The explainability of a model is fundamentally a human perception that can vary between individuals. As such, we categorize the explanations and transparency factors into their respective trustworthiness categories. Importantly, we do not categorize the model methods into the trustworthiness categories because it depends on what data type is being relayed to the user and how they are interpreting it. We note, transparency is not a factor in the framework but rather a degree to which the performance, purpose, and process factors are salient to the human. In traditional black-box models, the purpose and process factors are largely obfuscated from the user, as the user does not have knowledge of how the AI/ML formed its decision. Applying various XAI methods increases the transparency of the purpose, process, and/or performance perceptions, which is the ultimate goal of XAI.

We begin at the bottom of Figure 4 with the most objective classification. The model methods outlined in the section above describe the broad aspects of the methods used to instantiate explainability into an AI/ML model. Importantly, we can classify explanation types and model methods as objective factors, whereas the type of transparency is more subjective. Model methods provide different types of explanations depending on how they are structured and what information they provide. This information is objective because it displays some aspect of the AI/ML that is not influenced by the user's perceptions but rather a description of the method being used. Similarly, the type of explanation is also objective because the factors focus on the

information being provided. In other words, the XAI is providing information about what it is interpreting from the data (data explanation) or how it processed the data (rationale explanation), or both. It is in the transparency factors that the human perceives and interprets the information. This is highly subjective because there are many individual differences between people that influence how transparency factors will be recognized and perceived. These differences are illustrated in Figure 4 with the black arrow demonstrating the continuum of objectivity and subjectivity. For example, a logistic regression model that provides a rational explanation through a SHAP algorithm may result in different beta weights for each predictor, given the weight of the other predictors when they are not present. However, this interpretation of the resultant beta weights is reliant on the user's experience with logistic regression. A user who has little-to-no knowledge of logistic regression may not perceive the differences in beta weights as revealing much about the model, but a user experienced with logistic regression may find the resulting SHAP output helps to clarify the underlying methods of the model. As such, the subjectivity of the transparency factors necessitates a psychological theory, as they are ultimately subjective perceptions. To further elucidate the theoretical underpinnings of each subjective psychological perception we outline the perception, its types of transparency, and the data types they are built on.

5.1 Performance

The first trustworthiness perception we discuss, performance, may be the most relevant and salient aspect of trustworthiness. Although researchers do not traditionally think of performance as an XAI construct, it is a key aspect of XAI. The performance construct of trustworthiness is informed by the safety and performance explanations provided by the system. The safety and performance explanations communicate the system's dependability, reliability, and competence. Interestingly, performance can also be a double-edged sword. An AI/ML algorithm that has a high reliability rate can lead to complacency and overreliance on the system (Parasuraman & Manzey, 2010). Research has demonstrated that lowering the reliability can reduce the complacency of the user on the system (Banks et al., 2018; Parasuraman & Manzey, 2010).

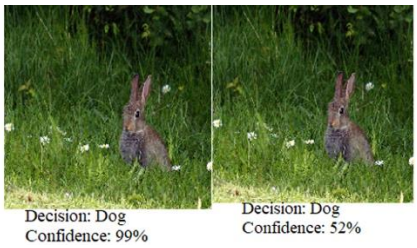

Figure 5. Left picture is traditional ML. Right Picture is adversarial distance confidence intervals.

One problem with traditional AI/ML systems is that even when they are incorrect, they have high confidence in their decisions. Recent research advances have harnessed adversarial distances to create more accurate confidence intervals of the AI/ML model's decisions (Bennette et al., 2020; Tomsett et al., 2020). These AI/ML models are trained to understand that there may be images outside of the classifications it was trained on, leading to more accurate confidence intervals when the AI/ML algorithm encounters datapoints outside of its training dataset. For example, Figure 5 illustrates the results of an AI/ML algorithm that has been trained to classify cats and dogs. As illustrated on the left-hand side of Figure 5, the traditional AI/ML classifies the image as a dog and has high confidence that the image is a dog. The newer adversarial distance algorithm, shown on the right-hand side, still classifies the image as a dog, but the confidence is much lower given that it does not meet all the criteria of a dog image.

Importantly, these new AI/ML systems that apply the adversarial distance algorithms may be able to solve the reliability issue mentioned above. In other words, an AI/ML algorithm may be able to correctly classify cats and dogs, but any other image that it has not been trained on, such as a rabbit, can lead to low confidence in the machine's decision, alerting the operator that the system does not know how to handle a datapoint and decision authority should transition over to the operator. These more accurate performance metrics may increase the transparency of the reliability and competence of an AI/ML system. These increases in transparency metrics will, in turn or by extension, facilitate a more accurate perception of the AI/ML's performance.

5.2 Purpose

Second, we discuss the purpose dimension of trustworthiness. As noted above, the purpose dimension is concerned with the user's understanding of the applicability of the system. This purpose dimension is informed by both the responsibility and impact explanations. The responsibility explanations concern who developed the system and who is accountable for the implementation and the decisions of the system. The impact explanations can help to elucidate the broader

implications of using the AI/ML in a larger contextual manner. These two explanations facilitate the perceived helpfulness transparency factor. If the AI/ML is viewed as helpful, the system will be used and widely applied in the context, but if it acts as a hinderance to the user, it may be removed from the model. For example, the AI/ML model used in Figure 5 that is trained to classify images of cats and dogs but is shown images outside of its training set, such as the rabbit shown, will incorrectly classify the new images. Users may perceive the AI/ML as being improperly utilized, decreasing purpose perceptions and overall trust (Lee & See, 2004). Although this may not seem problematic for an animal image binning task, as the risk of using the system's output increases, the appropriateness of the AI/ML and the need for accuracy may increase, such as in military contexts.

5.3 Process

Third, XAI provides insight into how AI/ML algorithms make decisions, which represents the process aspect of system trustworthiness. The majority of the XAI literature is focused on providing increased rationale and data explanations, which are inherently process perceptions. These types of explanations rely on the transparency methods of understandability, comprehensibility, and interpretability. These transparency methods elucidate a perception of how the system functions, which is inherently a user perception. Giving the operator an understanding of how the AI/ML functions can facilitate trust in the system, particularly because the user may be aware of situations when the AI/ML may fail. For example, understanding what information the model uses from a gradient saliency map can inform the user if the model is using aspects of an image that are not relevant (e.g., if a model focuses on the background of an image such as grass, instead of the animal it is supposed to categorize; see Figure 2). The rationale and data explanations facilitate an understanding of the aspects of the data the AI/ML is using. However, the link between process and trust may not be a direct linear relationship. Increased transparency about the model's process may not always facilitate increased trust or greater explainability of the system. For instance, as more transparency about the model's process is provided, it can increase the operator's cognitive workload (Bainbridge, 1983). As such, the process construct may serve as a curvilinear relationship with trust, such that increases in the transparency of the process can lead to increased trust in the system to a certain point. After that point, there may be too much cognitive workload, such that either over- or under-reliance may occur because of the distracting nature of the underlying process information (ICO & Turing, 2019). Therefore, it is important to determine whether and when process information should be conveyed to the user.

5.4 Relationship of trustworthiness factors

We note that the performance, purpose, and process factors are not necessarily orthogonal in nature. Although the factors inform different aspects of trust and explainability, they are predicated on the actual information the model is providing. The complexity of the information can lead to several different trustworthiness factors. For example, the AI/ML model that classifies animal images, which we mentioned above, may facilitate performance and purpose factors. If the model was trained on images of cats and dogs, but during the use of the system it encounters images that are outside of its training data set, such as the rabbit shown in Figure 5, the user may be informed about two trustworthiness aspects of the model, performance and purpose. The purpose of the model was to classify cats and dogs, but it is now being used outside of its appropriate or intended application which results in lower performance. The user can ascertain the model is not appropriate for the data set it is being used on, thereby causing lowered performance. All of this can be determined from the simple knowledge acquisition of how the model was trained.

5.5 Differential and temporal effects

Although Lee and See (2004) delineate much of the human-machine trust process, we expand on their interpretation by discussing varying degrees and extents to which the constructs are relevant in XAI situations (i.e., performance, purpose, and process). We postulate the trustworthiness constructs are all related but have differential relevance depending on the situation. First, we view transparency of performance as a necessary condition for understanding and trust in AI/ML (and also machines in general). It is hard to imagine a scenario in which an operator does not care about the performance of the algorithm or machine. Indeed, the performance of the system provides the necessary information about the system to update the feedback loop of the operator that Lee and See mention. Conversely, the constructs of purpose and process may not always be necessary components for trust in a system. A good example of this can be found in predictive modeling in health care. A hospital started using predictive modeling to identify patients with a high probability of being readmitted (Health Care Innovations Exchange, 2009). Despite using the system for several years, the only aspect known to the users was the performance. Banerjee et al (2018) later developed

a deep learning predictive model of short-term life expectancy in hospital care by examining medical records which added additional explainabilty through an interactive graphical tool that helped physicians understand the predictions. As such, the performance of the system was always present to the user, but the basis for the decisions (i.e., the process) was lacking until the graphical tool was developed.

The case of the hospital predictive model raises important questions about if and when purpose and process information is important to the operator. We contend that it is typically only when an AI/ML algorithm performs an incorrect or unanticipated behavior that purpose and process information is important. As illustrated with the complacency in machines literature, if the machine performs too well the user will over rely on the machine (Parasuraman & Manzey, 2010). We note the difference between the two outcomes. An incorrect behavior is a decision in which the AI/ML makes a mistake or error. For example, the classification of the rabbit as a dog in Figure 5 is a mistake. Additionally, the operator may also inquire into the purpose of the AI/ML upon receiving an error. The operator may explore the dataset on which the AI/ML was trained and realize the model is only useful for classifying images of cats and dogs, given its training. The operator may want to then either expand the training data to include more variance in data types (i.e., a greater variety of animals) or continue processing while monitoring for images that fall outside of the system's trained purpose. An unanticipated behavior, on the other hand, is a decision given by the AI/ML model the operator was not expecting but may not be incorrect. For example, an AI/ML algorithm employed by Target Corporation correctly identified a teenager as being pregnant and began sending her coupons for baby supplies. The teen's father confronted Target but soon after found out that his daughter was indeed pregnant. Although the algorithm's purpose was to identify customers that may be pregnant based on their buying habits, the identification of a teenager who was pregnant was an unanticipated but correct result (Hill, 2012).

Operators of a system that performs reliably may not be interested in the process factors that lead to the reliable performance. It is after the algorithm makes a mistake or has an unintended consequence that operators will want to delve into the processes of the algorithm. Indeed, this follows much of the history of the development of AI/ML (McCoy et al., 2022). It is only recently that researchers have begun to explore the process functions of the AI/ML algorithms to gain a better understanding of how the algorithms err, so as to ultimately improve performance and trust in the system.

Of particular interest to AI/ML errors/mistakes are newly developed adversarial distance algorithms (Bennette et al., 2020) described in Section 5.1 Example-based methods. AI/ML systems that can reliably alert the operator to issues with classification that may have a differential effect on performance than previous reliability research would suggest. Researchers have found that decreasing the reliability of the system actually improves performance in a task when paired with a human operator. This is due to the operator's willingness to exert more effort monitoring the machine, for the operator knows the machine is not 100% reliable. The adversarial distance algorithms may provide information to the user about the confidence of the AI/ML model's decision. If the AI/ML system is reliable in this context (i.e., the AI/ML correctly classifies data it has been trained on and alerts the operator to data that falls outside of its training classifications), it may be a good compromise between too-high and too-low reliability, avoiding complacency and underuse, respectively (Hill, 2012).

5.6 Risk as a moderator

There are also differences in the risk associated with the use of an AI/ML model. The amount of risk associated with the system's decision is one important consideration when evaluating the level of transparency that should be provided with an AI/ML's decision. When considering performance, purpose, and process aspects of transparency, the decision or output's impact (e.g., safety considerations) will largely determine the information provided to the user. Using an AI/ML model for medical diagnoses, for example, will require a higher level of performance and a more accurate representation of true system performance to the user than an AI/ML model being used to predict a suggested friend on a networking website (Arrieta, 2022). Additionally, declining an applicant's loan request based on the decisions of an AI/ML model would require the financial institution to disclose each of the factors that affected the decision, as one example (Bibal et al., 2021).

6. Future Research

First, we note the current paper focused on relatively simple applications of the model to AI/ML contexts, specifically image classification. We utilized these examples to clarify the concepts across different psychological processes. As such, future research should apply these psychological principles to a variety of XAI methods. Second, we note the relative dearth of literature on the psychological perceptions of XAI models that are currently being developed. Research would do well to test the efficacy of XAI in relating information to the user. Specifically, psychological

research with users that assess perceptions through a variety of metrics such as Likert scales, open-ended questions, and possibly physiological metrics could help researchers understand the benefits of XAI, if any. Research in this area can also influence the development of XAI, such that model methods that do not provide any useful information to user perceptions may be abandoned for more fruitful endeavors that increase transparency and trust in the systems. Third, future research should explore when the user wants clarification of the underlying processes from the AI/ML. We theorized above that transparency of performance is the most salient aspect of the interaction and only after a mistake or failure will the user want more information. Research should explore the temporal effects theorized in the paper. Lastly, future research should employ the proper terminology to help understand the impact of the XAI on the relevant construct. In other words, if research is exploring XAI, care should be taken to discuss the data explanation, transparency type, and possibly the underlying psychological process the XAI is created to facilitate.

7. Conclusion

The current paper sought to create an overarching framework of XAI. In this paper we categorized the previous taxonomies according to the information they provided about the AI/ML model they were designed to explain. Specifically, we categorized the previous taxonomies based on their degree of subjectivity, how they relate to each other, and how they inform different levels of AI/ML transparency. The model from this paper can help researchers more accurately define what aspects of the XAI are influential in human perception and through which mechanism(s) the XAI model is hypothesized to influence perceptions. The formation of a unifying theoretical model (shown in Figure 4) allows researchers from various fields to better understand how XAI methods inform the user from a psychological perspective. This viewpoint is vital to the advancement of the AI/ML and XAI fields because the notion of explainability within these domains is intrinsically psychological. XAI model developers can use this framework to better ascertain whether the model they are creating provides the information that it is intended to. From this point, researchers can empirically study how this information impacts the user and their perceptions of the AI/ML system, for example the system's trustworthiness. Lastly, we note the lack of research in trust and reliance on AI/ML. Little research has been conducted on the trustworthiness perceptions of AI/ML or XAI models. Researchers in the computer science field have recommended utilizing XAI to increase transparency in the models (Rajabiyazdi & Jamieson, 2020). However, no research to date has explored whether the XAI models actually increase operator perceptions of transparency. Research would do well to examine the psychological impact of the model methods currently being developed to ensure they do add to the understanding of the users.

8. References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Agency for Healthcare Research and Quality. (2022, Aug. 20). Health care innovations exchange. https://www.ahrq.gov/innovations/index.html

Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd ed). MIT Press.

Arrieta, A. B. (2022). On the Design, Implementation and Application of Novel Multi-disciplinary Techniques for explaining Artificial Intelligence Models.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Bainbridge, L. (1983). Ironies of automation. In *Analysis, Design and Evaluation of Man–Machine Systems* (pp. 129-135). Pergamon.

Banerjee, I., Gensheimer, M. F., Wood, D. J., Henry, S., Aggarwal, S., Chang, D. T., & Rubin, D. L. (2018). Probabilistic prognostic estimates of survival in metastatic cancer patients (PPES-Met) utilizing free-text clinical narratives. *Scientific Reports*, *8*(1), 1-12.

Banks, V. A., Plant, K. L., & Stanton, N. A. (2018). Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. *Safety Science*, *108*, 278-285.

Bennette, W., Maurer, K., & Sisti, S. (2020). Harnessing Adversarial Distances to Discover High-Confidence Errors. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-10). IEEE.

Bibal, A., Lognoul, M., De Streel, A., & Frénay, B. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, *29*(2), 149-169.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Chen, J. Y., & Barnes, M. J. (2014). Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13-29.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics* (pp. 296-299). Auerbach Publications.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155-1170.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In *Machine Learning in Radiation Oncology* (pp. 3-11). Springer, Cham.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An Introduction to Statistical Learning: With Applications in R*. Spinger.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1-42.

Hill, K. (2012). How target figured out a teen girl was pregnant before her father did. Forbes. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=56584b416668

ICO & Turing. (2019). Explaining decisions made with AI: Draft guidance for consultation - Part 1: The basics of explaining AI. Wilmslow, Cheshire, UK: (Information Commissioner's Office and The Alan Turing Institute).

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1-8.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 8:4, 277-301.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

McCoy, L. G., Brenna, C. T., Chen, S. S., Vold, K., & Das, S. (2022). Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology*, *142*, 252-257.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98-119.

Meske, C., & Bunde, E. (2020). Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. In *International Conference on Human-Computer Interaction* (pp. 54-69). Springer, Cham.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT press.

Molnar, C. Interpretable Machine Learning. 2019. Available online: https://christophm.github.io/ interpretable-ml-book/ (accessed on 22 January 2019).

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905-1922.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, *375*(13), 1216.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (pp. 506-519).

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381-410.

Rajabiyazdi, F., & Jamieson, G. A. (2020, October). A review of transparency (seeing-into) models. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 302-308). IEEE.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*, 1-85.

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, *9*, 4843-4873.

Sheu, Y. H. (2020). Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Frontiers in Psychiatry*, 1091.

Suman, S., Roy, S., Prakash, S., Kumar Singh, R., & Kumar, S. (2020). Understanding and improvisation of human perceptions using artificial intelligence and machine learning. *Journal of Mathematical Sciences & Computational Mathematics*, *1*(2), 251-256.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*. 3

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., ... & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, *1*(4), 100049.

Zhou, J., Gandomi, A. H., Chen, F. & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics, 10(5),* 593.