ABSTRACTION, VALIDATION, AND GENERALIZATION FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE

Scott Cheng-Hsin Yang
Department of Mathematics and Computer Science
Rutgers University
101 Warren Street, Newark, NJ 07102
scott.cheng.hsin.yang@gmail.com

Tomas Folke
Department of Mathematics and Computer Science
Rutgers University
101 Warren Street, Newark, NJ 07102
tomas.folke@rutgers.edu

Patrick Shafto
Department of Mathematics and Computer Science
Rutgers University
101 Warren Street, Newark, NJ 07102
patrick.shafto@rutgers.edu

April 30, 2021

ABSTRACT

Neural network architectures are achieving superhuman performance on an expanding range of tasks. To effectively and safely deploy these systems, their decision-making must to be understandable to a wide range of stakeholders. Methods to explain AI have been proposed to answer this challenge, but a lack of theory impedes the development of systematic abstractions which are necessary for cumulative knowledge gains. We propose Bayesian Teaching as a framework for unifying explainable AI (XAI) by integrating machine learning and human learning. Bayesian Teaching formalizes explanation as a communication act of an explainer to shift the beliefs of an explainee. This formalization decomposes any XAI method into four components: (1) the inference to be explained, (2) the explanatory medium, (3) the explainee model, and (4) the explainer model. The abstraction afforded by Bayesian Teaching to decompose any XAI method elucidates the invariances among them. The decomposition of XAI systems enables modular validation, as each of the first three components listed can be tested semi-independently. This decomposition also promotes generalization through recombination of components from different XAI systems, which facilitates the generation of novel variants. These new variants need not be evaluated one by one provided that each component has been validated, leading to an exponential decrease in development time. Finally, by making the goal of explanation explicit, Bayesian Teaching helps developers to assess how suitable an XAI system is for its intended real-world use case. Thus, Bayesian Teaching provides a theoretical framework that encourages systematic, scientific investigation of XAI.

***K*eywords** Explainable AI, Bayesian Teaching, Design patterns, Cognitive Science, Human Computer Interaction

1 Introduction

Over the past decade neural network architectures have had impressive performance gains, reaching human or even super-human performance in many tasks, including speech recognition, translation, and image classification [1]. These architectures have the potential to revolutionize many human activities, including logistics, medicine, and law [2, 3, 4]; however, the responsible and safe deployment of these systems depend on them being understandable to human stakeholders [5]. Two solutions have been suggested to this problem: one is to design systems that are inherently interpretable or transparent, which often involves a compromise in performance; the second is to develop bespoke solutions to explain the decision-making of an obscure system post-hoc [6]. In this paper, we present a third approach in

which explainability is analyzed as a problem of facilitating understanding of AI systems by humans. Thus, we propose a general approach to explaining AI systems by explicitly analyzing the problem of providing information that enables a human to understand and predict the AI.

The literature on explainable AI has exploded in recent years [7], but there is still a dearth of coherent theoretical frameworks of XAI techniques [8], and the taxonomies that do exist are based on the technological substrate underlying the explanation techniques as opposed to their pragmatic goals. This lack of theory hobbles XAI research because it obscures what lessons can be safely transferred between studies and applications, and which components need to be re-validated in new contexts. As a consequence, it reduces both the speed of knowledge accumulation and deployment of safe, explainable AI systems across sectors. Additionally, most XAI solutions tend to be designed by software engineers for engineers, and as such do not consider how to explain the target system to non-technical users [9, 10, 11, 12]. This is problematic because successful explanation clearly depends on the users and their goals [13], and if an AI system is successfully deployed, software engineers will be a small subset of the users.

Explainable AI is a complex problem, with both technological and psychological components. A theoretical framework that formulates the problem of XAI in structured and normative ways may surface associations between different methods and results that previously seemed disconnected. Such a framework also decomposes XAI problems into abstractions that represent fundamental components and dependencies, that can be validated separately. Furthermore, such a modular approach would support deployment, as it would allow formal testing as to what sub-components of explanation generalize to what contexts. We propose Bayesian Teaching, which formalizes explanation as a social act between a teacher and a learner, as such a framework. In the following section we will explain how Bayesian Teaching abstracts any XAI system into the following four components: (1) the target inference, (2) the explanatory medium, (3) the explainee model, and (4) the explainer model. For concreteness, we show how to apply Bayesian Teaching to decompose popular classes of XAI methods. Then, we illustrate how the decomposed parts can be validated semi-independently through user studies, and reflect on how Bayesian Teaching promotes human-centeredness in XAI research and application. Finally, we discuss generalization of the decomposed parts, including remarks on the manipulation and recombination of the components.

2 Bayesian Teaching

Bayesian Teaching formalizes explanation as a communication act between the explainer (teacher) and the explainee (learner) by the following equation:

$$P_T(x|\Theta) = \frac{P_L(\Theta|x)P(x)}{\sum_{x' \in \Omega} P_L(\Theta|x')P(x')}. \tag{1}$$

The equation describes how a teacher P_T should select an explanation x to best explain a target inference Θ, contingent on their model of the learner P_L. Specifically, it says that the probability of choosing an explanation x to explain the target inference Θ is proportional to the probability that the explanation x would lead the learner model P_L to the target inference Θ. Thus, Bayesian Teaching explicitly decomposes the explanation generation process into four components: the model or inference to be explained Θ; the explanation x; the learner model, which is captured by the posterior $P_L(\Theta|x)$; and the teacher model, which is captured by the selection posterior $P_T(x|\Theta)$. The explanation is selected from a pre-specified set Ω, in which each element has a prior probability $P(x)$ of being selected.

The aim of XAI is to improve human users' understanding of AI systems. As such, a successful explanation should shift the user's belief to increase the fidelity between their internal model of the AI system and the AI system itself. Bayesian Teaching provides a formal account of such intentional belief-shifting via explanation. One implication of the belief-shifting perspective is that the success of an explanation is determined by how much it shifts a user's belief in a desired direction. Bayesian Teaching specifies the components—Θ, x, $P_L(\Theta|x)$, and $P_T(x|\Theta)$—required to quantitatively model this shift. Below we define and explain each of these components.

The target inference Θ is an aspect of the model that human users wish to understand. Possible target inferences range from global aspects of the model, such as model parameters, to intermediate components, such as the model's latent variables [14], to local targets, such as the model's prediction on a particular data point [15]. Local targets can be curated into a curriculum to inform model behavior on a holistic level. The size and complexity of the curriculum capture the trade-off between explanation completeness and explanation complexity. In general, the target Θ is related to the behavior of the model to be explained. Specifically, one should consider how the target Θ aligns with the actual use cases in deployment, such as debugging, verification, or acceptance testing.

The explanation x is the object provided to the end user to induce understanding about Θ. Common explanation media include instances from the training data, features of the data (e.g. saliency maps), and simplified models that accurately describe the target model for some subset of the problem space. Toward the goal of enhancing understanding, a key

consideration when picking an x is ease of processing by human users. Appropriate media can often be derived from the model's training data, as they are typically curated by and hence understandable to humans. Using XAI of image classification as an example, one can take the training images at different level of granularity to produce different types of x, including the images themselves [16, 15], regions of an image [17], or pixels of an image [18]. Other media that are easy to process are intuitive decision rules, which can be captured as distilled rule sets or even entire models, such as decision trees and linear models as is done in mimic learning.

The learner model $P_L(\Theta|x)$ is a computational model that describes how the user makes inferences about Θ when given the explanations x. All XAI methods have a learner model, explicit or implicit. We advocate making the learner model explicit to allow validation of this crucial component. The quality of the explanation generated depends on the quality of the learner model used. An inaccurate learner model will lead to unreliable and confusing explanations because the mapping between x and Θ would be inaccurate. Conversely, a perfect learner model could lead to optimal explanation. The specifications of x, Θ, and $P_L(\Theta|x)$ provide the input, output, and the assertions to be tested in a validation. In much of the XAI literature the learner model is embodied as a loss function. The loss function composes of two parts: a *mapping* from x to a Θ' that exists in the same mathematical space as Θ, and the loss part that can be used to assign a probability to the target Θ based on how different Θ' and Θ are [19].

The teaching model $P_T(x|\Theta)$ specifies the explanation-generation process. As Equation 1 suggests, this selection process is largely determined by the learner model. This is intuitive because a good teacher should consider the learner when selecting an explanation. To find the optimal explanation, one could search for the x that maximizes $P_L(\Theta|x)$, and hence $P_T(x|\Theta)$. This solution is equivalent to finding the x that maximizes the numerator of Equation 1 and thus avoids the computation of the denominator. Other approaches to inference include sampling from the $P_T(x|\Theta)$, which will provide a sense of the relative effectiveness of near-optimal explanations.

The Ω and $P(x)$ terms describe additional constraints on the explanations considered in the explanation-generation process. Ω specifies the pool of explanations to select from, and $P(x)$ specifies which element in the pool is more likely *a priori*. Factors that influence the design of Ω include cognitive load and intuition on regions on interest. Returning to the image classification example, one may want to show a limited number of images as explanations to avoid cognitive overload or limit saliency maps to only the lungs in an x-ray image for pulmonary disease detection. In the literature, $P(x)$ is often used to control the complexity of the explanation. For example, if x are decision trees, shallower trees are given higher prior probability than deeper ones. $P(x)$ can also capture the cost for constructing the x in terms of computational resources, manual labor, etc.

3 Decomposition via abstraction

In the previous section we have identified abstractions central to Bayesian Teaching. Here we demonstrate how these abstractions facilitate decomposition of existing methods into component parts. We consider three popular classes of XAI methods: explanation-by-examples, explanation-by-features, and mimic learning. These three methods differ with regards to what explanatory medium they use (the x term), which is the most common distinguishing factor between current XAI methods. However, all also make commitments regarding the target inference Θ, the learner model $P_L(\Theta|x)$, and the sampling of explanation in $P_T(x|\Theta)$, as we illustrate in the following subsections.

3.1 Explanation-by-examples

Explanation-by-examples seeks to explain the behavior of a target model by presenting a subset of cases from the training data that strongly influenced the model's inference. For instance, to explain why a classifier categorized a certain image as a cat as opposed to other animals, examples of cat may be provided to show what the classifier considers as prototypical cats. The explanation-by-examples approach has many desirable properties: It is fully model-agnostic and applicable to all types of machine learning [20, 21, 22]; it is domain- and modality-general [23, 24]; and it can be used to generate both global explanation [25, 16, 14, 26, 27] and local explanation [28, 29, 30]. In the context of Bayesian Teaching, explanation-by-examples is obtained by setting x to be a data point from the training data [31]. Below we use the Bayesian Teaching framework to decompose two existing explanation-by-examples methods into the components described in Section 2.

Suppose you have an image classifier and you would like to understand how the it represents its target classes. In a previous study based on Bayesian Teaching [14], the authors address this issue by finding a few examples from the training data that captures the target model's class representations. Here the target Θ is the latent class means of a probabilistic linear discriminant analysis (PLDA) model trained on the whole training dataset. The explanation x is a set of three images from the training dataset for each class. The learner model, $P_L(\Theta|x)$, computes the probability that a PLDA (the mapping) assigns to Θ when trained on x instead of the full dataset. The teaching is based on max selection

from $P_T(x|\Theta)$. The Bayesian Teaching paper [14] also proposes that the goodness of x can be evaluated by whether the explanations help humans predict the target model's classification in a two-alternative-forced-choice (2AFC) task.

Similar to the previous method for explaining the class representation of a classifier, in the maximum-mean-discrepancy-critic (MMD-critic) method proposed in [16], the authors aimed to find a small subset of data that can represent the entire data set (as opposed to the latent means of the classes). They did this by modelling the data distributions which allows for the comparison between the target distribution and the learner-model induced distribution needed for example selection. Here the model to be explained is a kernel function on top of a deep neural net (DNN), and the target Θ is the data distribution of a particular class induced by this model. The explanation x is a set of images from the training data, with Ω restricting the set size and the classes from which the images can be sampled. The learner model, $P_L(\Theta|x)$, is a function that describes how similar the data distribution induced by the kernel function (the mapping) on x is to Θ. The teaching process of selecting x to induce the desired data distribution Θ is done via max selection. More specifically, the work presents two sub-targets: "prototype" and "criticism." The prototype selection aims to make the distribution induced by the learner model on x match Θ as much as possible, whereas the criticism selection aims to make the two differ as much as possible. The MMD-critic paper [16] also presents an evaluation task that measures how much the explanations help humans predict the model's predictions in terms of accuracy and response speed. To evaluate the model with human participants, the authors bridged the gap between Θ (a data distribution) and the evaluated class prediction by a nearest-neighbor-classifier based on the kernel distance in the DNN's final feature space.

3.2 Explanation-by-features

Explanation-by-features seeks to explain AI decisions by drawing attention to sub-components of an instance of the data that influence the output decision. In other words, this class of methods sets the x to be a constituent of a data point, such as regions of an image, phrases in a document, or elements in a vector. Saliency maps are a form of explanation-by-features used in image classification, and is among the most popular XAI methods in recent years [32]. Below we decompose two existing methods for generating saliency maps through the lens of Bayesian Teaching.

The Randomized Input Sampling for Explanation (RISE) method for generating saliency maps is a simple, model-agnostic method that only requires access to the output probabilities — and not the internal workings — of the target model to be explained [33]. The target Θ is the predicted class label of a test image from the target model, typically a convolutional DNN. The explanation x is a mask over the test image: pixels that are salient for the classification are fully unmasked. The learner model, $P_L(\Theta|x)$, outputs the predictive probability of the class label specified in Θ by passing the test image masked by x into the target model (the mapping). Teaching selects a mask by finding the *expectation* over the mask for a particular class. While finding a mask that maximizes $P_T(\Theta|x)$ is possible, that mask is likely to focus on a single salient region, whereas the expectation of the mask is likely to capture *all* salient regions. Indeed, despite differences of algorithmic approach, RISE can be viewed as a special case of Bayesian Teaching [15].

In the SHapley-Additive-exPlanation (SHAP) method presented in [18], the authors generate feature saliency values, which are the weights of a linear model that locally matches the target model's inferences. Here, the target Θ is to perfectly match the target model's predictive distribution on a test data point. The explanation x is the saliency of features, such as individual pixels of an image or words in a document. The learner model, $P_L(\Theta|x)$, assumes an additive linear model (the mapping), where the saliency values x are the weights of the linear model. The learner model assumes two additional constraints referred to as missingness and consistency. Given the form of the model and the target Θ to perfectly match a distribution, the authors prove that there is only one solution (ie., one set of weights) that satisfy these constraints. This translates to the teaching process, $P_T(x|\Theta)$, being a delta function on the solution. The SHAP paper [18] evaluates the goodness of x by asking human participants to assign weights to certain features in order to compare the human-assigned weights to those generated by the method. Note that the SHAP method is closely related to the LIME method of [17], which we will discuss in the following subsection. This connection provides a link between how explanation-by-features and mimic learning may morph into each other.

3.3 Mimic learning

Mimic learning, sometimes referred to as model distillation [34], is a class of explanation methods where the behavior of a complex, obscure model is approximated by a simpler model that is easier for humans to interpret [35]. Popular examples are approximating the local behavior of a deep neural network with linear models [17] or decision-trees [36]. In terms of the components introduced in Section 2, the explanation x is usually also the parameters of the learner model $P_L(\Theta|x)$ in mimic learning.

The pioneering LIME method [17] uses linear models to approximate the behavior of the target model locally. Here, the target Θ is the decision boundary of the target model in the neighborhood of a test data point, where neighborhood is defined by a combination of the kernel function and data augmentation used. The explanation x is the weights of a

linear model. This linear model maps the weights x to a linear decision boundary. The learner model, $P_L(\Theta|x)$, is a function that quantifies how well the mapped decision boundary matches the decision boundary of the target model. The teaching component of weight selection is done via maximizing $P_T(x|\Theta)$. To transform the weights into a form that can be easily processed by end users, the authors presented the features—such as regions of an image or words in a document—that have positive weights. [1] The LIME paper [17] presents three high-level tasks to evaluate explanations: participants were tasked with competency testing (predicting which classifier generalizes better), feature debugging (identifying harmful features), and model anomaly detection (identifying classifier irregularities).

The work in [36] distills a neural network into a soft decision tree. The target inference Θ is the predictive distribution given by the target DNN on a set of test data points. The explanation x is a soft binary decision tree, which needs to be further visualized when provided to a human end-user. The learner model, $P_L(\Theta|x)$, is a function that quantifies the match between the prediction distribution from the soft binary decision tree with parameter x (the mapping) and that from the target DNN to be explained. $P(x)$ is set to favor trees that have high-entropy in the paths taken over the target test dataset. The teaching process of selection x is done via maximizing $P_T(\Theta|x)$.

3.4 Summary

In this section we have demonstrated how existing XAI methods can be broken down into components according to the abstraction provided by Bayesian Teaching. Such demonstration highlights that all XAI methods necessarily make commitments about Θ, x, $P_L(\Theta|x)$, and $P_T(x|\Theta)$, and that these commitments can be made explicit. Having done the decomposition, one can now validate and ground each component, as discussed in the next section. The decomposition resulting from the abstraction provides a template to reason about which insinuations of the different components can fit together. Component-specific validation tests the extent to which a component would generalize across selections of other components and use cases. Therefore, the abstraction, validation, and generalization made possible by Bayesian Teaching underscore the usefulness of our framework in systematizing research and development.

4 Validation of components

In software development a *unit test* is a piece of code that invokes one unit of work and checks whether that unit operates as intended. When developing a new function it is considered good practice to write a unit test prior to writing the function itself [37]. One reason for starting with designing the unit test is that it assures the programmer that they are trying to solve a well-specified problem and that they will know when they have succeeded. Unit-testing in XAI is desirable for similar reasons. The strict control a unit-test offers in software development might be unrealistic for much of XAI work, but as an analogy it captures what we should aim for. Specifically, we want to systematically test the dependency between inputs and outputs of a specific XAI component. This is different from a software unit test, in that components cannot always be individually evaluated the way software units can, but a test can still be designed to evaluate one specific component. Just as unit-tests allow for robust integration of functions, appropriate evaluation of XAI components supports their generalization beyond the original contexts in which they are introduced.

In software development, a successful unit-test determines that the evaluated unit produces the desired input-output dependency and nothing else. As we have alluded to throughout this article, the aim of XAI is to move the human user's beliefs so that they accurately capture the AI system; hence, any unit test of an XAI component must be a user study. The specific design of the user study and the metrics recorded should be informed by the real-world use case the XAI system aims to solve: such as tweaking an AI system prior to release, improving human-AI teaming, or enabling effective auditing. When AI is to influence high-stakes decisions in areas such as law and medicine, it might be helpful to specify an assertion statement that puts thresholds on the metrics evaluated. For example, users need to accurately catch the AI's mistakes 95% of the time, or they must have a normalized ranking loss below .05 when ranking the importance of the input features on a given decision. The fact that one component passes a unit-test does not imply that it is suitable to explain any XAI model, for any user, with any goal, just like a software function passing a unit-test does not guarantee it will work in every software package. However, a successful unit-test provides structured evidence that one component of the XAI solution meets some clearly defined criterion, and specifies the input-output interface that ensures the proper working of that component. This information enables researchers and engineers to effectively reason about whether a component is appropriate for their application, and if it is, they can confidently integrate it into their application, leading to significant reductions in development time.

[1] Strictly speaking, this visualization invokes another Bayesian Teaching problem where the x are the features, and the rest of the components are left to be specified.

4.1 Validation of Target Inference

Evaluation of the target inference Θ requires x because x is the communication medium between the user and XAI system. However, in this case x is auxiliary in that it does not *need* to come from a sensible learner model, hence making the test independent of the learner-model component. The assumption of independence of the learner model is reasonable when the range of x involved in the task can be efficiently covered by uniform sampling. Such evaluation tests whether it is possible to achieve Θ by *some* x, but does not speak to the optimality of a *specific* x.

To evaluate the modularity of Θ, we can run a user study to measure a particular user metric on a given set of Θ's and x's. For example, one can design the set of Θ's to be the predicted classes of three test images from three different classes, the set of x's to be example sets with different set size (e.g., 2, 3, or 4 images in a set), and the user-study metric to be the probability that the user would predict a Θ given an x. If all the x's result in the same rank-order of Θ's on the performance metric (e.g., if human's prediction is always poor on a particular class regardless of the number of examples shown), we can conclude that that Θ is independent of the set of x's considered. On the other hand, if the rank-order of Θ varies by x, the two should always be considered as a unit. Generalizing further, if the independence of Θ is established over a wide range of x, it increases the probability that it will work similarly with a novel, yet untested x; otherwise, each novel pairing need to be explicitly evaluated.

In evaluating Θ, it is also important to distinguish between two forms of belief-updating in the context of XAI application. When auditing AI systems, the assumption is that the human user has access to the ground truth. Consequently, the XAI solution should help the user to develop an accurate model of the AI's "beliefs," captured by the user's ability to predict the AI's generalization behavior. In this case, Θ would be focused on predictive behavior, as in most of Table 1. When collaborating with AI systems, humans may have worse access to the ground truth than the AI. Consequently, the XAI and AI systems should work together to improve the human perception of the ground truth. In this case, in addition to understanding the AI, Θ should also include how the explanations lead human users to the ground truth. Providing explanation for teaching the ground truth returns Bayesian Teaching to its root in pedagogical modelling and shows how XAI and pedagogy are closely related.

4.2 Validation of Explanation

The medium of explanation can be evaluated partly independently from all other components of Bayesian Teaching. The reason for this partial independence is that explanation media can be evaluated along a dimension that is task-independent: ease-of-processing. It is self-evident that all else being equal, an explanation that is easy for humans to process should be preferred. Ease-of-processing can be decomposed into two primary elements: alignment with the human cognitive-perceptual system, and complexity. Different representations of the same underlying information can vary in human ease-of-processing. For example, humans can interpret probabilistic information more accurately when it is presented as natural frequencies rather than conditional probabilities [38], especially if the natural frequencies are visualized as icon arrays [39, 40]. Research on data-visualisation and risk-perception has studied how to optimize information presentation for human understanding [41, 42, 43], and XAI researchers could benefit from implementing these lessons.

Whereas information should always be presented in a way that is maximally interpretable to humans, information complexity involves a trade-off. More-complex information comes at a processing cost, but complexity can also add value. The solution to this trade-off is constrained by the complexity of the target inference to be explained Θ and the processing capacity of the target users. Domain experts tend to have a greater tolerance for complexity in their area of expertise, possibly because they have developed strategies to chunk task-relevant information more efficiently [44, 45]. To summarize, the appropriate complexity of an explanation is determined by the complexity of Θ, as well as the capacity of the users, as such complexity need to be evaluated with reference to a specific Θ, on the target population for the intended XAI solution.

There have been some attempts to create unit-tests for explanation media without reference to a specific learner model [46, 12]. For example, Lage and colleagues evaluated explanation media that varied on multiple complexity metrics. Their evaluation consisted of three different tasks: prediction (predicting the AI system's decision), verification (determining whether the system made an accurate decision), and counter-factual analysis (determining whether changing a single input feature would alter the system's classification). They measured the response time, accuracy, and subjective satisfaction of human participants in each condition. This provides a strong template for how to evaluate explanation media, but their results are not easily interpretable because they did not account for the complexity trade-off discussed previously.

4.3 Validation of Learner Model

The accuracy of the learner model $P_L(\Theta|x)$ is essential for effective explanation in the Bayesian Teaching framework. In order to be effective the learner model should capture both the user's belief prior to exposure to explanation, as well as the inferential processes the user applies to update their beliefs given the explanation. Thus, both the prior belief $P_L(\Theta)$ (implicit in a Bayesian learner model) the posterior beliefs $P_L(\Theta|x)$ should be evaluated. Because of this focus on human belief, both the development and the optimization of learner models depend on lessons from cognitive science as well as computer science. How well the learner model aligns with actual users can be evaluated by the *fidelity* between the modelled response to a given explanation x and the user's actual response. This involves specifying a Θ, sampling x, computing $P_L(\Theta|x)$, and running a user study that measures the fidelity between the learner model and the actual intended users. When assessing the calibration of the learner model, it is important to cover a wide range of $P_L(\Theta|x)$ [15]. This implies evaluating both explanations that the learner model predict will improve user understanding and explanations expected to be detrimental. If the learner model accurately predicts the full range of user behavior in response to explanation, it can be considered to have passed the key test.

Different users vary in their prior beliefs, inferential biases, and goals. As such, it is often unrealistic to develop one general learner model that captures all users well. Instead, we propose to develop different learner models for different user classes, such as AI engineers and clinicians. The modular nature of the Bayesian Teaching framework often allows for varying the learner model $P_L(\Theta|x)$ to fit the current user, while keeping Θ and x constant.

Aside from providing targeted explanations, formal learner models can add value to explainable AI by encoding general human inferential biases. For example, recent evidence from human reinforcement learning suggests that people learn more from information that supports their existing beliefs, relative to information that contradicts them [47, 48]. A well-designed learner model should incorporate such biases, so that they can be accounted for and leveraged for effective explanation. We recommend a modular encoding of such biases that can easily interface with any learner model, treating it as a meta-model, so as to speed up the development cycle of new XAI solutions.

4.4 Validation of Teacher Model

The evaluation of the teaching process, $P_T(x|\Theta)$, by definition depends on the specified x, Θ, and probability P_T. Equation 1 shows that teaching is fully determined by Θ, x, the learner model given by $P_L(\Theta|x)$, Ω, and $P(x)$. Thus, validation of these components implies the validation of the teaching process. By virtue of Bayes' rule, a user task suitable for evaluating the learner model will also be suitable for evaluating the teaching process. The former should focus on covering a wide range of $P_L(\Theta|x)$, while the latter should ensure the achievement of a particular $P_L(\Theta|x)$.

Additionally, Bayesian Teaching defines the effectiveness of an explanation by the extent to which it shifts a user's belief towards a target inference. The belief-shifting framework puts an upper bound on explanation effectiveness: If the user holds the target belief prior to being exposed to the explanation, there is no way to measure a potential positive impact of the explanation. As such, it only makes sense to test XAI interventions when there is a misalignment between user beliefs about the target AI system and the ground truth.

4.5 Validation on Use Cases

The appropriate target inference Θ is determined by the use case the user wishes to solve. To give an example, suppose a doctor wishes to understand why an AI image classifier diagnosed a particular patient with cancer, the XAI system may want to expose the decision-boundary around that particular image. In this case, a target inference that focuses on the decision boundary would be more suitable than one that focuses on matching the modeled data distribution.

There can be multiple ways to design the Θ and x to address the same use case. The pros and cons of each design often need to be evaluated empirically with user studies. Returning to the previous example of understanding the decision boundary around a data point, one can also set up a Bayesian Teaching problem to show what variation of masks would change the class label on the test data point from one to the other [c.f., 30]. This design would correspond to setting Θ to be the different class labels (e.g., positive and negative) and x to be a set of masks. Yet another way to explain the decision boundary is to give example data points from both side of the decision boundary. In this case, Θ could be the predicted label of the test data point, x would be a set of data-label pairs, and Ω or $P(x)$ can be used to enforce data-label pairs from both classes [49].

Use cases are often arranged in a hierarchy, where the higher-level use case depends on the performance in the lower-level use cases. This hierarchy can lead to a hierarchy of nested Bayesian Teaching problems. Returning to our medical example, a chief radiologist at the same hospital might want to prioritize what images should be passed along to a radiologist and what images can be automatically classified by the AI system. Here, the target inference relate to the general decision-boundaries of the model (not a specific image), with a special focus on determining regions

where the decision-boundaries are poorly aligned with the ground truth (cases when the AI system are likely to make mistakes). One may tackle this use case by designing a two-level Bayesian Teaching problem, where the lower level problem aims to explain decision boundary around a data point as described above, and the higher level problem aims to curate the appropriate data points to cover the overall boundary. In the higher level problem, Θ can thus be the overall decision boundaries, x is a set of local decision boundaries, and the learner model describes how the local decision boundaries are stitched together.

5 Generalization

In Section 2 we showed that Bayesian Teaching provides an abstract template to think about XAI by highlighting that explanation is a kind of goal-directed communication between the XAI system and human. In Section 3 we showed how to break down existing XAI methods into components according to this abstract template. From the decomposition, two complementary threads emerge. The first is the validation of components. In Section 4 we argued that validation should be done and that *modular* testing of certain components is possible. The second thread is the generalization of the components by means of recombining them to form novel XAI methods, which we discuss below. The two threads are complementary in that validated components would promote the validity of the recombination, *even absent a holistic testing of the system*. However, because evaluation of individual components is rarely done, below we simply entertain the recombination and leave their validity as a research direction.

Table 1 shows the components from the decomposition via abstraction in Section 3. Using Table 1, we make some remarks about trends that surfaced from the decomposition, including novel methods that could be formed by recombination.

	Target inference Θ	Explanation x	Learner model $P_L(\Theta\|x)$	Human evaluation task
(i)	Latent variables of the target model (e.g., class means of PLDA) [14]	Images from training data [14, 16]	PLDA [14]	Predicting the target model's predictions [14, 16]
(ii)	Modeled data distribution of a particular class [16]	Mask over target image [33, 18]	Kernel function on DNN [16]	User-assigned salience [18]
(iii)	Predicted class labels on target data points [33]	Model parameters [17, 36]	DNN classifier [33]	Competency testing: which classifier generalizes better [17]
(iv)	Predictive distributions given by the target model on target data points [18, 36]	Features (e.g., regions of an image) [17]	Additive linear model [18, 17]	Feature debugging: identify harmful features [17]
(v)	The decision boundary of the target model in the neighborhood of target data points [17]		Soft decision tree [36]	Model anomaly detection: identify classifier irregularities [17]

Table 1: A list of the components obtained from existing XAI methods presented in [14, 16, 33, 18, 17, 36]. For brevity, the mapping of the learner model that maps x to Θ is listed, but the probability assignment is left out. In addition to the target inference Θ, explanaiton x, and learner model $P_L(\Theta|x)$ mentioned in Section 2, the human evaluation tasks used in these papers are listed. The teacher's selection $P_T(x|\Theta)$ is left out, because most of the methods investigated maximize that probability.

Explanation medium. Most XAI methods are distinguished or classified by the x component. We argue that a coherent theory that includes all necessary components is more desirable. The coherence and completeness of such a perspective not only clarifies the connections among XAI methods but also offer systematic guidance on how to manipulate and compose them effectively.

Teaching component. Most methods select explanations by maximizing $P_T(x|\Theta)$. In the likely case of uncertainty in learner model and complexity penalty, strictly maximizing can lead to consistent suboptimal explanation. In general, max selection is not robust when the modeled inference of the user does not fully match that of the actual user [50]. Probabilistic selection, though less convenient, can lead to better average performance in these cases, and naturally supports testing whether explanations that are predicted to be effective indeed are.

Θ and evaluation task. Not all Θ and evaluation task are equally well aligned. For example, the goal of explaining predicted class labels in Θ (iii) is very well aligned to the user task of predicting the target model's predictions in evaluation task (i). In contrast, the predictive label target in Θ (iii) and the selection of a competent model in evaluation task (iii) are not directly related. Such misalignment requires additional consideration: is it possible that a simple

modification of Θ could improve this alignment? For example, would it be better to modify Θ to be the *difference* between the two target models' predictive probability on some test data points? An alternative solution would be to select a curriculum of target data points that will help the user to figure out the more competent model. Such curricula can be fleshed out in another Bayesian Teaching problem, where the new x is a sequence of Θ, and the new Θ will explicitly specify measures of discrimination between the two models.

Recombination. If a Θ is on the level of the generalization behavior of the model in data space [e.g., Θ, in the Table (iii)–(v)], this Θ will work together with any x. This observation allows for the generation of new XAI methods from recombination. For example, one can find a soft decision tree [x, in the Table (iii)] to best match the decision boundary of a target model [Θ, in the Table (v)]. On the other hand, if Θ is on the level of a model's latent variable or parameters [e.g., Θ, in the Table (i)], both the x and the mapping of the learner model are limited to models with the same parametric form. In summary, this illustration of how Bayesian Teaching can decompose and recombine existing XAI methods is beginning to hint at how Bayesian Teaching can identify reusable, modular components akin to a software development process.

6 Conclusion

We have argued for the importance of the development of systematic abstractions, validation, and generalization for explainable AI. Whereas the current state of the art in the field depends on solutions that are problem-specific, rendering the study of XAI to be a series of unconnected engineering problems, we advocate for a systematic, scientific approach to XAI. We present a theoretical framework based on Bayesian Teaching, which unifies the human and machine aspects of the problem and is strongly supported by research in cognitive science. Bayesian Teaching introduces a collection of abstractions that facilitate systematic thinking about XAI, which allows for the decomposition of diverse prior approaches. The abstractions presented by Bayesian Teaching support systematic validation of components, a necessary aspect of modern software development, which ensures that component parts of our model behave as expected through component-wise user studies. We further argue that abstraction and validation together support generalization—the ability to recompose validated aspects of models into new XAI methods for rapid deployment on new tasks and domains. Through this systematic approach, Bayesian Teaching supports the a cumulative science of XAI that incorporates best practices of behavioral research and software development.

Acknowledgements

This material is based on research sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-17-2-0146 to P.S. and S.C.-H.Y. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

This work was also supported by DoD grant 72531RTREP, NSF MRI 1828528 to P.S. The methods described herein are covered under Provisional Application No. 62/774,976.

Conflict of Interest

The authors declare no conflict of interests.

References

[1] Terrence J Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48):30033–30038, 2020.

[2] John Armour, Richard Parnham, and Mari Sako. Unlocking the potential of ai for english law. *International Journal of the Legal Profession*, pages 1–19, 2020.

[3] Paras Malik Amisha, Monika Pathania, and Vyas Kumar Rathaur. Overview of artificial intelligence in medicine. *Journal of family medicine and primary care*, 8(7):2328, 2019.

[4] Sven Winkelhaus and Eric H Grosse. Logistics 4.0: a systematic review towards a new logistics system. *International Journal of Production Research*, 58(1):18–43, 2020.

[5] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. Explanation in ai and law: Past, present and future. *Artificial Intelligence*, page 103387, 2020.

[6] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[8] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou. Reviewing the need for explainable artificial intelligence (xai). *arXiv preprint arXiv:2012.01007*, 2020.

[9] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

[10] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[11] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.

[12] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*, 2018.

[13] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[14] Wai Keen Vong, Ravi B Sojitra, Anderson Reyes, Scott Cheng-Hsin Yang, and Patrick Shafto. Bayesian teaching of image categories. In *CogSci*, 2018.

[15] Scott Cheng-Hsin Yang, Wai Keen Vong, Ravi B. Sojitra, Tomas Folke, and Patrick Shafto. Mitigating belief projection in explainable artificial intelligence via bayesian teaching, 2021.

[16] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288, 2016.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, pages 1135–1144, 2016.

[18] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017.

[19] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[20] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 882–891, 2018.

[21] Baxter S Eaves Jr, April M Schweinhart, and Patrick Shafto. Tractable bayesian teaching. In *Big Data in Cognitive Science*, pages 74–99. Psychology Press, 2016.

[22] Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. Showing versus doing: Teaching by demonstration. In *Advances in neural information processing systems*, pages 3027–3035, 2016.

[23] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.

[24] Atsushi Kanehira and Tatsuya Harada. Learning to explain with complemental examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8603–8611, 2019.

[25] Been Kim, Cynthia Rudin, and Julie A Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pages 1952–1960, 2014.

[26] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[27] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.

[28] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

[29] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9291–9301, 2018.

[30] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451*, 2019.

[31] Scott Cheng-Hsin Yang and Patrick Shafto. Explainable artificial intelligence via bayesian teaching. *NIPS 2017 workshop on Teaching Machines, Robots, and Humans.*, 2017.

[32] Akanksha Atrey, Kaleigh Clary, and David Jensen. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv preprint arXiv:1912.05743*, 2019.

[33] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint*, 2018.

[34] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *arXiv preprint arXiv:2004.14545*, 2020.

[35] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[36] Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

[37] Michael Olan. Unit testing: test early, test often. *Journal of Computing Sciences in Colleges*, 19(2):319–328, 2003.

[38] Gerd Gigerenzer and Adrian Edwards. Simple tools for understanding risks: from innumeracy to insight. *Bmj*, 327(7417):741–744, 2003.

[39] Rocio Garcia-Retamero, Mirta Galesic, and Gerd Gigerenzer. Do icon arrays help reduce denominator neglect? *Medical Decision Making*, 30(6):672–684, 2010.

[40] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology*, 28(2):210, 2009.

[41] Alan M MacEachren, Anthony Robinson, Susan Hopper, Steven Gardner, Robert Murray, Mark Gahegan, and Elisabeth Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

[42] Ken Brodlie, Rodolfo Allendes Osorio, and Adriano Lopes. A review of uncertainty in data visualization. *Expanding the frontiers of visual analytics and visualization*, pages 81–109, 2012.

[43] David Spiegelhalter, Mike Pearson, and Ian Short. Visualizing uncertainty about the future. *science*, 333(6048):1393–1400, 2011.

[44] William G Chase and Herbert A Simon. Perception in chess. *Cognitive psychology*, 4(1):55–81, 1973.

[45] Fernand Gobet and Herbert A Simon. Expert chess memory: Revisiting the chunking hypothesis. *Memory*, 6(3):225–255, 1998.

[46] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

[47] Stefano Palminteri, Germain Lefebvre, Emma J Kilford, and Sarah-Jayne Blakemore. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology*, 13(8):e1005684, 2017.

[48] Tor Oreste Tarantola, Tomas Folke, Annika Boldt, Omar David Perez, and Benedetto De Martino. Confirmation bias optimizes reward learning. *bioRxiv*, 2021.

[49] Tomas Folke, Scott Cheng-Hsin Yang, Sean Anderson, and Patrick Shafto. Explainable ai for medical imaging: Explaining pneumothorax diagnoses with Bayesian Teaching. In *Proc. SPIE 11746, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III, 117462J*, Apr 2021.

[50] Pei Wang, Junqi Wang, Pushpi Paranamana, and Patrick Shafto. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, 33, 2020.