Knowledge-to-Information Translation Training (KITT): An Adaptive Approach to Explainable Artificial Intelligence

Robert Thomson[1(✉)] and Jordan Richard Schoenherr[1,2(✉)]

[1] Army Cyber Institute/Behavioral Science and Leadership Department, US Military Academy, West Point, USA
{robert.thomson, jordan.schoenherr}@westpoint.edu
[2] Department of Psychology, Institute for Data Science, Carleton University, Ottawa, Canada

Abstract. Modern *black-box* artificial intelligence algorithms are computationally powerful yet fallible in unpredictable ways. While much research has gone into developing techniques to interpret these algorithms, less have also integrated the requirement to understand the algorithm as a function of their training data. In addition, few have examined the human requirements for explainability, so these interpretations provide the right quantity and quality of information to each user. We argue that Explainable Artificial Intelligence (XAI) frameworks need to account the expertise and goals of the user in order to gain widespread adoptance. We describe the Knowledge-to-Information Translation Training (KITT) framework, an approach to XAI that considers a number of possible explanatory models that can be used to facilitate users' understanding of artificial intelligence. Following a review of algorithms, we provide a taxonomy of explanation types and outline how adaptive instructional systems can facilitate knowledge translation between developers and users. Finally, we describe limitations of our approach and paths for future research opportunities.

Keywords: Explainable AI · Knowledge translation · Adaptive instructional systems

"Be gracious… I think an explanation is well overdue." – Wilton Knight, Knight Rider

1 Introduction

Machine Learning and Artificial Intelligence algorithms (AIs) have revolutionized the speed and complexity at which humans (and our technologies) process information. The last 30 years has seen the rise of computers and increasing automation as complex technologies have become ubiquitous in modern society. With this explosion in computational power, many of the algorithms driving our technology have become so complex that we no longer understand how they process information and make decisions. This lack of interpretability leads to an inability to predict how such algorithms will behave as they approach the edges of their competencies, or even understand what

R. A. Sottilare and J. Schwarz (Eds.): HCII 2020, LNCS 12214, pp. 187–204, 2020.
https://doi.org/10.1007/978-3-030-50788-6_14

their competencies are [1]. As such, there are concerns whether the developers [2], let alone the average user, can reliably interpret the output of modern AIs. Without interpretability, it becomes challenging to trust AIs, especially when their outputs appear to reflect errors [3–5].

Compounding this is the requirement that many of these algorithms require millions of data points for training. These large datasets are themselves so complex that many implicit (or explicit) biases can go undetected [6, 7]. One of the most telling failures of automation was the widely reported case of Google Photos auto-tagging feature that had difficulty accurately classifying the faces of non-Caucasian persons [8, 9]. This was due in part to the fact that their training dataset comprised of substantially more photos of Caucasian faces than non-Caucasian [10]. Detecting and remediating bias in a dataset (i.e., *preprocessing* the data) is a human labor-intensive problem which is generally not well automated for all but the most structured data types (i.e., recently Google has developed AutoML to automatically preprocess text and video input).

Explainable artificial intelligence (XAI) attempts to provide explanations of the structure and operations of AIs in a form that humans can understand. However, explainability alone is not necessarily a solution [11]. In the present study, we consider 1) recent advances in XAI; 2) what kinds of explanations are intelligible to humans; and 3) how adaptive instructional systems (AIS) can be used to facilitate this process. We further differentiate between interpretability, explainability, and believability. We then present the Knowledge-to-Information Translation Training (KITT) framework that acts as an intermediary between humans and non-human learning systems.

2 Explainable AI

XAI has become a buzzword in the information science community. The DARPA XAI Program began in 2016 with a goal to develop novel techniques to create explainable algorithms or induce explainable models from current algorithms [12]. It is based on the assumption that the most computationally complex and opaque algorithms (*black box* algorithms) are the most powerful, and that there is currently a tradeoff between performance and interpretability [13, 14]. That is, most AIs exist on a continuum of interpretability, where the most powerful techniques (e.g., neural networks; reinforcement learning) are the least interpretable. This is increasingly problematic as these black box algorithms are currently supporting key decisions-making processes in the medical, financial, and criminal justice systems, all without being adequately validated. Crucially, there are instances of bias in these systems, such as the apparent racial discrimination in prison sentencing and parole recommendation for minorities accused of non-violent crimes [13, 15].

Interpretability has a number of potential benefits including increased reliability, resilience, and validity of AIs, as well as enhancing trust between AIs and human users. Reliable AIs are those that make unbiased decisions. Bias can be introduced from two sources: the nature of the learning algorithm, and the underlying training data. Assuming the algorithm is valid, interpretable models can aid in the identification of sources of implicit and explicit biases in training data. Similarly, interpretable algorithms let us find the source of bias, whether it be a function of the data or injected by

an adversary. By helping to find outlier data points, interpretable AIs increase the resiliency of algorithms. Valid AIs weigh evidence in the manner for which we intended. An example of an invalid AI would be an AI that overweighs the role of zip code (as a proxy for many socioeconomic factors) in parole decisions. Interpretable AIs make it possible to validate their underlying decision-making processes. Finally, humans are more likely to trust an AI's decisions when that AI can provide an explanation, especially when the AI had made an error. In fact, an informative explanation after detecting an error may increase trust in the AI, but only after sufficient trust has been earned [16].

What needs to be explained and how an AI's decisions should be explained are often left ill-defined, reflecting a *squishy* problem [17]. A common theme is to conflate interpretability with an explanation. That is, the ability of the AI to simply describe its decision-making process does not mean that it is intelligible, nor does it mean that such a description can be used to predict the future performance of the model. In fact, these underspecified descriptions can lead to developers having incorrect mental models of the AIs' decisions [18]. Consequently, the believability of an explanation in no way implies that it is valid.

A common requirement of AIs is the provision of why it failed to perform a task as expected. Multiple levels of description are generally available, yet researchers addressing this issue appear to frequently focus on incomplete explanations on only one possible level of analysis. For instance, a technique may break down an AI's decisions to a set of IF-THEN rules (e.g., 'IF probLEFT > probRIGHT THEN turn LEFT') and/or probabilities (e.g., 'I turned LEFT because it was 61% more likely to succeed'). This level of description is generally most relevant to AI developers with a focus on debugging the system. This is generally not the right level of analysis for a typical user, who respond best to causal explanations [5, 19–21].

Following a brief review of core features of AI, we consider a number of approaches to interpretability in XAI research (for a more in-depth review, see [22]).

2.1 What Needs to be Interpreted in Artificial Intelligence Systems

At their core, most AIs can be distilled into a clustering or categorization mechanism. Once provided with a set of inputs, processing occurs, resulting in the selection of an output or set of outputs. Interpretability is required at each of these three stages of processing: the inputs need to have a clear meaning, the processes need to explain how the inputs are transformed, and the outputs need to be meaningful units.

Early approaches to AI were predominantly symbolic and propositional (e.g., a set of IF-THEN rules operating over meaningful primitives). For instance, an early example of XAI, MYCIN, [23, 24] was developed in the context of medical consultation. Following earlier criterion developed for computerized systems, [25] the symbolic processing techniques was adopted "in order to be accepted by physicians, should be able to explain how and why a particular conclusion has been derived".

The challenge for symbolic and propositional systems is that their interpretability does not necessarily scale to more complex problems. As seen in, it can be readily intuited that all irises with petal length greater than 2.45 cm are setosas (i.e., IF

petalLength > 2.45 cm THEN setosa[1]). What if there were hundreds of nested decisions? In this case there would need to be some mechanism to distill the most relevant decisions to the user, such as a simplifying model of causation that describes the process. While the model would be considered interpretable since all decisions are available to the user, the model itself is not inherently explainable without additional processing. To be truly explainable, one needs to provide more than the set of primitive decisions. Instead, an explanation requires that a system distills the most relevant criterion that maximize predictability for the particular user's goals [26] (Fig. 1).

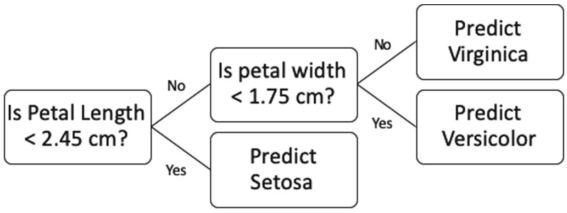

Fig. 1. A Decision Tree representation classifying Irises based on petal length and width. This example was derived from the *sklearn* Python package and visualized using the D3 tree visualization package.

The two approaches of the most productive AIs are *Reinforcement Learners* and *Neural Networks*. Reinforcement learners follow a common theme: there are a set of input states, and a reward associated with every possible action given those states. The *policy* of the learner is the set of learned actions for a given state based on maximizing the reward. In mathematical terms the policy is a map providing the probability of responding when in a given state. Rewards can be immediate or discounted, i.e., based on distal actions. Generally, the space is too vast to capture the set of all possible states and actions. Consequently, there is some sampling or exploration of the space to capture a reasonable set of actions. Similar to the issue of Decision-Trees, classical reinforcement learners are technically interpretable in that each decision is expressed by the policy, but not necessarily explainable at a higher level (e.g., finding hierarchical patterns in states).

Neural networks are loosely inspired by the firing of neurons within the brain. *Deep* neural networks, utilize mathematical inputs (i.e., the input layer vector) which are then processed through distributed layers of 'hidden' states to active the output layer, with these hidden states being updated by some learning signal (see Fig. 2). Supervised networks are trained on labeled data (data with a ground-truth answer) and are generally used for prediction tasks, while unsupervised networks generally cluster data which has no 'correct answer' and are used to cluster data to find potentially meaningful relations. Most neural networks are very data hungry, so finding good labeled

[1] Decision trees get more complex when the decisions are probabilistic and become less explainable even though they are still technically interpretable.

data can be a challenge. There are also semi-supervised techniques which utilize substantially less training data. These networks use the limited training data to guide the initial clustering for the larger amount of unlabeled data. The challenge is that these hidden layers are not generally interpretable, take substantial training data, and have hundreds to thousands of manipulable parameters [27].

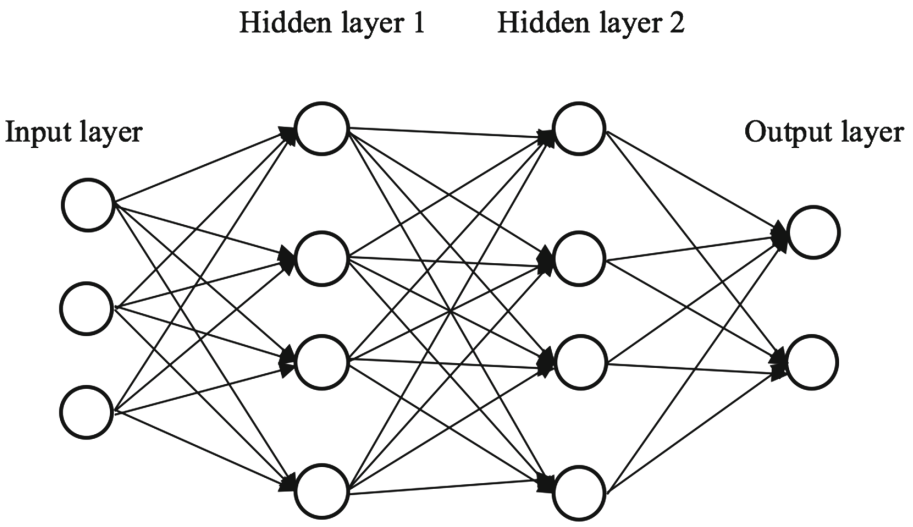

Fig. 2. Deep Neural Network. This example shows an input and output layer as well as two hidden layers. Layers are composed of a set of nodes, each with its own set of connections.

2.2 Approaches to XAI: Models of Interpretability

While there are hundreds of techniques being developed to improve the interpretability of AIs, there are three key approaches which we will discuss: *attention filters*, *model induction*, and *exemplar presentation.*

Attention Filters

In a comparable manner to theories of human attentional processing, the focus of attentional filter techniques is to highlight a subset of features that have made the most significant contributions to decision or classification process. In images, this typically represents a selection of 'salient' pixels within an image as a sort of heat map (e.g., see Fig. 3). In addition to heatmaps, image data has also been explained by captioning the image with automatic labelling of objects and/or actions, and by producing hierarchical description graphs representing some relationships between the objects and/or actions within the image [26, 28]. While heatmaps provide some information about the underlying focus of an algorithm, they are not interpretable in and of themselves. Instead, they rely on humans to interpret whether the focus is on the relevant labeled object or a potentially spurious correlated feature (e.g., incorrectly *perceiving* a person in a riding outfit sitting on a chair as that person riding a horse due to their outfit being

the most salient feature of the decision). As such, the user is the one generating the explanation (Fig. 3).

Fig. 3. Heatmap of a horse from the HMB-51 dataset using a convolutional neural network

Outside of image data, salient features have been intuited by visualizing salient clusters of high-dimensional representations using t-distributed stochastic neighbor embedding (t-SNE), which is a non-linear dimension-reduction technique. This has been used to discover missing competencies in algorithms in game playing algorithms [29, 30]. These techniques seek to interpret the high-dimensional space of the algorithms' *cognition*[2] in a fashion analogous to using neuroimaging (e.g., fMRI) to intuit human cognitive processes.

Model Induction

Another approach to interpretability reflects an examination of model induction processes. The primary goal of model induction is to train a separate interpretable AI to make the same decisions as the non-interpretable AI, to some level of abstraction [27, 30–32]. These interpretable models include simple linear models, rule lists [33, 34], or decision trees (although see [29] for the use of cognitive models to approximate AI

[2] For the sake of analogy, we will refer to the black-box processes and output, such as the hidden layers in a neural network leading to the output layer, as an algorithm's cognition.

decisions). Perhaps the most well-known is LIME (Local Interpretable Model-agnostic Explanations), which samples the decision/classification space of a non-interpretable model and creates an interpretable local representation that approximates the performance of the non-interpretable model [31]. In plainer terms, LIME focuses on interpreting a given prediction as opposed to a perfectly interpretable *global* model. In the case of reinforcement learning, it is possible to decompose the reward signal to better understand the most salient features that impacted the reward by sampling the decision space [35].

The focus of many model induction methods is not to make models inherently more explainable in and of themselves, but instead to make their decisions interpretable to some level of fidelity without guaranteeing that the interpretable model captures any of the actual cognitive processes of the non-interpretable model.

Exemplar Presentation

Similar to attentional techniques, exemplar presentation techniques such as Bayesian Teaching [36] present a subset of data is provided to the user which represent the prototypical exemplar of a given decision. Thus, to understand why a given image was classified as a bird, several training examples most representative of the bird decision are presented to the user. The user should then be able to intuit some of the underlying cognitive processes of the model. This technique is based on being analogous to how children learn in an education setting [37]. At no point is any cognition of the underlying AI model directly interpreted, it is up to the user the intuit the cognition by finding commonalities within the examples.

2.3 Challenges for Existing Approaches

Most existing approaches to AI are not attempting to directly make more interpretable models, instead, they focus on either developing some approximate models which are more interpretable, or they try to expose some salient features which are being used by the model in some capacity, then have the user induce the underlying cognitive processes of the model. Each style has several overlapping limitations. Perhaps the most telling is scale: developing interpretable models is computationally intractable. Each technique generally uses some kind of subsampling or approximation in order to make either the decision-space more understandable (in the case of t-SNE dimensionality reduction), or to simply reduce the computational complexity of the interpretable model so it may be trained in a reasonable amount of time.

Any time one makes an approximation, some amount of information is generally lost. In the case of very complex decision-spaces, it becomes more likely that significant parts of the decision-space will not be explored, and resultingly the interpretable model will be incomplete. Consequently, any attempts to generalize from the current decision might result in an error. In fact, there is evidence that providing limited explanations causes users to overly trust their model and make incorrect predictions of their model's future behavior [18]. Furthermore, attentional and exemplar techniques rely on the user to generate their own mental model of the AI model's cognitive processes. This projects the user's human biases into their perception of the AI model.

Finally, model induction techniques tend to focus on only one level of interpretability, so the provided explanation may not in fact be a proper explanation, and may not be presented at the right level based on the user's goals and expertise. For

instance, imagine a very complex model running a nuclear power plant. It is possible to break down this model into a very complex decision-tree consisting of hundreds decisions, however, if the AI initiates an emergency shutdown the kind of explanation might be different based on whether it is the developer, manager, or engineer using the system. To date, most development has focused on the software developer so they may be able to better debug the system [15].

3 Explanations

Bridging the gap between developers who focus on interpretability and users requiring explainability necessitates a consideration of what makes an explanation [38]. In general, explanations consist of two parts: an explanandum (a phenomenon that needs to be explained) and the explanans (the statements purporting to explain the phenomenon). In the context of XAI, explanandum can consist of the algorithm (i.e., an input-output mapping), or any given subset of processes that define an algorithm's function. We will demonstrate that the requirements of explainability are relative to the recipient of the explanation.

3.1 Model Criterion: Interpretability, Explainability, and Believability

Information and computer scientists have considered a number of types of explanations [20, 26, 39–41]. For instance, Marr [39] suggested three independent levels of explanation for understanding systems: the computation level, the algorithmic and representational level, and the implementation level. Importantly, in proposing this distinction, Marr was neither concerned with knowledge translation between information and computer scientists and the general public, nor was he interested in understanding how individuals explain human behaviors to themselves and others, i.e. folk psychology. In contrast to Marr, Samek and Müller [26] consider explanation in terms of recipient, information content, and intentions of the explanation (questions that are answers and how explanation is used). This approach emphasizes the relative nature of explanation.

Doran, Shultz, & Besold [42] provide another approach to classifying XAI systems. They identified two distinct approaches to explaining black box systems (or, *opaque systems*) where input-output mappings are not accessible to the user. *Interpretable systems* reflect systems wherein more technical details are provided to users for them to understand the operations of a system. In their example, Doran et al. contrast regression models wherein outputs and variable weights can be contrasted with deep neural networks that are unlikely to be interpretable. In contrast, *comprehensible systems* symbols such as linguistic markers or images are presented along with the output. They note that users must rely on their "own implicit knowledge and reasoning about them". Doran et al. note that interpretable and comprehensible systems reflect improvements relative to the explainability of opaque systems.

Earlier work by Lipton [1] also highlights the benefits of interpretability. In the context of AI, Lipton notes that "the task of interpretation appears underspecified … [with the] demand for interpretability [arising] when there is a mismatch between the

formal objectives of supervised learning (test set predictive performance) and the real-world costs in a deployment setting." Reviewing the literature, Lipton suggests that there are four criteria for interpretability: trust, causality, transferability, informativeness, and fair and ethical decision-making. Each of these criteria, however, is considerably variability in terms of how they are defined and, indeed, interpreted. In this way, these criteria are often externally defined relative to the system due to features of the social environment [19]. Thus, considering the possible explanation that users have access to can help understand their current mental model and promote the development of a more sophisticated understanding of these systems [43].

3.2 Typology of Explanations

Users' needs and goals are crucial to developing effective XAI. Earlier philosophical taxonomies provide insight into what types of explanations are possible. For instance, Aristotle was the first to attempt to systematically differentiate kinds of explanations (i.e., material, formal, efficient cause, and final cause). In that early philosophers relied on their own experience and observations, their formulations of explanation might reflect formalized human intuitions (i.e., folk theories). More recently, Dennett [44] provided a framework that assumes three broad approaches to explanation: the physical stance, design stance, and intentional stance. These stances concern causal/mechanical explanations, functional explanations, and intentional explanations, respectively. For Dennett, explanations are used in order to *predict* the behavior of biological and nonbiological systems through analogical reasoning. Adopting a similar framework, we define three primary explanation types below as well as provide examples of other specific explanation subtypes that demonstrates the inclusivity of this framework.[3]

Causal/Mechanical. The broadest class of explanation is the causal explanation. A casual explanation consists of an appeal to an underlying causal relationship between objects, entities, forces, and events. They take the form X caused Y to do A. For instance, a drone did not find the target after System X failed to classify a target object. Most explanations offered in academic disciplines in the humanities, physical and social science likely fall into this category.

Historical. Historical explanations appeal to past events as causing an event or an entity to engage in an action, e.g., Event X resulted in Event Y. For instance, an algorithm produced a racial bias in an output variable because it was trained on a biased datasets.

Reductive. A reductive explanation appeals to a lower-order property of a system to explain a higher-order property, e.g., Feature A resulted in Action Y. For instance, the autonomous vehicle hit another car because its proximity sensory was not working. In the case of reductive explanations, the explainer selects a subset of features from a total set, ideally, those that are most strongly associated with an outcome.

[3] We acknowledge that there is potentially overlap between some explanation types.

Functional/Teleological. A second class of explanation is the functional explanation. Functional explanations appeal to the function, end, or purpose of objects, entities, and events, e.g., X did Y because of Function A. For instance, a surveillance camera feed is flagged because it detected that an individual was potentially carrying a weapon. Here, the function (detect) qualifies why a specific video feed was brought to the user's attention. Functional explanations reflect a more general form of explanation relative to causal or mechanical. Namely, in the detection example, the actual data processing of the detection algorithm is *not* provided as part of the explanation.

Formal or Mathematical. Formal or mathematical explanations can be understood as special cases of functional explanations. Formal explanation appeal to formal principles embedded within an extant logical framework, e.g., X was observed because of Y, where Y is a defined principle. For instance, the program failed to compile because there was an illegal command on Line 5.

Subsumption. Subsumptive explanations appeal to an ontological category of an objects, entities, and events, e.g., X has A because it is a member of Category Y. For instance, a developer noted that an app worked perfectly because it was optimized for the latest release of the Android operating system. In this case, a specific kind of app worked because it was running on a specific kind of operating system.

Macro-to-Micro. A macro-to-micro explanation reflects the application of properties from a macro-level phenomenon to a micro-level phenomenon, e.g., an application failed to work because of a specific line of code that was inserted into the script, like a computer failing to boot because it had the wrong drivers installed. Here, analogical reasoning is used to understand a micro level phenomenon (i.e., insertion of a code) by means of a macro-level phenomenon (i.e., available drivers).

Intentional. The third class of explanation is the intentional explanation. These explanations appeal to the mental states and motivations of an autonomous agent, e.g., X did Y because it wanted or needed A. For instance, an autonomous vehicle swerved away from the patch of icy because it *wanted* to avoid a crash. Here, an analogy is provided based on the user's knowledge of a known domain (i.e., folk psychology) to help understand an unknown domain (i.e., system operations). Intentional explanations reflect the most simplified form of explanation in that they ignore specific causal mechanisms (e.g., software and hardware) as well as the overall function (i.e., what the system was designed to do).

Anthropomorphic Explanations. Often used synonymously with intentional explanations, anthropomorphic explanations emphasize the human-like qualities of a non-human entity. In addition to intentional states, anthropomorphic explanations tend to include affective responses (e.g., happy, sad, mad), behaviours (e.g., 'acting up', 'stubbornness'), or human social bonds (i.e., they're friends, they like one another). For instance, two systems might not like talking one another.

Meta-Explanation. In general, meta-explanations reflect a kind of explanation that is used to understand the failure of lower-order explanations. Meta-explanations appeal to the structure of a scenario, including communication and argumentation between agents, e.g., X believes A while Y believes B, leading them to misunderstand

in situations like W. For instance, the team thought that a drone failed to drop its payload because of a software issue. When the team examined the code, they instead realized it was a hardware issue. In that humans are involved in meta-explanations, they frequently reflect a special case of intentional explanations that reflect collective or conjoint intentionality. In this way, XAI might be seen as an effort in meta-explanation, e.g., users do not understand systems failures because they do not understand the operations of the underlying code.

3.3 The Problem of XAI

Having defined the kinds of algorithms the clustering or categorization mechanism that define AI and the explanations types that are intelligible to model developers and users, the problem of XAI becomes one of knowledge translation. Throughout this process, information will be lost, but the essential features and functions of algorithms must be maintained. Two paths are possible: knowledge-to-information translation and information-to-knowledge translation.

Knowledge-to-Information Translation. Knowledge-to-information translation reflects an implementation problem: a developer wants to have a machine perform a given task (knowledge) and they therefore need to create a code that can be used by a system to perform a task given a set of constraints (information). This is the problem faced in information and computer science training and education: developers have a goal-state in mind but need to determine what code satisfies those operations for any given set of functions. This likely reflects the approach to XAI inspired by Marr [39], wherein computational knowledge is translated into an algorithm, or an algorithm is implemented using software and hardware.

When clients or users are considered, the nature of an effective explanation is unclear. Although the client has specific needs (e.g., for a website or software that performs a specific function), they need to communicate this to the developer. Similarly, an application user might want to have software perform a specific task but does not know what command will produce the desired result. In these cases, a client or user without knowledge of information sciences likely has an intentional model of what they want a problem to do, but they do not understand the underlying formal principles. Moreover, if they believe that a system has operated successfully, they will not question the results.

Information-to-Knowledge Translation. In contrast to knowledge-to-information translation, XAI reflects an information-to-knowledge translation problem. Namely, when developers or users are presented with an error, they likely want to understand *why* the output had those properties. However, they might not require or want a comprehensive causal explanation of the phenomenon which would likely require more knowledge of information science. Instead, the user wants to be able to understand the output in a manner that is compatible with their existing knowledge. The "essential" features or functions that can account for satisfaction or violation of goals. For instance, policymakers likely wish to understand how an algorithm functions in order to ensure that

the results it provides are not biased (e.g., discriminating based on social categories or socioeconomic status). In this way, the provision of an explanation is part of a learning process.

Finally, explanations are inherently reductive. Explanations are offered as a means to understand a system by simplifying and systematizing its operations. Human information processing follows a similar path. Research suggests that there are shifts from implicit to explicit representations [63] and that conscious processing results in progressively more abstract, simplified representations [64, 65]. By translating information to knowledge, we provide an approach which parallels human cognition.

4 Knowledge-to-Information Translation Training (KITT)

In order to train learners to understand AI, we must account for 1) the objectives of explainability and XAI, 2) what constitutes a good explanation, and 3) the process of learning and communicating information such that it relatable to a learner's knowledge. We refer to this approach as knowledge-to-information translation training, or KITT. Central to this approach is the distinction between interpretability (understanding the output and operations of a system), explainability (using analogical reasoning to predict one system by means of knowledge of another), and believability (acceptance of an analogy but without understanding its basis).

When implemented, the KITT framework assumes that training and learning is mediated through an intelligent tutoring system (ITS; e.g., [45–48]) which can use multiple criteria to monitor and assess expert and non-expert performance. KITT can be implemented in a number of ways, e.g., a standalone training AIS or integrated into an AI used for research, data processing, or consultation. The KITT framework assumes that learners who simply wish to understand an AI would prefer simple explanations (e.g., intentional explanations) but over time, learners will require more in-depth knowledge of the operations of a system. This reflects an adaptive learning process.

4.1 Explanatory Scaffolding Process

Interpretability. Learning has been defined as social scaffolding, wherein an educator assesses a learner's current mental representation of a problem space and identifies what is both relevant and novel to a learner [49, 50]. AISs can address this learning process [43]. In order to accomplish this, the KITT framework assumes that we must first assess the learner's current state of knowledge within the domain of information science (i.e., declarative and procedural knowledge pertaining to the modelling approach being considered) and then provide additional information to extend this knowledge. Unlike other approaches to XAI, KITT assumes that the kind of explanation provided can be quite variable depending on what kind of predictions and level of understanding is desired.

For information science, knowledge-to-information translation will consist of systematically learning core curriculum areas (e.g., programming languages, strategies for

debugging, system operations). In these cases, the focus will likely be on causal/mechanistic explanations. Programmers will be taught principles of causation based on the functions of programming languages and the constraints of hardware, i.e., command X results in operation Y, or a PC will crash if program X runs due to insufficient RAM. When debugging a program, they will focus on historical explanations and reductive explanations, i.e., the program failed to run a subroutine due to a failure to close a loop in the code. This makes the learning process comparable to learning neuroscientific principles (e.g., structure of neurons, cell assemblies, fiber tracts).

Explainability and Believability. Developing a bottom-up understanding of programs and systems is not always possible or desirable. For instance, in the case of a typical user, clinician, or policy analyst that wants to verify that an approach is valid, deep descriptive knowledge is not required. Making a similar point, Marr [39] notes that reductive approaches to explanation like "[n]europhysiology and psychophysics have as their business to describe the behavior of cells or of subjects, not to explain such behavior" (p. 15). The KITT framework assumes that interpreting information and translating it into knowledge is facilitated through a process of analogical reasoning using relationships found in known mechanical, functional, and intentional systems. However, in contrast to the relatively simple nature of scaffolding required for developers, KITT requires the selection of an explanation that will be intelligible to the user. It must therefore highlight functional similarities between the operations of an AI and a familiar domain of explanation.

The goals of XAI parallel explanation in psychological science. Namely, psychological phenomena reflect the output of the interactions of neurons. A straightforward means to understand the information-to-knowledge translation processes is to consider how explanations of psychological phenomena are presented to non-experts. Studies of explanations of psychological phenomena have suggest that the addition of mechanistic relationships (e.g., the use of irrelevant neuroscientific evidence in explaining psychological phenomena) can make an explanation more believable [51, 52]. However, Schoenherr, Thomson, & Davies [53] replicated this effect using valid and invalid general mechanistic explanations (i.e., X was drawn to Y by a force) suggesting that neuroscientific explanations themselves might not be the basis for this effect. Instead, mechanistic explanations might appear to be more believable regardless of their validity.

The believability of an explanation does not imply that it facilitates learning. Science educators have considered the utility of certain kinds of explanations [54–58]. For instance, in the context of biology and evolutionary theory, teleological explanations are frequently used, e.g., a feature has adapted *for* survival in an environment [59, 60]. Moreover, evidence suggests that certain kinds of explanations can be used to scaffold learners' understanding. For instance, Tamir and Zohar [57] found that while 10th and 12th graders accepted anthropomorphic formulations (82%), they did not necessarily believe that plants (29%) or animals (62%) actually had intentionality. More recently, studies have replicated these findings noting discrepancies between factors that predict learning and their acceptance and use of these explanations and their associated terms [61, 62]. Thus, following Dennett's [44] proposition, intentional explanations might be

more intelligible to learners while learners themselves simultaneously understand the analogical basis for these statements.

In summary, while mechanistic explanations might be more believable, intentional explanations likely provide a more principled means to start a learner's training in understanding AI. Learners might be capable of acquiring superior functional knowledge concerning the operations of a system when framed in intentional turns. Namely, when a non-expert first understands what a system is trying to do (i.e., its intended function), the system can be decomposed into separable, functional units. This is analogous to a student studying cognitive psychology who learners about memory, attention, and decision-making. In the case of some AIs (e.g., deep neural nets), this functional division might not fall neatly along these lines. Rather, functions might reflect a larger number of specific, heterogeneous processes relative to human cognition, e.g., feature identifiers for eyes, image size, luminosity, brow shape. Moreover, different approaches to AI might be more inherently explainable than others, e.g., symbolic and propositional systems relative to deep neural nets. In some cases, functions might be unspecifiable in human terms.

5 Conclusion and Caveats

XAI reflects an approach to AI that allows experts and non-experts to understand and predict the output and operations of artificial systems. Rather than assuming that a single kind or level of analysis is appropriate, we suggest that an effective solution to XAI requires a consideration of the learner's knowledge [19] and that this knowledge and the goals of the user will changes over time. This is not reflected in other frameworks like Marr's [39] that is likely more appropriate when providing explanations to those in information science. Instead, the cognitive science of explanation needs to be used to inform those processes [38]. We have provided a taxonomy based on Dennett [44] that suggest three levels of explanation: causal/mechanistic, functional/teleological, and intentional.

Specifically, while an AI might be interpretable to experts, this does not imply that operations of the systems are accurately understood. Instead, explainability requires the ability to relate information in one domain to another. Simplified explanations can be based on basic features of human intentionality (i.e., wants, desires, goals) or more specific analogies based on human information processing systems (e.g., attention, memory, categorization). The selection of the level that is desirable will depend on the knowledge and goals of the user as well as the nature of the AI.

Knowledge-to-information and information-to-knowledge translation need not be exclusive paths. In a comparable manner to cognitive neuroscience, bottom-up and top-down reasoning can be developed conjointly. Namely, users can have both an ability to interpret a system at a low-level while also having high-level explanations of the function of a system. This process reflects *abductive reasoning* wherein multiple, possible explanations are plausible based on the kinds of explanations that can provide predictive models. In that XAI reflects a squishy problem, we assume that satisficing criteria should be used to judge the abilities of learners wherein a learner's knowledge is assessed in a relative, rather than absolute, manner [43].

Finally, we have described the use of the KITT framework in a manner that is amenable to an AIS. However, like all learning tasks, KITT can be adapted to an empirical approach to do research in XAI. Namely, the accuracy and retention of information concerning an AI being assessed by developers can be used to determine 1) the complexity of explanation that provides a high level of predictive accuracy, as well as 2) the kinds of explanations that are facilitate learning and believability.

Acknowledgements. Research was sponsored by the Army Research Laboratory and was accomplished under the Cooperative Agreement Number W911NF-19-2-0223. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for the Government purposed notwithstanding any copyright notation herein.

References

1. Lipton, Z.C.: The mythos of model interpretability. In: ICML Workshop on Human Interpretability in Machine Learning, New York (2016)
2. Bhatt, U., et al.: Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability and Transparency, pp. 648–657 (2020)
3. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Int. J. Hum. Comput. Stud. **58**(6), 697–718 (2003)
4. Andras, P., et al.: Trusting intelligent machines: deepening trust within socio-technical systems. IEEE Technol. Soc. Mag. **37**(4), 76–83 (2018)
5. Rossi, F.: Building trust in artificial intelligence. J. Int. Aff. **72**(1), 127–134 (2019)
6. Caliskan, A.B.J., Narayanan, A.: Semantic derived automatically from language corpora contain human-like biases. Science **6334**(356), 183–186 (2017)
7. Zou, J., Schiebinger, L.: AI can be sexist and racist - it's time to make it fair. Nat. Comments **559**, 324–326 (2018)
8. BCC: Google apologises for photos app's racist blunder. BBC (2015). https://www.bbc.com/news/technology-33347866. Accessed 15 Dec 2019
9. Kasperkevic, J.: Google says sorry for racist auto-tag in photo app. The Guardian (2015). https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app. Accessed 14 Dec 2019
10. Hern, A.: Google's solution to accidental algorithmic racism: ban gorillas. The Guardian (2018). https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people. Accessed 15 Dec 2019
11. Edwards, L., Veale, M.: Slave to the algorithm: why a right to an explanation is probably not the remedy you are looking for. Duke Law Technol. Rev. **16**, 18–84 (2017)
12. Gunning, D.: DARPA XAI BAA. DARPA (2016). https://www.darpa.mil/attachments/DARPA-BAA-16–53.pdf. Accessed 20 Feb 2020
13. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions, and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
14. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
15. Deeks, A.: The judicial demand for explainable artificial intelligence. Columbia Law Rev. **119**(7), 1829–1850 (2019)

16. Yin, M., Wortman, V., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019)
17. Straunch, R.: Squishy problems and quantitative method. Policy Sci. **6**, 175–184 (1975)
18. Lakkaraju, H., Bastani, O.: "How do I fool you?": manipulating user trust via misleading black box explanations. In: Proceedings of AAAI/ACM Conference on AI, Ethics, and Society (2020)
19. Miller, T.: Artif. Intell. **267**, 1–38 (2019)
20. Hoffman, R., Klein, G., Mueller, S.: Explaining explanation for "Explainable AI". In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, pp. 197–201 (2018)
21. Gilpin, L., Bau, D., Yuan, B., Baiwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of IEEE 5th International Conference on Data Science and Advanced Analytics, pp. 80–89 (2018)
22. Došilović, F., Brčić, M., Hlupić, N.: Explainable artificial intelligence: a survey. In: Proceedings of 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 210–215 (2018)
23. Fagan, L.M., Shortliffe, E.H., Buchanan, B.G.: Computer-based medical decision making: from MYCIN to VM. Automedica **3**, 97–108 (1980)
24. Shortliffe, E.H.: Computer-Based Medical Consultations: MYCIN. Elsevier/North Holland, New York (1976)
25. Gorry, G.A.: Computer-assisted clinical decision making. Methods Inf. Med. **12**, 45–51 (1973)
26. Samek, W., Müller, K.-R.: Towards explainable artificial intelligence. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700, pp. 5–22. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_1
27. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
28. Aditya, S.: Explainable image understanding using vision and reasoning. In: Proceedings of Thirty-First AAAI Conference on Artificial Intelligence (2017)
29. Somers, S., Mtisupoulos, C., Lebiere, C., Thomson, R.: Explaining the decisions of a deep reinforcement learners with a cognitive architecture. In: Proceedings of International Conference on Cognitive Modeling (2018)
30. Somers, S., Mitsopoulos, K., Lebiere, C., Thomson, R.: Cognitive-level salience for explainable artificial intelligence. In: Proceedings of International Conference on Cognitive Modeling, Montreal (2019)
31. Ribeiro, M., Singh, S., Guestrin, C.: "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2016)
32. Ras, G., van Gerven, M., Haselager, P.: Explanation methods in deep learning: users, values, concerns and challenges. In: Escalante, H.J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., van Gerven, M. (eds.) Explainable and Interpretable Models in Computer Vision and Machine Learning. TSSCML, pp. 19–36. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4_2
33. Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., MacNeille, P.: A Bayesian framework for learning rule sets for interpretable classification. J. Mach. Learn. Res. **70**(18), 1–37 (2017)
34. Keneni, B., et al.: Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. IEEE Access **7**, 17001–17016 (2019)

35. Erwig, M., Fern, A., Murali, M., Koul, A.: Explaining deep adaptive programs via reward decomposition. In: Proceedings of International Joint Conference on Artificial Intelligence - Working on Explainable Artificial Intelligence (2018)
36. Yang, S., Shafto, P.: Explainable artificial intelligence via Bayesian teaching. In: Proceedings of 31st Conference on Neural Information Processing Systems, Long Beach (2017)
37. Shafto, P., Goodman, N., Griffiths, T.: A rational account of pedagogical reasoning: teaching by, and learning from, examples. Cogn. Psychol. **71**, 55–89 (2014)
38. Keil, F.C., Wilson, R.A., Wilson, R.A.: Explanation and Cognition. MIT Press, Cambridge (2000)
39. Marr, D.: Vision: A Computational Approach. Freeman & Co., San Francisco (1982)
40. Biran, O., Cotton, C.: Explanation and justification in machine learning: a survey. In: Proceedings of IJCAI-2017 Workshop on Explainable Artificial Intelligence (XAI) (2017)
41. Park, D.H., Hendricks, L.A., Akata, Z., Schiele, B., Darrell, T., Rohrbach, M.: Attentive explanations: justifying decisions and pointing to the evidence. arXiv preprint arXiv:1612.04757 (2016)
42. Doran, D., Schulz, S. Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 (2017)
43. Schoenherr, J.R.: Adapting the zone of proximal development to the wicked environments of professional practice. In: Proceedings of HCII 2020, Copenhagen, HCI International (2020)
44. Dennett, D.: The Intentional Stance. MIT Press, Cambridge (1987)
45. Anderson, J.R., Gluck, K.: What role do cognitive architectures play in intelligent tutoring systems? In: Klahr, V., Carver, S.M. (eds.) Cognition Instruction: Twenty-Five Years Progress, pp. 227–262. Lawrence Erlbaum Associates, Mahwah (2001)
46. Nwana, H.S.: Intelligent tutoring systems: an overview. Artif. Intell. Rev. **4**, 251–277 (1990)
47. Ohlsson, S.: Some principles of intelligent tutoring. Instr. Sci. **14**, 293–326 (1986)
48. Polson, M.C., Richardson, J.J.: Foundations of Intelligent Tutoring Systems. Psychology Press (2013)
49. Vygotsky, L.S.: Thought and Language. MIT Press, Cambridge (1934/1986)
50. Vygotsky, L.S.: Mind in Society: The Development of Higher Mental Processes. Harvard University Press, Cambridge (1930–1934/1978)
51. Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E., Gray, J.R.: The seductive allure of neuroscience explanations. J. Cogn. Neurosci. **20**, 470–477 (2008)
52. Rhodes, R.E., Rodriguez, F., Shah, P.: Explaining the alluring influence of neuroscience information on scientific reasoning. J. Exp. Psychol. Learn. Mem. Cogn. **40**, 1432–1440 (2014)
53. Schoenherr, J.R., Thomson, R., Davies, J.: What makes an explanation believable? Mechanistic and anthropomorphic explanations of natural phenomena. In: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society. Cognitive Science Society, Boston (2011)
54. Bartov, H.: Teaching students to understand the advantages and disadvantages of teleological and anthropomorphic statements in biology. J. Res. Sci. Teach. **18**, 79–86 (1981)
55. Talanquer, V.: Explanations and teleology in chemistry education. Int. J. Sci. Educ. **29**, 853–870 (2007)
56. Talanquer, V.: Exploring dominant types of explanations built by general chemistry students. Int. J. Sci. Educ. **32**, 2393–2412 (2010)
57. Tamir, P., Zohar, A.: Anthropomorphism and teleology in reasoning about biological phenomena. Sci. Educ. **75**, 57–67 (1991)

58. Zohar, A., Ginossar, S.: Lifting the taboo regarding teleology and anthropomorphism in biology education—heretical suggestions. Sci. Educ. **82**, 679–697 (1998)
59. Bardapurkar, A.: Do students see the selection in organic evolution? A critical review of the causal structure of student explanations. Evol. Educ. Outreach **1**(3), 299–305 (2008)
60. Ziegler, D.: The question of purpose. Evol. Educ. Outreach **1**, 44–45 (2008)
61. Barnes, M.E., et al.: Teleological reasoning, not acceptance of evolution, impacts students' ability to learn natural selection. Evol. Educ. Outreach **10**(1), 7 (2017)
62. Thulin, S., Pramling, N.: Anthropomorphically speaking: on communication between teachers and children in early childhood biology education. Int. J. Early Years Educ. **17**, 137–150 (2009)
63. Karmiloff-Smith, A.: Beyond Modularity. MIT Press/Bradford Books, Cambridge (1992)
64. Zeki, S.: The disunity of consciousness. Trends Cogn. Sci. **7**, 214–218 (2003)
65. Dehaene, S., et al.: Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn. Sci. **10**(5), 204–211 (2006)