Explainable Artificial Intelligence: A Review of the Literature

Adrian Salazar Gomez

Abstract

Most of the currently used AI techniques can not provide information about the decision making process followed to produce a decision. However, these techniques are currently employed in critical areas such as medicine and public security and is hard to know whether the produced decisions are unfair, bias, or wrong. In response to this issue, explainable Artificial Intelligence is a multidisciplinary research area focused on solving the problems resulting from using opaque AI methodologies. Explainable artificial intelligence techniques allow users to understand and interact with the results given by a an artificial intelligence techniques. In an effort to increase the awareness of explainable artificial intelligence and encourage its development, we provide examples of critical areas where non-transparent AI techniques are currently used. In this literature review, we show the most relevant developments and point out the major challenges in the development of explainable artificial intelligence techniques. Finally, we provide research suggestion to improve explainable artificial intelligence methodologies.

Introduction

Explainable Artificial Intelligence techniques are able to explain their decisions to users. These AI techniques are fundamental for the safe and trusted application of AI in critical areas. The development of Explainable Artificial Intelligence is in an early stage and depend on multiple research areas such as social sciences, artificial intelligence, and human computer interaction [Miller, 2018]. The current research trends in the explainable artificial intelligence is focused on trying to understand what characteristics make an artificial Intelligence explainable [Ribeiro et al., 2016] and on the development of competitive artificial intelligence techniques that can return explanations about their behaviour [Gunning, 2017]. This review aims to increase the awareness about the necessity of developing explainable artificial intelligence techniques. Additionally, we aim to classify and define the main development areas in explainable artificial intelligence. Then provide and classify the most important methodologies within each research area in explainable AI.

This review is divided into five parts. We start the review by describing the development of Artificial Intelligence techniques while stressing the necessity of developing Explainable Artificial Intelligence techniques. Additionally, this part identifies different research disciplines that are contributing towards the full development of Explainable Artificial Intelligence techniques. The second part discuss the advances in the research area focus on the definition of explainability. In the third part, we cover the major explainable AI methodologies. We focus our attention into the classification of the methodologies and the most important development within each area. The fourth part identifies promising research directions for the development of Explainable Artificial Intelligence techniques. Finally, the fifth chapter summarises the Explainable Artificial Intelligence key points.

1 Why Explainable Artificial Intelligence

The intelligent systems based on Machine Learning have recently experienced an outstanding popularity because their ability to solve complex problems with a performance that can rival with humans [Mnih et al., 2013]. Resulting in its application into multiple fields such as autonomous cars [Tian et al., 2018], email filters [Guzella and Caminhas, 2009] or medicine [Esteva et al., 2017]. However, these machine learning-based models are limited by their inability to provide explanations of a decision. These limitations have raised questions about the extent in which models should be trusted and the fields where these techniques should be applied [Biran and McKeown, 2017]. The lack of trust in machine learning models and the interest of big institutions such as DARPA [Gunning, 2017] and the European Union [Goodman and Flaxman, 2016] in developing Artificial Intelligence techniques with a performance that can rival current machine learning techniques but also with the capacity to produce explanations has encouraged the research community to develop Explainable Artificial Techniques.

Explainable Artificial Intelligence (XAI) techniques have the ability to explain their behaviour. Hence, the user not only will be able to understand the set of rules that leads the AI to take decisions but also will be able to identify specific information that the AI used to take a specific decision. Additionally, in XAI users can validate the outcome of these techniques which ensures a high level of satisfaction and trust that allow the application of these techniques into critical domains.[Biran and Cotton, 2017].

The research and development of explainable Artificial Intelligence techniques covers multiple disciplines. These disciplines might variate in the literature; however, the most frequent classification of the disciplines is provided by DARPA for its Explainable AI (XAI) program [Gunning, 2017] which identifies two big research areas. The first path is focused on the technical side of the explainable Artificial Intelligence, its goal is to develop Artificial Intelligence techniques that can produce understandable explanations. The second research path is focused on the concept of explainability and covers topics such as the characteristic that a good explanation should have, the elements that an explainable AI should explain, and how to convey the right explanation. In this review, part 3 covers the technical parts and part 2 covers the the explainability components in the literature.

2 Explanations in Explainable Artificial Intelligence

2.1 Concept of explanation and explainable AI

The definition of explanation has many approaches [Lombrozo, 2006] [Miller, 2018]. However, one common and simple approach to define explanation in Artificial Intelligence is as an statement that can answer why-type questions; particularly, why and why-should questions [Gilpin et al., 2018].

Most of the works in explainable AI suggest that, in order to be fully explainable, an explainable AI has to hold certain characteristics. First, an explainable AI should provide justifications of why a decision was made and why this decision is better than other decisions. Second, an explainable AI should be able to interact with the user and allow the user to explore the solution space. Finally, specially in areas such as planning, an explainable AI should be able to explain why a decision should not be taken [Gilpin et al., 2018] [Fox et al., 2017]. However, despite the suggested characteristics are the most common among the literature, this characteristics might change depending on the area of study since a consensus has not been reached yet[Gunning, 2017].

2.2 Elements of an explanation

In AI, an explanation conveys two basic elements; interpretability and completeness. Interpretability is defined as the explanation of a process in way that human can understand it. Completeness is the capacity of an explanation to mimic the actions carried out in a process. There is a trade-off between these two concepts. Normally, in AI, the techniques with high performance provide explanations that are barely interpretable by users. Whereas,low performance methodologies tend to provide more interpretable explanations[Gilpin et al., 2018]. Consequently, one of the biggest challenges in explainable AI is to elaborate interpretable and complete explanations.

There is a lack of consensus, when defining interpretability. Despite the availability of a general concept of interpretability, there is not an agreed specific definition of it. This results in a lack of metrics to evaluate interpretability [Lipton, 2016]. Some studies have tried to clarify the concept of interpretability and the factors that have influence on it. Works such as [Kim, 2015] and [Ridgeway et al., 1998] propose a close relationship between explainability and the capacity of generate trust. However, as indicates in [Gilpin et al., 2018] the concept of trust in AI is not well defined. [Dragan et al., 2013] relates interpretability with the AI's capacity to discover causality in the structure of the data set. [Lou et al., 2012] uses interpretability as a synonym of understandability and intelligibility; also labelled as a model transparency. The formulation of the interpretability concept is necessary for the further development of explainable machine learning models. The literature review of [Lipton, 2016] covers the elements that the research community has considered to define interpretability and puts all of them together to approximate the concept of interpretability.

Overall, the development of explainable AI techniques needs the side development of accurate metrics. To then, be able to represent the interpretability of an explanation. The development of these metrics is in parallel to the development of the interpretability concept.

3 Explainable Artificial Intelligence methodologies

The research community has been working in the development of explainable AI methodologies for a long time. However, the constant increase in the applications of AI in critical fields, the concerns of important institutions such as the European Union and DARPA, and the public concerns about the unethical use of AI have fostered an exponential increase in the number of studies about the topic during the recent years [Adadi and Berrada, 2018]. Simple methodologies such as decision lists [Rivest, 1987] decision trees [Breiman et al., 1984] or explanations outputs for specific planning models [Kambhampati, 1990] were one of the first families of methodologies to be used with the goal of producing explanations. Then, because the good performance of machine learning models and the increasing complexity in the AI applications, most of the studies in the area suggested methodologies to enable explainable elements in machine learning methodologies or to develop machine learning methodologies with the inner capacity to provide explanations. However, other methodologies have been recently formulated to address the explainable AI problem such as explainable AI planning [Fox et al., 2017]. In the following sections, we cover the main methodologies and developments within each methodology.

3.1 Explainable machine learning

The initial research in explainable machine learning models was focused on using interpretable models such as decision trees, decision lists, decision sets, additive models, and rule-base models [Lakkaraju et al., 2017]. However, the high performance of the non-explainable machine learning models and its applicability into multiple cases encouraged the development of new competitive explainable machine learning techniques [Gunning, 2017].

We have divided this section into three parts. First, we classify the main methodologies to enable explications in opaque machine learning models; additionally, we indicate the main developments in each classification. Second, we emphasise the relevance of the explication methodologies in deep learning and we cover the main developments. Finally, we point out the main thoughts of the research community for further research in the development of explainable machine learning.

3.1.1 Classification of explainable machine learning methodologies

The research community has proposed numerous approaches to classify the explainable machine learning methodologies. The DARPA challenge suggests three approaches to create explainable machine learning models: Deep Explanation, Interpretable Models, and Model Induction. Deep explanation is the modification of deep learning model and neural networks into explainable models. The Interpretable Models approach consist on the creation of new machine learning algorithms with explainable features. Finally, model induction is the development of add-on techniques to be used alongside non explainable models to infer explanations. In this literature review, based in an screening of the literature, We suggest a classification of the methodologies based on three characteristics; the application moment of the methodology, the scalability of the methodology , and whether the information provided by the explainable methodology approximates a local or global behaviour of the machine learning model.

The classification we propose is in line with the classification suggested by [Gunning, 2017] but no focus in deep learning models. [Lakkaraju et al., 2017], [Lipton, 2016], and [Adadi and Berrada, 2018] suggest similar classifications to the ones we use in this study. Finally, is it worth mentioning that the classifications we use are not exclusive between them but are mutually exclusive between the labels within each classification in most of the cases. A methodology can be classified as highly scalable and produce global approximations but a methodology can not be classified; to best of our knowledge, as a model-agnostic and model specific methodology. Table 1 labels the works we reviewed into the suggested classifications.

3.1.1.1 Methodology classification based on the scope of the approximation

One common approach to enable explanations in machine learning is by producing approximations of the opaque machine learning techniques by employing explainable models to generate similar result as the original machine learning model. This type of techniques are labelled as model approximation techniques [Gilpin et al., 2018]. Approximations can be local or global depending on the model behaviour that the approximations covers. If the approximation produces an explanation considering the whole model and the logic of the entire model, the approximation is global. Whereas, if the approximation explains the the behaviour of an individual set of predictions, the a approximation is local. [Lakkaraju et al., 2017].

Within the proposed global approximation methods [Lakkaraju et al., 2017] proposes BETA as a framework to construct global explanations by generating decision sets from the model feature space. Additionally, the technique ensures the presented decisions set represents well the global behaviour of the original model by using optimisation techniques. BETA is able to generate model approximation of user's inputs to observe the behaviour of the model in situations that are of interest of the user and explore the solution set. In the same direction, [Yang et al., 2018] proposes a global approximation methodology that creates a interpretation tree built on local approximations resulting from an optimisation process. The tree structure reveals the global behaviour of the machine learning model. Finally, [Valenzuela-Escárcega et al., 2018] proposes to combine learning approaches from the representation learning field with bootstrapping techniques. The methodology outputs a scored list of the global pattern in the original model. Other works that propose global methodologies are [Letham et al., 2015] who uses sparse generative models to produce rule-sets with explanations and [Nguyen et al., 2016] who develops a methodology to explain the behaviour of

deep neural networks globally through visualizations. In general, most of the global approximations methodologies reconstruct the original models by selecting the best feature space with optimisation techniques and then using simple explainable methods in the resulting feature space.

Within the development of local approximations techniques, [Ribeiro et al., 2016] proposes LIME. This methodology uses an explainable model to approximate individual predictions by perturbing data individual data instances and observing the variations in the predictions of the original model. [Koh and Liang, 2017] approximates models by using influence functions and modifying the inputs during the model fitting process. Finally, [Baehrens et al., 2010] approximates the local behaviour of a model by using gradient explanation vectors while uncovering the most important features for the predictions. Other local approximation models are presented in [Lei et al., 2018].

Overall, all the works in either local and global approximation indicate the existence of a trade off between approximation interpretability and approximation fidelity. Hence, the more specific and detailed explanations outputs given by the approximation the lower the capacity of the approximation to mimic the result of the original model[Lakkaraju et al., 2017] [Ribeiro et al., 2016]. In general, most of the approximation studies indicate that the approximation has lower performance as the original model. However, approximation are crucial to understand how the original model predicts [Lakkaraju et al., 2017]. Most of the approximation studies aim to approach Neural Networks [Adadi and Berrada, 2018]. However, most of the approximations can be adapted to other machine learning models. Finally, [Letham et al., 2015] points out the necessity to create a measure to represent the fidelity of the approximation to the real machine learning models.

3.1.1.2 Methodology classification based on the moment where the methodology is applied

Depending on the application moment of the technique, explainable machine learning models are model-intrinsic or post-hoc. If the methodology that enables explainable characteristics take action during the operation of the machine learning model or if the methodology is part of the model, we label the methodology as a model intrinsic. On the other side, if the methodology takes action after the generation of the machine learning model in a backward-like process, the methodology is post-hoc [Lipton, 2016].

Within the model intrinsic methodologies, there are machine learning methodologies that can report information because the original design of the methodologies allows it. These, methodologies tends to be simple and limited in complex scenarios; in this classification, we have methodologies such as decision trees [Breiman et al., 1984], linear models and probabilistic models [Lou et al., 2012]. This models received wide attention before the application of machine learning into complex scenarios. Another kind of model-intrinsic methodologies suggests to modify the original architecture of the opaque models such as neural networks to produce explanations. In this direction, the studies are mostly focus in modifying artificial neural networks [Adadi and Berrada, 2018]. [Choi et al., 2016] suggests a neural network able to provide explainable elements by constraining the weights assigned to the neural networks inputs. [Simonyan et al., 2013] suggest a explainable methodology for convolutional neural networks for image classification; the methodology visualises the most relevant elements of the image. Other intrinsic methodologies are [Goudet et al., 2018], [Louizos et al., 2017], [Dong et al., 2017], [Palm et al., 2018], and [Santoro et al., 2017]. Few approximations methods are model-intrinsic. Whereas almost all the model-specific methodologies are model-intrinsic methodologies.

Post-hoc methods provide multiple kinds of explanations like numeric explanations [Xu et al., 2014], text explanantions [Krening et al., 2017] or visual explanations [Mordvintsev et al., 2015]. In overall, most of the approximation methods we describe in the previous section and almost all the model-agnostic methodologies that we describe in the next section are post-hoc. Most of the methods currently developed are model-agnositc and post-hoc. Table 1 provides more post-hoc methodologies.

3.1.1.3 Methodology classification based on the scalability of the methodology

The explainable machine learning methodologies are model-specific or model agnostic methods depending on the applicability scope of the methodologies. If the methodology is applicable into multiple machine learning models, the methodology is model agnostic. On the other side, model-specific methods are specifically designed to work in specific models [Ribeiro et al., 2016].

The main advantage of the model agnostic methodologies is their flexibility; these methodologies are applicable to every model. In most of the cases, the studies apply model-agnostic methods into artificial neural networks. However, despite its flexibility, model-agnostic methods do not produce perfect representations of the real model; hence, losing compactness and accuracy [Lakkaraju et al., 2017]. Most of the recently developed methodologies in explainable machine learning are model-agnostic [Adadi and Berrada, 2018]

[Adadi and Berrada, 2018] and [Lipton, 2016] classify the model agnostic methodologies into four categories: visualisation methodologies , knowledge extraction, influence methods, example based explanations. Visualisation methodologies represent visually the patterns within the black box models. Some works in visualisation include [Ribeiro et al., 2016] and [Yang et al., 2018] where both methodologies producing tree-shape explanations of the results. Knowledge extraction or text explanations methodologies extract verbal explanations such as rule-base-logic from the original model, as an example [Lakkaraju et al., 2017] present a knowledge extraction methodology that generates rule based explanations. Influence methods explain the model behaviour through studying the influence of the inputs in the model predictions by perturbing the original inputs. The works of [Fisher et al., 2018] and [Bien et al., 2011] presents influence methods that condense the importance of the inputs into numerical values. Is it worth mentioning that influence methodologies can be used alongside other explainable methodologies to improve the performance of the explanations since these methodologies are able to reduce the input dimension and improve the model compactness [Tibshirani, 1996]. Feature selection processes can be applied before [Hall, 1999] or during [Xu et al., 2014] the model fitting process into the data-set, Finally, example based explanations explain the model behaviour by presenting significant data-set instances and the correspondent predictions such as in [Bien et al., 2011]. In this category, [Wachter et al., 2017] presents *conceptual explanations* as a methodology to indicate the minimum changes in the data required to change the output of the machine learning methodology.

Most of the approximations we defined before are model agnostic and most of the post-hoc methodologies are model-agnostic. For example, approximation methodologies such as [Koh and Liang, 2017], [Lakkaraju et al., 2017], and [Ribeiro et al., 2016] are model agnostic. Model agnostic can be post-hoc such as[Ying et al., 2019] who develops a model agnostic method to produce explanations in graph machine learning methods. However, model agnostic methods cannot be model-intrinsic methodologies.

On the other hand, the model specific methodologies interprets the model directly and tends to maintain the original model performance. However, these methods limits the range of models we can use . Most, of the relevant developments in the model specific methodologies cover neural networks and deep learning techniques. Most of the model agnostic methodologies are model intrinsic methodologies We cover these methodologies in section 3.1.2

Overall, there are multiple approaches to classify explainable machine learning methodologies. Is worth mentioning the overlap among some classification we covered. We consider the necessity to elaborate a standard classification to facilitate the understanding of the topic and foster the development of explainable machine learning methodologies.

3.1.2 Explanations in deep learning models

In this part we discuss in detail the evolution of the methodologies that enable explainable characteristics in artificial neural networks (ANN) and Deep Networks. This focus on ANNs is in response of the 'Deep Explanation' research direction that [Gunning, 2017] suggests to develop explainable AI methodologies. [Gilpin et al., 2018] points out three direction in which the research community has transformed the deep learning method into explainable methdologies; explaining the data processing process in deep learning structures, through the representations of the data in the networks, and through the implementation of *explanantion-producing systems*.

The methodologies that explain ANNs through the explanation of the data processing process explains why an specific neural network input lead to a specific output. Hence, these methods try to simplify the complexity of the network to make them explainable. To do so, there are two approaches, by approximating the neural networks with simple models as suggested previously in part 3.1.1.1 or by using a salience map [Gilpin et al., 2018]. Among all the approximations, there are model-intrinsic approximations for neural networks such as [Schmitz et al., 1999] who proposes ANN-DT as a methodology to specifically approximate ANNs with decision trees. Whereas, the salience mapping approaches create multiple neural networks with different inputs selected by occlusion [Zeiler and Fergus, 2014]; resulting in map with the features of the data-set area that triggers changes in the network. Most of the processing methodologies are evaluated by how similar are to the original neural network [Ribeiro et al., 2016].

A representation of a deep network explains the role and structure of the data within the neural network structure. Depending on the granularity of the analysis, network representations focus on how the information works in the single units of the networks and in the layers of the artificial network [Gilpin et al., 2018]. To represent the behaviour of the neurons, studies have proposed to visualise the inputs that maximise the activity of the neurons [Zhou et al., 2014] or by assessing the ability of a neuron in solving a specific task [Bau et al., 2017]. The visualisation of the layers consist on the representation of the output of a single vector. The most common approach is to test the performance of a trained layer into other similar problem in what is called transfer learning [Sharif Razavian et al., 2014]. In overall, representation methodologies does not explain the global behaviour of the neural network architecture but we can use them to detect bias behaviours in the neural network structures

Finally, explanation-producing systems are model-intrinsic methodologies that modify the deep learning architecture to make it explainable. The techniques usually follow three approaches. First, through attention networks that learn and display the weight of the inputs [Xiao et al., 2015]. A second approach is through disentangled representations; this approach identifies relevant elements in the data; the concept is similar to a feature selection process. However, in this case the methodology is embedded into the neural network architecture. One example in this category is Beta-VAE that identifies explainable factors in generative neural networks [Higgins et al., 2017]. The last approach is through training neural networks with actual explanations. This approach generates human readable information resulting from a training process with a data-set that contains human written explanations [Antol et al., 2015].

Overall, the complete explanation of deep learning models is far from be reached [Gilpin et al., 2018]. So far, despite the numerous advances, there is no a best approach to enable explainable characteristics in neural network structures. Additionally, we have noticed a lack of deep learning methodologies able to interact with the user and produce human-readable explanations. Finally, [Gilpin et al., 2018] indicates the necessity of developing evaluation metrics for explainability.

Concluding explainability in machine learning methods, despite the efforts in enabling explainable features in machine learning, there is a shortage in the development of methodologies to transform the output of these models into human-readable explanations. Additionally, to the best of our

knowledge, [Lakkaraju et al., 2017] is the only methodology that allows in some extent iteration with the user.

During the screening of the literature, we noticed a shortage of classifications for explainable machine learning methodologies. [Gunning, 2017], [Zeng et al., 2018],[Adadi and Berrada, 2018], and [Gilpin et al., 2018] propose different classification with certain elements in common such as the differentiation between models with inherit explainable characteristics and methodologies that adapt opaque machine learning methodologies. However, the classifications are not uniform. Consequently, we encourage the research community to further research in the development of a common classification for explainable machine learning models.

Classification of the method	Methods
Model agnostic	[Koh and Liang, 2017], [Lakkaraju et al., 2017], [Ribeiro et al., 2016], [Yang et al., 2018], [Koh and Liang, 2017], [Lei et al., 2018], [Ying et al., 2019], [Xu et al., 2014], [Krening et al., 2017], [Mordvintsev et al., 2015], [Bien et al., 2011], [Wachter et al., 2017]
Model specific	[Choi et al., 2016] , [Nguyen et al., 2016], [Goudet et al., 2018], [Louizos et al., 2017], [Dong et al., 2017], [Palm et al., 2018], [Santoro et al., 2017], [Simonyan et al., 2013], [Adadi and Berrada, 2018], [Simonyan et al., 2013]
Instrinsic	[Choi et al., 2016], [Goudet et al., 2018], [Louizos et al., 2017], [Dong et al., 2017], [Palm et al., 2018], [Santoro et al., 2017]
Post-hoc	[Koh and Liang, 2017], [Lakkaraju et al., 2017], [Ribeiro et al., 2016], [Yang et al., 2018], [Koh and Liang, 2017], [Ying et al., 2019], [Xu et al., 2014], [Krening et al., 2017], [Mordvintsev et al., 2015], [Simonyan et al., 2013], [Bien et al., 2011], [Wachter et al., 2017]
Global-approximations	[Lakkaraju et al., 2017], [Yang et al., 2018], [Valenzuela-Esc arcega et al., 2018], [Letham et al., 2015] , [Nguyen et al., 2016], [Baehrens et al., 2010]
Local-approximations	[Ribeiro et al., 2016], [Koh and Liang, 2017], [Lei et al., 2018] , [Baehrens et al., 2010], [Ying et al., 2019]

Table 1: Classification of the reviewed methodologies

3.2 Planning in explainable artificial intelligence

A big proportion of the methodologies suggested to solve the explainable AI problems belong to the Machine learning field [Adadi and Berrada, 2018]. However, other AI areas have been proposed to make AI transparent. Explainable artificial intelligence planning (XAIP) is an area within AI planning that address the explainable AI problem [Fox et al., 2017]. AI planning studies the development of computational models to reach a goal state from an initial state thought a set of actions based on planner designs in response of the exploration of the environment conditions [Russell and Norvig, 2016]. Explainable Artificial Intelligence planning addresses the design of trusted planners that are able to interact with human while its decision making process remains transparent. The data independence and the transparent nature of XAIP are the mayor advantages against other explainable AI methodologies. Whereas, one of the mayor challenges in XAIP is to transform planner output's into explanations [Fox et al., 2017].

The development of techniques to explain planners behaviour is not recent [Kambhampati, 1990]. However, most of this initial explanations in planning were focus on specific planning methodologies and the explanations were aimed to planning experts [Chakraborti et al., 2017b]. The increase in interest in explainable AI and the use of planners in critical domains such as traffic control [Vallati et al., 2016], robotics [Cashmore et al., 2018], and healthcare [Canal et al., 2018] have fostered the development of studies focused on the development of explainable planners. [Fox et al., 2017] formalises the concept of explainable planning, analyses the suitability of AI planning to solve the explainable AI challenge, and presents an overview of the opportunities and challenges in the development of XAIP methodologies.[Langley et al., 2017] presents a similar concept to the concept presented in [Fox et al., 2017] but oriented to autonomous intelligent agents. The ideas in [Chakraborti et al., 2017b] and [Zhang et al., 2017] complement the ideas in [Fox et al., 2017].

Explainable AI planning is closely related with other research areas in planning such plan explanation, model reconciliation, and plan applicability. Plan explanations aims to make humans to understand plans generated by planners [Sohrabi et al., 2011]. Model reconciliation aims to reconcile differences between human models and planners models [Chakraborti et al., 2017b]. *Plan Explicability* studies the human interpretation of plans [Seegebarth et al., 2012]. Explainable AI planning takes elements from these research areas to create planner explanations for highly complex scenarios[Borgo et al., 2018].

The characteristics of AI planners make easier to elaborate explainable methodologies. First, AI planners are model-based and are independent of the domain dynamics; this allows researchers to focus on the development of explanations. Second, planners generate action-observation pair sets which facilitates causality explanation. Third, planners decisions process follows an fixed transparent criteria which can be queried to know the decision source of each action. Additionally, AI planners can perform in complex situation where no data is available [Fox et al., 2017].

Despite the potential capacity of planners to facilitates explanations, the transformation of planner actions into human-interpretable explanations remains as one of the principal challenges in XAIP. [Fox et al., 2017] presents a road-map with elements that need to be considered when designing explainable planners. The following sections describes in detail each of this XAIP challenges proposed by[Fox et al., 2017] and highlights studies that face the corespondent challenges.

3.2.1 Action causality and human-readable explanations

First, planners must be able to demonstrate the causality of their decision and explain this causality to non-planning researchers despite the complexity of the environment. [Seegebarth et al., 2012] presents a methodology to generate raw planner explanations in real time; the explanations are proofs in a first order-logic formulae as a result of employing a hybrid planning system. Additionally, the work suggests the consideration of the knowledge of user in planning knowledge and the presentation methodology as elements to influence the explanation appropriateness. [Bidot et al., 2010] suggests to improve the explanation of planners decision process by developing explainable interfaces for planners using spoken dialog systems.[Sohrabi et al., 2011] employs domain knowledge to develop a methodology that generate preferred explanations; the work formalises the concept of preference and implements into planner language to produce plans that consider preferred explanation. [Rosenthal et al., 2016] contributes directly to the elaboration of explanations in human-readable form by using verbalisation in planning and produce human-readable narratives of the planner actions. This study presents the variable verbalization algorithm to segment the agent plans into fractions that are translated into utterances by mapping the plan segment into a verbalisation space using a function.[Belvin et al., 2001], [Bohus et al., 2014], and [Thomason et al., 2015] have proposed methodologies to transform autonomous systems actions into narrations. Finally, [Chakraborti et al., 2017a] proposes the elaboration of visualisations of key component in the the plan to elaborate human-readable explanations in planning. Overall, the natural evolution to tackle this challenge is the combination of the techniques to show actions causality and techniques to generate human-readable information from planning inputs.

3.2.2 Action justification

Second, planners have to justify why the planner plan's are better than the alternatives suggested by humans. To solve this problem,[Borgo et al., 2018] presents XAI-PLAN, a methodology to fill the gap between the planner plans and the action suggested by the user and explain why the selected plan is the best. XAI-PLAN allows users to explore a finite set of alternative actions and compare the alternative plan performances with the original planner's plan. The works in model reconciliation aim to solve this problem. [Chakraborti et al., 2017b] suggests to transform the human model into

a model easy to compare with the planner plan with a series of updates [Zhang et al., 2017] tries to solve the problem by formalising the concepts explicability and predictability compatible with planners to generate plans that can be understood and compared by humans.

3.2.3 Comparison of human suggestions and planner suggestions

The third challenge in XAIP is to enable planners with methodologies to easily contrast information from users and planners. [Fox et al., 2017] proposes to use a planning algorithm to plan until a specific state where the user can compare the suggested plans with the planner validator VAL presented in[Fox et al., 2005].

3.2.4 Explain the user why an action is not feasible

Fourth, planners must be able to explain why an action is not possible. Normally, an action is not possible because the state conditions do not allow it or because the action drives to the plan failure [Fox et al., 2017]. To explain that an action is not possible in the planner, [Fox et al., 2017] proposes to use the planner validator presented in [Fox et al., 2005] as a method to detect whether the state does not meet the prerequisites for the action. [Hoffmann et al., 2014], [Steinmetz and Hoffmann, 2016], and [Bäckström et al., 2013] suggest model-checking methodologies to prove whether a specific action drives the plan to failure. Despite the existence of methodologies to prove that an action is not feasible, there is not methodology to produce explanations for nor planning-experts[Fox et al., 2017].

3.2.5 Explain why re-planning is required

Fifth, planners have to be prepared to detect and explain why the elaboration of a new plan is required and why a new plan is not necessary. The literature has elaborated techniques to know whether a re-planning is necessary. [Fox et al., 2017] proposes to use the proposes filter violation techniques as showed in [Cashmore et al., 2015] where the method updates a knowledge base about the new environment and uses filters to find re-planning instances. For the same problem, [Molineaux et al., 2012] suggests *discover history* as an algorithm that explores previous observations in form of historical logs to predict future states; this algorithm detect re-planning instances and provides justification aimed to planner researchers. The presented methodologies lack in the elaboration of explanations for non-planner professionals.

3.2.6 Challenges in explainable AI planning

Because the early stage in which XAIP is currently, there are numerous aspects that XAIP needs to cover. [Fox et al., 2017] suggest to expand XAIP to temporal planning, probabilistic planning , planning in uncertain situations. Additionally, [Fox et al., 2017] points out the necessity of formalising the concept of explanaibility in general and in planning. [Seegebarth et al., 2012] mentions the necessity of performing studies to find ways to measure the quality of an explanation. [Borgo et al., 2018] believes that further research is necessary in the assessment of the context to deliver meaningful explanations and assess the the explanation impact on the user trust. [Chakraborti et al., 2017a] indicates that further research needs to be done in the the model acquisition process. In overall, all the XAIP works consider the necessity to revise previous planning methodologies and re-adapt them into the explainable planning framework. Another general concern in the XAIP literature is to bring ideas from other disciplines such as psychology or social-sciences to fill the gap between explanations and planners output. A good starting point to know the main characteristic of an explainable planner is in the road-map proposed in [Fox et al., 2017].

As a conclusion, explainable artificial intelligence planning has brought the attention of the research community attention as a methodology to solve the AI explainable problem because its natural capacity to be explainable, the increasing use of planning, and its capacity to solve situations where learning is not available. However, XAIP is in a early stage and requires the formulation of concrete methodologies that considers previous work from multiple research areas to eventually be able to provide a methodology/framework to solve a whole AI problem while providing trusted explanations.

4 Further development in explainable AI

In our literature review, we have identified that most of the explainable machine learning methodologies are focused on enabling transparency characteristics in advanced machine learning models. However, there are few studies aiming to transform machine learning outputs into human readable format and there are less studies aiming to allow interaction between the machine learning method and the final user. We would like to recommend the development of interactive methodologies. To do so, machine learning research can get inspiration from XAIP where we have found more studies focus on user-AI iteration.

In our literature review, we have identified the potential of explainable planning; particularly, in situations where learning is not possible. However, we have noticed that most of the works in XAIP are conceptual. In response, we would like to recommend the development of formal methodologies to explain the causality of an action into human-readable form or planning methodologies that can have complex interaction with the suggestion the user. Additionally, we would like to see works that implements the concept of planning into probabilistic planning and temporal planning. pecially in areas such as robotics, p. and in situations . The fully formulation of the problem. The development of a whole framwork to provide explanantion to the user.

In overall, we did not find many formal techniques to transform any kind of input provided from an artificial intelligence technique into a human readable explanation. To do so a research in the natural language generation might be helpful. Additionally, we found particularly interesting the idea of explanation through visualisation which might be useful for Human-AI Teaming by the implementation of visualisations into the emerging virtual reality technologies.

Finally, most of the studies in the literature remark the necessity of defining a metric to measure the level of explainability or interpretability. Ideas to produce these metrics can be inspired from other areas such as logic, argumentation, fuzzy logic, and provenance.

5 Conclusion

In this literature review, we have explained why the development of explainable artificial intelligence techniques is crucial for the further development of artificial intelligence and its ethical implementations. We have identified the basic elements that makes an artificial intelligence explainable and we have remarked areas in the definition of explainability that need a consensus among the research community. After a screening of the literature, we detected that most of the literature is focus on enabling transparent and explainable features in machine learning methodologies. However, other AI methodologies such as AI planning can contribute to solve the explainable AI challenge. Finally, we have identified numerous challenges the research community need to tackle to create a fully explainable AIs while competing in performance with opaque artificial intelligence methodologies; thus, solve one of the most critical problems in the current AI field.

References

[Adadi and Berrada, 2018] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

[Antol et al., 2015] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

[Bäckström et al., 2013] Bäckström, C., Jonsson, P., and Ståhlberg, S. (2013). Fast detection of unsolvable planning instances using local consistency. In *Sixth Annual Symposium on Combinatorial Search*.

[Baehrens et al., 2010] Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and MÃžller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831.

[Bau et al., 2017] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549.

[Belvin et al., 2001] Belvin, R., Burns, R., and Hein, C. (2001). Development of the hrl route navigation dialogue system. In *Proceedings of the first international conference on Human language technology research*.

[Bidot et al., 2010] Bidot, J., Biundo, S., Heinroth, T., Minker, W., Nothdurft, F., and Schattenberg, B. (2010). Verbal plan explanations for hybrid planning. In *MKWI*, pages 2309–2320. Citeseer.

[Bien et al., 2011] Bien, J., Tibshirani, R., et al. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4):2403–2424.

[Biran and Cotton, 2017] Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 8.

[Biran and McKeown, 2017] Biran, O. and McKeown, K. R. (2017). Human-centric justification of machine learning predictions. In *IJCAI*, pages 1461–1467.

[Bohus et al., 2014] Bohus, D., Saw, C. W., and Horvitz, E. (2014). Directions robot: in-the-wild experiences and lessons learned. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 637–644. International Foundation for Autonomous Agents and Multiagent Systems.

[Borgo et al., 2018] Borgo, R., Cashmore, M., and Magazzeni, D. (2018). Towards providing explanations for ai planner decisions. *arXiv preprint arXiv:1810.06338*.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. wadsworth int. *Group*, 37(15):237–251.

[Canal et al., 2018] Canal, G., Alenyà, G., and Torras, C. (2018). Adapting robot task planning to user preferences: an assistive shoe dressing example. *Autonomous Robots*, pages 1–14.

[Cashmore et al., 2018] Cashmore, M., Fox, M., Long, D., Magazzeni, D., and Ridder, B. (2018). Opportunistic planning in autonomous underwater missions. *IEEE Transactions on Automation Science and Engineering*, 15(2):519–530.

[Cashmore et al., 2015] Cashmore, M., Fox, M., Long, D., Magazzeni, D., Ridder, B., Carrera, A., Palomeras, N., Hurtos, N., and Carreras, M. (2015). Rosplan: Planning in the robot operating system. In *Twenty-Fifth International Conference on Automated Planning and Scheduling*.

[Chakraborti et al., 2017a] Chakraborti, T., Fadnis, K. P., Talamadupula, K., Dholakia, M., Srivastava, B., Kephart, J. O., and Bellamy, R. K. (2017a). Visualizations for an explainable planning agent. *arXiv preprint arXiv:1709.04517*.

[Chakraborti et al., 2017b] Chakraborti, T., Sreedharan, S., Zhang, Y., and Kambhampati, S. (2017b). Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*.

[Choi et al., 2016] Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512.

[Dong et al., 2017] Dong, Y., Su, H., Zhu, J., and Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4306–4314.

[Dragan et al., 2013] Dragan, A. D., Lee, K. C., and Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 301–308. IEEE Press.

[Esteva et al., 2017] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115.

[Fisher et al., 2018] Fisher, A., Rudin, C., and Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the rashomon perspective. *arXiv preprint arXiv:1801.01489*.

[Fox et al., 2005] Fox, M., Howey, R., and Long, D. (2005). Validating plans in the context of processes and exogenous events. In *AAAI*, volume 5, pages 1151–1156.

[Fox et al., 2017] Fox, M., Long, D., and Magazzeni, D. (2017). Explainable planning. *CoRR*, abs/1709.10256.

[Gilpin et al., 2018] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: an approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*.

[Goodman and Flaxman, 2016] Goodman, B. and Flaxman, S. (2016). European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.

[Goudet et al., 2018] Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. (2018). Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer.

[Gunning, 2017] Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.

[Guzella and Caminhas, 2009] Guzella, T. S. and Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206–10222.

[Hall, 1999] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[Higgins et al., 2017] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

[Hoffmann et al., 2014] Hoffmann, J., Kissmann, P., and Torralba, A. (2014). " distance"? who cares? tailoring merge-and-shrink heuristics to detect unsolvability. In *ECAI*, pages 441–446.

[Kambhampati, 1990] Kambhampati, S. (1990). A classification of plan modification strategies based on coverage and information requirements. In *AAAI 1990 Spring Symposium on Case Based Reasoning*. Citeseer.

[Kim, 2015] Kim, B. (2015). *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology.

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

[Krening et al., 2017] Krening, S., Harrison, B., Feigh, K. M., Isbell, C. L., Riedl, M., and Thomaz, A. (2017). Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55.

[Lakkaraju et al., 2017] Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.

[Langley et al., 2017] Langley, P., Meadows, B., Sridharan, M., and Choi, D. (2017). Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*.

[Lei et al., 2018] Lei, J., GSell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

[Letham et al., 2015] Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371.

[Lipton, 2016] Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

[Lombrozo, 2006] Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

[Lou et al., 2012] Lou, Y., Caruana, R., and Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM.

[Louizos et al., 2017] Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456.

[Miller, 2018] Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

[Mnih et al., 2013] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

[Molineaux et al., 2012] Molineaux, M., Kuter, U., and Klenk, M. (2012). Discoverhistory: Understanding the past in planning and execution. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 989–996. International Foundation for Autonomous Agents and Multiagent Systems.

[Mordvintsev et al., 2015] Mordvintsev, A., Olah, C., and Tyka, M. (2015). Inceptionism: Going deeper into neural networks.

[Nguyen et al., 2016] Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395.

[Palm et al., 2018] Palm, R., Paquet, U., and Winther, O. (2018). Recurrent relational networks. In *Advances in Neural Information Processing Systems*, pages 3372–3382.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). 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. ACM.

[Ridgeway et al., 1998] Ridgeway, G., Madigan, D., Richardson, T., and O'Kane, J. (1998). Interpretable boosted naïve bayes classification. In *KDD*, pages 101–104.

[Rivest, 1987] Rivest, R. L. (1987). Learning decision lists. *Machine learning*, 2(3):229–246.

[Rosenthal et al., 2016] Rosenthal, S., Selvaraj, S. P., and Veloso, M. M. (2016). Verbalization: Narration of autonomous robot experience. In *IJCAI*, pages 862–868.

[Russell and Norvig, 2016] Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[Santoro et al., 2017] Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976.

[Schmitz et al., 1999] Schmitz, G. P., Aldrich, C., and Gouws, F. S. (1999). Ann-dt: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*, 10(6):1392–1401.

[Seegebarth et al., 2012] Seegebarth, B., Müller, F., Schattenberg, B., and Biundo, S. (2012). Making hybrid plans more clear to human users-a formal approach for generating sound explanations. In *Twenty-Second International Conference on Automated Planning and Scheduling*.

[Sharif Razavian et al., 2014] Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813.

[Simonyan et al., 2013] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[Sohrabi et al., 2011] Sohrabi, S., Baier, J. A., and McIlraith, S. A. (2011). Preferred explanations: Theory and generation via planning. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[Steinmetz and Hoffmann, 2016] Steinmetz, M. and Hoffmann, J. (2016). Towards clause-learning state space search: Learning to recognize dead-ends. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[Thomason et al., 2015] Thomason, J., Zhang, S., Mooney, R. J., and Stone, P. (2015). Learning to interpret natural language commands through human-robot dialog. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[Tian et al., 2018] Tian, Y., Pei, K., Jana, S., and Ray, B. (2018). Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering*, pages 303–314. ACM.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Valenzuela-Escárcega et al., 2018] Valenzuela-Escárcega, M. A., Nagesh, A., and Surdeanu, M. (2018). Lightly-supervised representation learning with global interpretability. *arXiv preprint arXiv:1805.11545*.

[Vallati et al., 2016] Vallati, M., Magazzeni, D., De Schutter, B., Chrpa, L., and McCluskey, T. L. (2016). Efficient macroscopic urban traffic models for reducing congestion: a pddl+ planning approach. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[Wachter et al., 2017] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):2018.

[Xiao et al., 2015] Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., and Zhang, Z. (2015). The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850.

[Xu et al., 2014] Xu, Z., Huang, G., Weinberger, K. Q., and Zheng, A. X. (2014). Gradient boosted feature selection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 522–531. ACM.

[Yang et al., 2018] Yang, C., Rangarajan, A., and Ranka, S. (2018). Global model interpretation via recursive partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1563–1570. IEEE.

[Ying et al., 2019] Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). Gnn explainer: A tool for post-hoc explanation of graph neural networks. *arXiv preprint arXiv:1903.03894*.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

[Zeng et al., 2018] Zeng, Z., Miao, C., Leung, C., and Chin, J. J. (2018). Building more explainable artificial intelligence with argumentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Zhang et al., 2017] Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H. H., and Kambhampati, S. (2017). Plan explicability and predictability for robot task planning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1313–1320. IEEE.

[Zhou et al., 2014] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2014). Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.