Argumentation and Explainable Artificial Intelligence - A survey

Alexandros Vassiliades\textsuperscript{1,2}, Nick Bassiliades\textsuperscript{1}, Theodore Patkos\textsuperscript{2}

\textsuperscript{1}Aristotle University of Thessaloniki, School of Informatics, Ethnikis Aminis 36, Thessaloniki, Greece, 54621
E-mail: valexande@csd.auth.gr, nbassili@csd.auth.gr
\textsuperscript{2}Institute of Computer Science, Foundation for Research and Technology - Hellas, Vasilika Vouton 100, Heraklion, Greece, 70013
E-mail: patkos@ics.forth.gr

Abstract

Argumentation and eXplainable Artificial Intelligence (XAI) are closely related, as in the recent years Argumentation has been used for providing Explainability to AI. Argumentation can show step by step how an AI System reaches a decision, it can provide reasoning over uncertainty, and can find solutions when conflicting information is faced. In this survey, we elaborate over the topics of Argumentation and XAI combined, by reviewing all the important methods and studies, as well as implementations that use Argumentation to provide Explainability in AI. More specifically, we show how Argumentation can enable Explainability for solving various types of problems in decision making, justification of an opinion, and dialogues. Subsequently, we elaborate on how Argumentation can help in constructing explainable systems in various applications domains, such as in Medical Informatics, Law, the Semantic Web, Security, Robotics, and some general purpose systems. Finally, we present approaches that combine Machine Learning and Argumentation Theory, towards more interpretable predictive models.

1 Introduction

Explainability of an Artificial Intelligence (AI) system (i.e., tracking the steps that lead to the decision), was an easy task in the early stages of AI, as the majority of the systems were logic based. For this reason, it was easy to provide transparency to their decisions by providing explanations, and therefore to gain the trust of their users. This changed in the last twenty years, when data driven methods started to evolve and became part of most AI systems, giving them computational capabilities and learning skills that cannot easily be reached by means of logic languages alone. Eventually, the steadily increasing complexity of computational evolution of AI methods resulted in more obscure systems.

Therefore, a new research field appeared in order to make AI systems more explainable, called eXplainable Artificial Intelligence (XAI). The graph presented in Figure\textsuperscript{1} showing the Google searches that contain the keyword XAI is very interesting, as it shows that people’s searches are steadily increasing since the mid of 2016, indicating the interest in explaining decisions in AI (the picture was part of the study Adadi & Berrada (2018)). Capturing an accurate definition of what can be considered an explanation is quite challenging, as can be seen in (Miller (2019)). Among many definitions, some commonly accepted ones are:

- An explanation is an assignment of causal responsibility (Josephson & Josephson (1996)).
- An explanation is both a process and a product i.e. it is the process and the result of a Why? question, (Lombrozo (2006)).
- An explanation is a process to find meaning or create shared meaning, (Malle (2006)).

Providing explanations to an AI system has two directions; the first one is to gain trust in a system or convince a user, and the other is for the scientists to understand how a data driven model reaches a
decision. The first case has many real word implementations which explain the decision of a system to the user. A significant amount of work can be found in the fields of Medical Informatics (Holzinger et al. (2017), Tjoa & Guan (2019), Lamy et al. (2019)), Legal Informatics (Waltl & Vogl (2018), Deeks (2019)), Military Defense Systems (Madhikermi et al. (2019), Keneni et al. (2019), Core et al. (2006, 2005)), and Robotic Platforms (Sheh (2017), Anjomshoae et al. (2019)). In the second case, the studies try to enhance transparency in the data driven model (Bonacina (2017), Gunning & Aha (2019), Yang & Shafto (2017), Samek & Müller (2019), Fernandez et al. (2019)); in some cases also visualization are used (Choo & Liu (2018)).

Figure 1: Google Searches with XAI (Adadi & Berrada (2018)).

In the last decade, Argumentation has achieved significant impact to XAI. Argumentation has strong Explainability capabilities, as it can translate the decision of an AI system in an argumentation procedure, which shows step by step how it conclude to a result. Every Argumentation Framework (AF) is based upon well defined mathematical models, from which the basic definitions are close to Set Theory, extended with some extra properties between the elements. The advantages of Argumentation, which give aid to XAI, are that given a set of possible decisions, the decisions can be mapped to a graphical representation, with predefined attack properties that subsequently will lead to the winning decision, and will show the steps that were followed in order to reach it.

This study provides an overview over the topics of Argumentation and XAI combined. We present a survey that describes how Argumentation enables XAI in AI systems. Argumentation combined with XAI is a wide research field, but our intention is to include the most representative relevant studies. We classify studies based on the type of problem they solve such as Decision Making, Justification of an opinion, and Dialogues between Human-System or System-System scenarios, and show how Argumentation enables XAI when solving these problems. Then, we delineate on how Argumentation has helped in providing explainable systems, in the application domains of Medical Informatics, Law Informatics, Robotics, the Semantic Web (SW), Security, and some other general purpose systems. Moreover, we get into the field of Machine Learning (ML) and address how transparency can be achieved with the use of an AF. The contributions of our study are the following:

(i) We present an extensive literature review of how Argumentation enables XAI.
(ii) We show how Argumentation enables XAI, for solving problems in Decision Making, Justification, and Dialogues.
(iii) We present how Argumentation has helped build explainable systems in the application domains of Medical Informatics, Law, the Semantic Web, Security, and Robotics.
(iv) We show how Argumentation can become the missing link between ML and XAI.

The remainder of this study is organized as follows. Section 2 presents the motivation and the contributions of this survey. Section 3 discusses Related Works and describes other surveys related to Argumentation or XAI. Section 4 contains the background needed for the terms in the subsequent sections. In Section 5 we present how Argumentation enables Explainability in Decision Making, Justification, and Dialogues. Moreover, we present how agents can use Argumentation to enhance their Explainability skills and what principles they must follow, in order not to be considered biased. Subsequently, Section 6 shows how Argumentation helped build explainable systems in various application fields. Section 7 elaborates on studies that combine Argumentation and ML, in order to show how
Argumentation can become the missing link between ML and XAI. Finally, we discuss our findings and conclude in Section 8.

2 Motivation

Argumentation Theory is developing into one of the main reasoning methods to capture Explainability in AI. The quantity of theoretical studies that use Argumentation to enable Explainability, as well as the plurality of explainable systems that use Argumentation to provide explanations in so many application fields that are presented in this survey, are proofs for the importance of Argumentation Theory in XAI. Nevertheless, previous surveys over Argumentation either do not point out its Explainability capabilities, or present the Explainability capabilities of Argumentation only for a specific domain (see Section 3). Therefore, we believe that there is a need in the literature for a survey over the topic of Argumentation and XAI combined, for various problem types and applications domains.

Firstly, we want to present an extensive literature overview of how Argumentation enables XAI. For this reason, we classify studies based on the most important practical problem types that Argumentation solves, such as decision making, justification of an opinion, and explanation through dialogues. Our goal is to show how Argumentation enables XAI in order to solve such problems. We believe that such a classification is more interesting for the reader that tries to locate which research studies are related to the solution of specific problem types.

Secondly, we want to point out the capabilities of Argumentation in building explainable systems. We can see that any AI system that chooses Argumentation as its reasoning mechanism for explaining its decision can gain great Explainability capabilities, and provide explanations which are closer to the human way of thinking. Henceforth, using Argumentation for providing explanations makes an AI system friendlier and more trustworthy to the user. Our goal is to show that Argumentation for building explainable systems is not committed to one application domain. Therefore, we present an overview of studies in many domains such as Medical Informatics, Law, the SW, Security, Robotics, and some general purpose systems.

Finally, our intention is to connect ML, the field that brought to the surface XAI, with Argumentation. The literature review over studies that combine Argumentation with ML to explain the decision of data driven models revealed how closely related those two fields are. For this reason, we wanted to show that Argumentation can act as a link between ML and XAI.

3 Related Work

In this section, we present surveys that are related to Argumentation or XAI. Our intention is to help the reader to obtain an extensive look in the field of Argumentation or XAI, and become aware of its various forms, capabilities, and implementations. One could read the surveys of (Modgil et al. (2013)) in order to understand the uses of Argumentation, (Baroni et al. (2011)) to understand how Abstract AFs are defined and their semantics, (Amgoud et al. (2008)) to understand how Bipolar AFs are defined, (Doutre & Mailly (2018)) to understand the dynamic enforcement that Argumentation Theory has over a set of constrains, and (Bonzon et al. (2016)) to see how we can compare set of arguments. Moreover, most studies in this section indicate what is missing in Argumentation Theory or XAI in general and how the gaps should be filled.

Argumentation is becoming one of the main mechanisms when it comes to explaining the decision of an AI system. Therefore, understanding how an argument is defined as acceptable within an AF is crucial. An interesting study to understand such notions is presented in (Cohen et al. (2014)), where the authors present a survey which analyzes the topic of support relations between arguments. The authors talk about the advantages and disadvantages of the deductive support, necessity support, evidential support, and backing. Deductive support captures the intuition: if an argument \( A \) supports argument \( B \) then the acceptance of \( A \) implies the acceptance of \( B \) and, as a consequence, the non-acceptance of \( B \) implies the non-acceptance of \( A \). Necessity support triggers the following constraint: if argument \( A \) supports argument \( B \) it means that \( A \) is necessary for \( B \). Hence, the acceptance of \( B \) implies the acceptance of \( A \) (conversely the non-acceptance of \( A \) implies the non-acceptance of \( B \)). Evidential support states that...
arguments are accepted if they have a support that will make them acceptable by the other participants in a conversation. Backing provides support to the claim of the argument. The authors showed that each support establishes different constraints to the acceptability of arguments.

An important survey for solving reasoning problems in a AF is introduced in (Charwat et al. (2015)), where the authors show the different techniques of solving implementation issues for an AF. The authors group the techniques into two different classes. Firstly, the reduction-based techniques where the argumentation implementation problem is transformed into another problem, a satisfiability problem in propositional logic (Biere et al. (2009)), or a constraint-satisfaction problem (Dechter et al. (2003)), or an Answer Set Programming (ASP) problem (Fitting (1992), Lifschitz (2019)). Reduction-based techniques have the advantages: (i) they are directly adapted with newer versions of solvers and (ii) they can be easily adapted to specific needs which an AF may need to obey. While, the other category called direct approaches refers to systems and methods implementing AF from scratch, thus allowing algorithms to realize argumentation specific optimizations.

Argumentation and ML is fused in (Longo (2016), Cocarascu & Toni (2016)). Longo (2016) in his study considers that any AF should be divided into sub-components, to make the training of ML classifiers easier when they are asked to build an AF from a set of arguments and relations between them. He considers that there should be five different classifiers, one for understanding the internal structure of arguments, one for the definition of conflicts between arguments, another for the evaluation of conflicts and definition of valid attacks, one for determining the dialectical status of arguments, and a last one to accrue acceptable arguments. Thus, he provides in his survey any ML classifier that has been built for each component, studies that have defined a formalization for the elements of any component, and studies that provide data for training. Nevertheless, he mentions that there is a lack of data to train a classifier for each sub-component. On the other hand, Cocarascu & Toni (2016) analyze the implementation of Argumentation in ML, categorizing them by the data driven model they augment, the arguments, the AF, and semantics they deploy. Therefore, they show real life systems of ML and Reinforcement Learning models that are aided by Argumentation in the scope of Explainability, in their survey elaborate over the topic of Abduction and Argumentation, which are presented as two different forms of reasoning that can play a fundamental role in ML. More specifically, the authors elaborate on how reasoning with Abduction and Argumentation, can provide a natural-fitting mechanism for the development of explainable ML models.

Two similar surveys are (Moens (2018), Lippi & Torroni (2016)), where the authors elaborate over the topic of Argumentation Mining (AM), from free text, and dialogues through speech. AM is an advanced form of Natural Language Processing (NLP). The classifiers in order to understand an argument inside a piece of text or speech, must firstly understand the whole content of the conversation, the topic of the conversation, as well as the specific key phrases that may indicate whether an argument exists. The aforementioned actions facilitate the identification of the argument in a sentence or dialog. Further analysis is necessary to clarify what kind of argument has been identified (i.e. opposing, defending, etc.). There are two key problems identified for AM systems in both surveys: (i) the fact that they cannot support a multi-iteration argumentation procedure, since it is hard for them to extract argument from a long argumentation dialogue, (ii) the lack of training data to train argument annotators, apart from some great efforts such as: The Debate, Debatepedia, Idebate, VBATES and ProCon. Moreover, (Moens (2018)) talks about studies where also facial expressions are inferred through a vision module to better understand the form of the argument. Another survey on AM is (Lawrence & Reed (2020)).

Finally, the explanation of Case-Based Reasoning (CBR) systems is explored in the survey of (Sormo et al.) (2005). Even though the authors do not include Argumentation in their survey, CBR works similarly to Case-Based Argumentation (CBA). Therefore, one could find interesting information about

http://www.debatepedia.org
https://idebate.org
http://vbate.idebate.org
https://www.procon.org
Explainability with CBR. Moreover, the authors extend the study of Wick & Thompson (1992), in which reasoning methods that take into consideration the desires of the user and the system are presented, in order to follow explanation pipelines. The pipelines capture the different methods that can help a system reach a decision:

(i) **Transparency**: Explain How the System Reached the Answer.
(ii) **Justification**: Explain Why the Answer is a Good Answer.
(iii) **Relevance**: Explain Why a Question Asked is Relevant.
(iv) **Conceptualization**: Clarify the Meaning of Concepts.
(v) **Learning**: Teach the user about the domain to state the question better.

On the other hand, the literature of surveys for XAI is also rich. A smooth introduction to XAI is the paper of Miller (2019), where the author reviews relevant papers from philosophy, cognitive psychology, and social psychology and he argues that XAI can benefit from existing models of how people define, generate, select, present, and evaluate explanations. The paper drives the following conclusions: (1) Why? questions are contrastive; (2) Explanations are selected in a biased manner; (3) Explanations are social; and (4) Probabilities are not as important as causal links. As an extension of these ideas we can see the extensive survey of Atkinson et al. (2020) on the topic of Explainability in AI and Law.

Fundamentally, XAI is a field that came to the surface when AI systems moved from logic based to data driven models with learning capabilities. We can see this in the survey of Adadi & Berrada (2018), where the authors show how XAI methods have developed during the last twenty years. As it was natural, data driven models increased the complexity of tracking the steps to reach a decision. Thus, the Explainability of a decision was considered as a “black box”. For this reason, a lot of studies have tried to provide even more Explainability to data driven models especially in the last decade. We can see many similar surveys that describe the Explainability methods which are considered state of the art, for the decision of various data driven models in Mozina et al. (2007), Došilović et al. (2018), Schoenborn & Althoff (2019), Guidotti et al. (2018), Collenette et al. (2020)). Nevertheless, there are many AI systems that still have not reached the desired transparency for the way they reach their decisions. A survey with open challenges in XAI can be found in the study of Das & Rad (2020).

Deep learning is the area of ML that is the most obscure to explain its decision. Even though many methods have been developed to achieve the desired level of transparency, there are still a lot of open challenges in this area. A survey that gathers methods to explain the decision of a deep learning model, as well as the open challenges, can be found in Carvalho et al. (2019). In this scope, we can find other more practical surveys that talk about Explainability of decision making for data driven models in Medical Informatics (Joa & Guan (2019), Pociuviću et al. (2020), or in Cognitive Robotic Systems (Anjomshoae et al. (2019)). The last three studies, present how data driven models give explanations for their decision in the field of Medical Informatics in order to recommend a treatment, to make a diagnosis (with the help of an expert making the final call), and image analysis to classify an illness, for example through magnetic resonance images to classify if a person has some type of cancer.

A theoretical scope on why an explanation is desired for the decision of an AI system, can be found in the study of Koshiyama et al. (2019), where the authors argue that a user has the right to know every decision that may change her life. Hence, they gather AI systems that offer some method of providing explanations for their decisions and interact with human users. This study was supported two years ago by the European Union which has defined new regulations about this specific topic (Regulation (2016)). Another, human-centric XAI survey is that of Paez (2019), where the author shows the different types of acceptable explanations of a decision based on cognitive psychology and groups the AI systems according to the type of explanation they provide. Moreover, the author reconsiders the first grouping based on the form of understanding (i.e., direct, indirect, etc) the AI systems offer to a human.

### 4 Background

The theoretical models of Argumentation obtained a more practical substance in 1958 by Toulmin through his book *The Uses of Argumentation* (Toulmin (1958)), where he presented how we can use Argumentation
to solve everyday problems. Yet, the demanding computational complexity limited the applicability of AFs for addressing real-world problems. Fortunately, during the last twenty years Argumentation Theory was brought back to the surface, new books were introduced (Walton [2005], Besnard & Hunter [2008]), and mathematical formalizations were defined. The Deductive Argumentation Framework (DAF) (Besnard & Hunter [2001]) is the first mathematical formalization describing an AF and its non-monotonic reasoning capabilities.

In this section, we will give the basic definitions of an Abstract Argumentation Framework (AAF) (Dung [1995], Dung & Son [1995]), and we will introduce several extensions of the AAF. More specifically, we are going to talk about the Structured Argumentation Framework (SAF) (Dung [2016]), the Label Based Argumentation Framework (LBAF) (Caminada [2008]), the Bipolar Argumentation Framework (BAF) (Cayrol & Lagasquie-Schiex [2005]), the Quantitative Bipolar Argumentation Framework (QBAF) (Baroni, Rago & Toni [2018]), and the Probabilistic Bipolar Argumentation Framework (PBAF) (Fazzinga et al. [2018]). All these frameworks are used in studies mentioned in the next sections, so to avoid over analysis we will give the definitions of the AAF, and we will mention the dimension being extended from its variations.

**Deductive Argumentation Framework:** DAF is defined only on first order logic rules and terms, and all the aforementioned frameworks are built upon it. DAF considers an argument as a set of formulae that support a claim, using elements from a propositional (or other type) language \( \Delta \). Thus, arguments in DAF are represented as \( \langle \Gamma, \phi \rangle \), where \( \Gamma \subseteq \Delta \) denotes the set of formulae and is called the support of the argument, which help establish the claim \( \phi \). The following properties hold: (i) \( \Gamma \subseteq \Delta \), (ii) \( \Gamma \vdash \phi \), (iii) \( \Gamma \not\vdash \bot \), and (iv) \( \Gamma \) is minimal with respect to set inclusion for which (i), (ii) and (iii) hold. Furthermore, there exist two types of attack, undercut and rebut.

(i) For any propositions \( \phi \) and \( \psi \), \( \phi \) attacks \( \psi \) when \( \phi \equiv \neg \psi \) (the symbol \( \neg \) denotes the strong negation).

(ii) Rebut: \( \langle \Gamma_1, \phi_1 \rangle \) rebuts \( \langle \Gamma_2, \phi_2 \rangle \), if \( \phi_1 \) attacks \( \phi_2 \).

(iii) Undercut: \( \langle \Gamma_1, \phi_1 \rangle \) undercuts \( \langle \Gamma_2, \phi_2 \rangle \), if \( \phi_1 \) attacks some \( \psi \in \Gamma_2 \).

(iv) \( \langle \Gamma_1, \phi_1 \rangle \) attacks \( \langle \Gamma_2, \phi_2 \rangle \), if it rebuts or undercuts it.

**Abstract Argumentation Framework:** An AAF is a pair \((A, R)\), where \(A\) is a set of arguments and \(R \subseteq A \times A\) a set of attacks, such that each \(a, b \in A\) the relation \((a, b) \in R\) means \(a\) attacks \(b\) (equivalently \((b, a) \in R\) means \(b\) attacks \(a\)). Let, \(S \subseteq A\) we call:

(i) \(S\) conflict free, if \(\forall a, b \in S\) holds \((a, b) \notin R\) (or \((b, a) \notin R\)).

(ii) \(S\) defends an argument \(a \in A\) if \(\forall b \in A\) such that \((b, a) \in R\), \(\exists c \in S\) and \((c, b) \in R\).

(iii) \(S\) is admissible if \(S\) is conflict free, and \(\forall a \in S, \forall b \in A\) such that \((b, a) \in R\), \(\exists c \in S\) holds \((c, b) \in R\).

Next, the semantics of AAF are specific subsets of arguments, which are defined from the aforementioned properties. But first we need to introduce the function \(F\), where \(F : 2^A \rightarrow 2^A\), such that for \(S \subseteq A\), \(F(S) = \{a \mid a\ is\ defined\ by\ S\}\). The fixpoint of a function \(F\) given a set \(S\), is a point where the input of the function is identical to the output, \(F(S) = S\).

(i) **Stable:** Let \(S \subseteq A\), \(S\) is a stable extension of \((A, R)\), iff \(S\) is conflict free and \(\forall a \in A \setminus S, \exists b \in S\) such that \((b, a) \in R\).

(ii) **Preferred:** Let \(S \subseteq A\), \(S\) is a preferred extension of \((A, R)\), iff \(S\) is maximal for the set inclusion among the admissible sets of \(A\).

(iii) **Complete:** Let \(S \subseteq A\), \(S\) is a complete extension of \((A, R)\), iff \(S\) is conflict free fixpoint of \(F\).

(iv) **Grounded:** Let \(S \subseteq A\), \(S\) is a grounded extension of \((A, R)\), iff \(S\) is the minimal fixpoint of \(F\).

**Structured Argumentation Framework:** SAF represents the arguments in the form of logical rules (Rule 1), and it introduces the constraints of preference between arguments. Firstly, we need to define some new concepts. A theory is a pair \((T, P)\) whose sentences are formulae in the background monotonic logic \((L, \vdash)\) of the form \(L \leftarrow L_1, \ldots, L_n\) where \(L, L_1, \ldots, L_n\) are ground literals. Henceforth, \(T\) is a set of ground literals, and \(P\) is a set of rules which follow the general form label : claim ← premise (Rule 1).

\[ r : a \leftarrow b_1, \ldots, b_n \]  

(1)
Argumentation and Explainable Artificial Intelligence - A survey

for $n \in \mathbb{N}$, and $a, b_1, \ldots, b_n \in \mathcal{T}$.

Rule $(I)$ is understood as if the facts $b_1, \ldots, b_n$ are true, then its claim $a$ is true, otherwise if any of the facts is false the claim is false. Additionally, if we have two rules $r, r'$ similar to Rule $(I)$, we define the preference of $r$ over $r'$ by\textit{preffer}(r, r'). The arguments in SAF have the same format similar to Rule $(I)$. Therefore, when we say we prefer an argument $a$ over $b$ we mean the relation\textit{preffer}(a, b). If the preference rules are removed from the framework then it is also called Assumption Based Argumentation Framework (ABA)\cite{dung2009}, where a set of assumptions (i.e. body of rule) support a claim (i.e. head of rule). Additionally, ABA can tackle incomplete information, because if we do not have all the literals from the body of a rule, we can make some assumptions, to fill the missing information.

Label Based Argumentation Framework: LBAF is a framework where the arguments are characterized by a label, which defines the acceptability of an argument. Briefly:

(i) an argument is labelled\textit{in} if all of its attackers are labelled\textit{out}, and is called acceptable.

(ii) an argument is labelled\textit{out} if at least one of its attackers is labelled\textit{in}.

(iii) an argument is labelled\textit{undec}, when we cannot label it neither\textit{in} nor\textit{out}.

Keeping this in mind, let $(\mathcal{A}, \mathcal{R})$ be an LBAF, then the framework can be represented through a function $\mathcal{L}: A \rightarrow \{\text{in}, \text{out}, \text{undec}\}$. The function $\mathcal{L}$, must be an injection (i.e. $\forall a \in \mathcal{A}, \exists L(a) \in \{\text{in}, \text{out}, \text{undec}\}$).

Bipolar Argumentation Framework: A BAF is a triplet $(\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-)$, where as before $\mathcal{A}$ is a set of arguments, $\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$ is a binary relation called support relation, and $\mathcal{R}^- \subseteq \mathcal{A} \times \mathcal{A}$ is a binary relation called attack relation. Therefore, $\forall a, b \in \mathcal{A}$ if $(a, b) \in \mathcal{R}^+$ we say the argument $a$ supports argument $b$, equivalently $\forall a, b \in \mathcal{A}$ if $(a, b) \in \mathcal{R}^-$ we say the argument $a$ attacks argument $b$.

Quantitative Bipolar Argumentation Framework: QBAF is an extension of BAF and is a 5-tuple $(\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-, \tau, \sigma)$, where $\mathcal{A}$, $\mathcal{R}^+$, $\mathcal{R}^-$ are the same as in BAF, and $\tau: \mathcal{A} \rightarrow \mathcal{K}$ is a base score function. The function $\tau$ gives initial values to the arguments from a preorder set of numerical values $\mathcal{K}$, meaning that $\mathcal{K}$ is equipped with a function $<$, such that $\forall a, b \in \mathcal{K}$ if $a < b$ then $b \not< a$. Another important component of QBAF is the strength of an argument, which is defined by a total function $\sigma: \mathcal{A} \rightarrow \mathcal{K}$.

Probabilistic Bipolar Argumentation Framework: PBAF is the last AF we will mention, and is also an extension of BAF. PBAF is quadruple $(\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-, \mathcal{P})$ where $(\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-)$ is a BAF, and $\mathcal{P}$ is a probability distribution function over the set $PD = \{(\mathcal{A}', \mathcal{R}'^+, \mathcal{R}'^-) \mid \mathcal{A}' \subseteq \mathcal{A} \land \mathcal{R}'^- \subseteq (\mathcal{A}' \times \mathcal{A}') \land \mathcal{R}'^+ \subseteq (\mathcal{A}' \times \mathcal{A}') \cap \mathcal{R}^+\}$. The elements in $PD ((\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-, \mathcal{P}))$ called possible BAFs, are the possible scenario that may occur and are represented through a BAF which was extended with probabilities.

5 Argumentation and Explainability

Explainability serves a much bigger goal than just the desire of computer scientists to make their system more transparent and understandable. Apart from the fact that Explainability is an aspect that justifies the decision of an AI system, it is also a mandatory mechanism of any AI system that can take decisions which affect the life of a person\cite{core2006}. For example, by making an automated charge to our credit card for a TV show that we are subscribers of, or booking an appointment to a doctor that we asked our personal AI helper to make last month, among others. The European Union has defined regulations that obligate a system with this kind of characteristics to provide explanations over their decisions\cite{regulation2016}. Therefore, in this section we describe how Argumentation enables Explainability. We can see that explaining a decision with argumentation is not something that emerged in recent years but existed in Argumentation Theory from its beginning\cite{pavese2019}.

We are going to divide this chapter based on the most important baselines that can provide explanation through Argumentation; Decision Making, Justification, and Dialogues. Moreover, we will present some basic studies that establish the field of XAI through Argumentation. But first we offer a literature review on how agents can use Argumentation, and the Ethics that Argumentation should follow. We consider these two subsections important, in order to show how XAI can be implemented in an agent through Argumentation, and what principles the agent must follow in order to be considered unbiased.
5.1 Agents and Argumentation

Single agent systems and multi agent systems can be built upon various forms of logic, such as first order logic (Smullyan (1995)), description logic (Nute (2001)), propositional logic (Buvac & Mason (1993)), and ASP (Fitting (1992), Lifschitz (2019)). One could read the book of (Wooldridge (2009)) to see the connections of Single Agent Systems (SAS) and Multi Agent Systems (MAS) with first order logic and other forms of logic, as well as to take a glimpse to Agents Theory in general.

At this point, we are going to describe the most important studies that share Argumentation Theory and Agents Theory either in SAS or MAS. One of the first studies that addressed this issue was that of (Kakas & Moraitis (2003, 2006)), in which the authors presented an AF to support the decisions of an agent. They consider a dynamic framework where the strength of arguments are defined by the context and the desires of the agent. Also, the concept of abduction is used by the agents in this framework. When they are faced with incomplete information, the agents can make hypotheses based on assumptions. Another important aspect is that of the personality of an agent. Based on definitions from cognitive psychology the authors give to each agent its own beliefs and desires translated into the AF as preferences rules. For instance let the two arguments $a = \text{"I will go for football after work\"}$, and $b = \text{"Bad news, we need to stay over hours today, to finish the project\"}$. Obviously, these two arguments are in conflict, thus we need a meta-argument preference hierarchy. If, additionally we knew that argument $b$ is stated by the employer, and argument $a$ by the employee, then it would hold $\text{prefer}(b, a)$. We can see implementations of such theoretical frameworks in (Panisson et al. (2014), Panisson & Bordini (2016)), where the authors use the AgentSpeak programming language to create agents that argue over a set of specific beliefs and desires using description logic, as well as in (Spanoudakis et al. (2016)) where agents argue in a power saving optimization problem between different stakeholders.

Next, Amgoud provides two studies (Amgoud & Serrurier (2007, 2008)) where agents are able to argue and explain classification. They consider a set of examples $X$, a set of classes $C$, and a set of hypothesis $H$ which is governed by a pre-ordered relation $\leq$ that defines which hypothesis is stronger. Then, an example $x \in X$ is classified in a class $c \in C$ by the hypothesis $h \in H$, and an argument for this statement is formalized as $a = (h, x, c)$. It is easily understood that other hypotheses could classify the same example to other classes, thus when all the arguments are created an AAF is generated.

An important need for agents which use AFSs is the capability of understanding natural language, and performing conversations with humans (Kemke (2006), Liu et al. (2019)). Understanding natural language and having a predefined protocol for conversation eases the exchange of arguments (Willmott et al. (2006), Panisson et al. (2015)), allows the agents to perform negotiations (Pilotti et al. (2015)), be more persuasive (Black & Atkinson (2011), Rosenfeld & Kraus (2016)), and to explain in more detail how they reached a decision (Laird & Nielsen (1994)).

The recent study of (Ciato et al. (2015)) proposes an AAF for an MAS focusing on the notions of Explainability and Interpretation. The authors define the notion of interpreting an object $O$ (i.e., interacting with an object by performing an action), as a function $I$ that gives a score in $[0,1]$ with respect to how interpretable the object $O$ is to the agent, when the agent wants to perform an action. The authors consider an explanation as the procedure to find a more interpretable object $x'$ from a less interpretable $x$. Thus, when a model $M : X \rightarrow Y$ maps an input set of objects $X$ to an output set of actions $Y$, the model is trying to construct a more interpretable model $M'$. An AAF can then use this procedure to explain why a set of objects is considered more interpretable.

5.2 Argumentation and Ethics

An agent in order to be trustworthy when it argues about a topic, it needs to be unbiased. For an agent to be considered unbiased, it must not support only its personal interest through the argumentation dialogue with other agent(s). Personal interests are usually supported through fallacies, such as exaggerations of the truth, or with unethical/fake facts. For this reason, the notions of ethics and argumentation are closely related to the problem of tackling biased agents (Correia (2012)). Moreover, knowing the ethics that an
agent follows when it argues is crucial as it enhances the Explainability, by allowing the opposing party to see how the agent supports its personal interests, and to accurately create counter arguments.

The combination of ethics and argumentation to construct unbiased agents who achieve argumentational integrity (Schreier et al. (1995)), is mostly used for persuading the opposing agent(s) (Ripley (2005)). Ethics in argumentation is also implemented in legal cases to conduct fair trials (Czubaroff (2007)), and in medical cases for patients privacy (Labrie & Schulz (2014)).

The existence of ethics in argumentation is very important in decision making problems that have conflicting interests between the participants. Especially in scenarios where the proper relations are mandatory, Ethics in Argumentation becomes a necessity. The authors in (Mosca et al. (2020)), propose a model where an agent works as a supervisor over decision making problems where conflicting interests exist, for sharing content online. The agent takes into consideration the personal utility and the moral values of each participant, and justifies a decision. A similar study is (Langley (2019)) for more generic scenarios.

E-Democracy is an evolving area of interest for governments wishing to engage with citizens through the use of new technologies. E-Democracy goal is to motivate young persons to become active members of the community by voting over decisions for their community through web applications, and argue if they disagree with a decision that is at stake. It is easily understood that ethics is an important aspect in the argumentation dialogues of an e-Democracy application. Citizens and the government should not be biased only in favor of their own personal interest but for the interest of the community. More specifically, citizens should think if a personal request affects negatively the other members of the community, while the government should consider if the decision that it is proposing has indeed positive results to the community or is only good for the popularity of the members of the government. In (Cartwright & Atkinson (2009), Wyner et al. (2012)), we can see many web applications for e-Democracy while in (Atkinson et al. (2005a)) e-Democracy is used to justify the proposal of an action.

The idea of e-Democracy goes one step further with (Wyner et al. (2012b)), where the authors propose a model to critique citizens proposals. The authors use Action-based Alternative Translating scheme (Wooldridge & Van Der Hoek (2005)) to accept or reject the justification of a proposal, and automatically provide a critique on the proposal, using practical reasoning. The critique is in the form of critical questions which are provided by the Action-based Alternative Translating scheme. Similarly, (Atkinson et al. (2011)) uses AF with values to critique citizens proposals, and (Wardeh et al. (2013)) provide web based tools for this task.

5.3 XAI through Argumentation

In this section, we will present some important studies that lead to the conclusion that Argumentation Theory is one of the most suitable models to provide explanations to AI systems.

The studies of Fan and Toni are very important in this field, as they provide a methodology for computing explanation in AAF (Fan & Toni (2014)) and ABA (Fan & Toni (2015a)). In the former, the authors define the notion of explanation as: given an argument $A$ defended by a set of arguments $S$, $A$ is called topic of $S$, and $S$ is the explanation of $A$. Then, they call compact explanation of $A$ the smallest $S$ with respect to subset relation, equivalently verbose explanation of $A$ the largest $S$ with respect to subset relation. Moreover, based on the notion of Dispute Trees they provide another method of explanation, with respect to the acceptability semantics of Dispute Trees. A Dispute Tree is defined as follows: (i) the root element is the topic of discussion, (ii) each node at odd depth is an opponent, (iii) each node at even depth is a proponent, and (iv) there does not exist node which is opponent and proponent at the same time. Then, Admissible Disputed Trees are those that each path from the root to the leaf has even length, and the root argument is called acceptable. Furthermore, Dispute Forest is the set of all Admissible Disputed Trees. It is easily seen, that the set of Dispute Forest contains all the explanations for why an argument $A$ is acceptable. While in the next study (Fan & Toni (2015a)) the authors extend these definitions to ABA.

Next, two studies present a formalization on how to model Explainability into an AF, in the context of solving scientific debates. In (Seselja & Straßer (2013)), the authors consider the formalization of an explanation as an extra relation and a set in the AF. More specifically, given an AF $(\mathcal{A}, \mathcal{R})$ an explainable
AF is a 5-tuple \((A, X, R, R', \tilde{R})\) such that \(X\) is the set of topics that receive arguments from \(A\) as explanations, the relation \(R'\) defines that an argument \(a \in A\) is part of the explanation of an element from \(x \in X\), and finally \(\tilde{R}\) states that some arguments may not exist simultaneously in an explanation for a topic because their co-occurrence brings inconsistencies. Therefore, an explanation is a set \(E\) that contains all the arguments from \(A\) that are connected through \(R'\) with an element from \(x \in X\), and do not co-exists in \(\tilde{R}\). On the same principles is the formalization of \(\text{Sakama (2018)}\). These formalizations were used for Abduction in Argumentation Theory, to model criticism inherent to scientific debates in terms of counter-arguments, to model alternative explanations, and to evaluate or compare explanatory features of scientific debates.

Case Base Reasoning (CBR) is one of the most commonly used methods in providing explanations about decisions of an AI system. Many studies use CBR to provide explanations in combination with Argumentation. In \(\text{Cyras et al. (2016)}\), the authors use CBR to classify arguments to a set of possible options, and when new information is inserted to the Knowledge Base (KB) the class of the argument may change.

Imagine we have the options \(O_1 = \text{book this hotel}, O_2 = \text{do not book this hotel}\), and our criteria are that the hotel should be close to the city center and cheap.

We find the hotel \(H\) to be close to the city center. Then, we have an argument for booking hotel \(H\), but when we look at the price, we see that it is too expensive for us. Then, we have a new argument not to book the hotel \(H\).

We can understand that this method method is close to explanation through dialogues, where each step adds new information to the KB. Therefore, the authors also provide an illustration of their framework with Dispute Trees. Another study that uses CBR and Argumentation is \(\text{Cyras et al. (2019)}\), where the authors give a framework that explains why certain legislation passes and others not, based on a set of features. They use the features of: (i) The Type, if the legislation is proposed by the Government, Private Sector, etc, (ii) The Starting House Parliament (it is a UK study thus the authors consider the House of Lords and the House of Commons), (iii) The number of legislations that are proposed, (iv) Ballot Number, and (v) Type of Committee. Another CBR model that classifies arguments based on precedents and features, for legal cases, is presented in \(\text{Bex et al. (2011)}\). The authors use a framework that takes in consideration information from the KB in order to classify the argument. More specifically, given a verdict that \(a\) person stole some money, under specific circumstances, the system must classify the argument if the defendant is guilty or not. The system will search for similar cases in its KB to make an inference based on important features such as type of crime, the details of the legal case, the age of the defendant, and if the defendant was the moral instigator.

The aspect of explanation of query failure using arguments is studied in \(\text{Arioua et al. (2014)}\). More specifically, the authors elaborate on query failures based on Boolean values in the presence of knowledge inconsistencies, such as missing and conflicting information within the ontology KB. The framework supports a dialectical interaction between the user and the reasoner. The ontology can also construct arguments from the information in the ontology on the question. The user can request for explanations on why a question \(Q\) holds or not, and it can follow up with questions to the explanation provided by the framework.

5.4 Decision Making with Argumentation

Argumentation is highly related to Decision Making. In fact, it has been stated that Argumentation was proposed in order to facilitate Decision Making \(\text{Mercier & Sperber (2011)}\). The contributions of Argumentation in Decision Making are plenty, with the most important being support or opposition of a decision, reasoning for a decision, tackling KBs with uncertainty, and recommendations.

The problem of selecting the best decision from a variety of choices is maybe the most popular among studies that combine Argumentation and Decision Making. In \(\text{Amgoud & Prade (2009)}\), the authors present the first AAF for Decision Making used by MAS. They propose a two step method of
mapping the decision problem in the context of AAF. Firstly, the authors consider beliefs and opinions as arguments. Secondly, pairs of options are compared using decision principles. The decision principles are: (i) **Unipolar** refer only to the arguments attacks or defenses; (ii) **Bipolar** take into consideration both; and (iii) **Non-Polar** are those that given a specific choice (the opinion), an aggregation occurs, such that arguments pros and cons disappear in a meta-argument reconsideration of the AAF. Moreover, the authors test their framework under optimistic and pessimistic decision criteria (i.e., a decision may be more desirable or less than other), and decision making under uncertainty. Decision Making under uncertainty is also presented in (Amgoud & Prade (2006)), where the authors try to tackle uncertainty over some decisions by comparing alternative solutions. Pessimistic and optimistic criteria are also part of the study.

In Zhong et al. (2014), the authors define an AF that takes into consideration information from the KB to make a decision using similar decisions. The framework firstly parses the text of the argument and extracts the most important features (nouns, verbs). Then, it compares with the decisions in the KB, and returns the most similar decision, with respect to the quantity of common features. The framework can back up its decision with assumptions on how similar the two cases are, and uses arguments which were stated for the similar case in the KB. On the other hand, in (Zeng, Fan, Miao, Leung, Jih & Soon (2018)), the authors use a **Decision Graph with Context** (DGC), to understand the context, in order to support a decision. A DGC is a graph, where the decisions are represented as nodes, and the interactions between them (attacking and supporting) as edges. The authors map the DGC in an ABA by considering decisions as arguments and the interactions between them as attack and support relations. Then, if a decision is accepted in the ABA it is considered a good decision.

Decision Making and Argumentation are also used to support and explain the result of a recommendation system (Friedrich & Zanker (2011), Rago et al. (2018)). Recommendation systems with Argumentation resemble feature selection combined with user evaluation on features. A recommendation system is 6-tuple \((M, AT, T, L, U, R)\) such that:

(i) \(M\) is a finite set of **items**.
(ii) \(AT\) is a finite, non-empty set of **attributes** for the items.
(iii) \(T\) a set of **types** for the attributes.
(iv) the sets \(M\) and \(AT\) are pairwise disjoint.
(v) \(\mathcal{X} = M \cup AT\).
(vi) \(L \subseteq (M \times AT)\) is a symmetrical binary relation.
(vii) \(U\) is a finite, non-empty set of **users**.
(viii) \(R : U \times \mathcal{X} \to [-1, 1]\) a partial function of **ratings**.

Mapping the recommendation system to an AF is done after the ratings have been given by a variety of users. Arguments are the different items from \(M\), and positive and negative ratings to the attributes related to an item from \(M\) as supports or attacks.

Decision Making for MAS in an ABA is presented in (Fan et al. (2014)). The authors consider that agents can have different goals, and decisions hold attributes that are related to the goal of each agent. In their case the best decision is considered as an acceptable argument in the joint decision framework of two different agents. Moreover, the authors define trust between agents, meaning that the arguments of an agent are stronger than the arguments of others in the scope of some scenario.

### 5.5 Justification through Argumentation

Justification is a form of explaining an argument, in order to make it more convincing, and persuade the opposing participant(s). Justification uses means of supporting an argument with background knowledge, defensive arguments from the AF, and external knowledge. One important study in this field is (ˇCyras et al. (2017)), where with the help of ABA and Dispute Trees the authors show how easy it is to justify if an argument is acceptable or not, just by reasoning over the Dispute Tree. Similarly, (Schulz & Toni (2016)) provide two methods of Justification for a claim that is part of an ASP, both using correspondence between ASP and stable extensions of an ABA. The first method relies on **Attack Trees**. The authors
consider an Attack Tree as: given an argument \( A \), the root of the tree, and the children being the attackers of \( A \), iteratively each node in the tree has as children only its attackers. An Attack Tree is constructed for the stable extension of an ABA, and is using admissible fragments of the ASP. If the literals that form the argument are part of the fragment then the argument is justified. The second justification method relies on the more typical method of checking if there exists an Admissible Dispute Tree for the argument.

Preference rules are usually used to justify the acceptability of an argument. We can see such studies in \( \text{Melo et al. (2016), Cerutti et al. (2019)} \). Acceptability of an argument is easily explained through preference rules, due to the fact that preference rules are a sequence of preferences between logic rules. \( \text{Melo et al. (2016)} \) present preferences over arguments formed from information of different external sources by computing the degree of trust each agent has for a source. The authors define the trust of a source \( \phi \) as a function \( tr(\phi) \in [0,1] \). Given an argument \( A \) with supporting set \( S = \{ \phi_1, \ldots, \phi_n \} \) from different external sources, the trust of an argument is given by Equation (2), where \( \otimes \) is a generic operator (i.e., it could be any operator according to the characteristics of the problem we try to solve).

\[
tr(A) = tr(\phi_1) \otimes \ldots \otimes tr(\phi_n)
\] (2)

Moreover, the authors consider two different types of agent’s behaviors: (i) Credulous agents trust only the most trustworthy source (the one with the biggest score from \( tr \)), and (ii) Skeptical agents consider all the sources from which they received information. Their study was based on \( \text{Tang, Cai, McBurney, Sklar & Parsons (2012)} \), where the authors also define trust of arguments in MAS. \( \text{Cerutti et al. (2019)} \) designed the \( \text{ArgSemSAT} \) system that can return the complete labellings of an AAF, and is commonly used for the justification of the acceptability of an argument. \( \text{ArgSemSAT} \) is based on satisfiability solvers (SAT), and its biggest innovations are: (i) it can find a global labelling encoding which describes better the acceptability of an argument, (ii) it provides a method where if we compute first the stable extensions we can optimize the procedure of computing the preferred extensions, and (iii) it can optimize the labelling procedure and computation of extensions of an AAF, with the help of SAT solvers and domain specific knowledge.

Justification for Argument Based Classification has been the topic of the study in \( \text{Thimm & Kersting (2017)} \). The authors propose a method of justifying the classification of a specific argument, based on the features that it possesses. For instance, \( X \) should be classified as a penguin because it has the features \( \text{black, bird, not(fly), eatsfish} \). An advantage of using classification based on features is that it makes explanation an easy task.

One common method to justify an argument is by adding values to the AF. There are cases where we cannot be conclusive that either party is wrong or right, in a situation of practical reasoning. The role of Argumentation in a case like this is to persuade the opposing party rather than to lead to mutually satisfactory compromises. Therefore, the values that are added to an AF are social values, and whether an attack of one argument on another succeeds depends on the comparative strength of the values assigned to the arguments. For example, consider the argument \( A1 \) from \( \text{BBC} \), and the argument \( A2 \) from \( \text{Fox News} \).

\[
A1 = \text{The weather tomorrow will be shinny}
\]

\[
A2 = \text{The weather tomorrow will be rainy}
\]

Obviously, those two arguments are in conflict; yet, we wish to reach to some conclusion about the weather, even an uncertain one, in order to plan a road trip. In this case, adding social values to the AF will solve the problem. For instance, a naive way is to define a partial order by relying on an assignment of Trustworthiness: if we trust information arriving from BBC more than from Fox News, we can use this order to reach to the conclusion. Another, way is to take a third opinion and consider valid the argument that is supported by two sources. These ideas were implemented in an AF by Bench-Capon in \( \text{Bench-Capon (2003a, 2002)} \), where the author extends an AAF by adding values (AFV). Subsequently, AFVs were used to solve legal conflicts \( \text{Bench-Capon (2003b), Bench-Capon et al. (2005)} \), to infer inconsistency between preferences of arguments \( \text{Amgoud & Cayrol (2002a)} \), and to produce acceptable arguments \( \text{Amgoud & Cayrol (2002b)} \).
The notions of AFVs were extended by Modgil in Hierarchical Argumentation Frameworks (Modgil (2006a, 2006b, 2009)). Intuitively, given a set of values \(\{a_1, \ldots, a_n\}\), an extended AAF, (i.e., attacks and defence relations exist), is created for each value \(\{(A_1, R_1), \ldots, (A_n, R_n)\}\). \(A_i\) contains arguments whose value is \(a_i\) and \(R_i\) attacks which are related to the arguments of \(A_i\), for \(i \in \{1, \ldots, n\}\). This mapping helped to accommodate arguments that define preferences between other arguments, thus incorporating meta level argumentation based reasoning about preferences at the object level. Additionally, the studies of (Coste-Marquis et al. (2012a, 2012b) and (Bistarelli et al. (2009)) depict more accurately how the social values are translated into numerical values in a single AAF, extending the studies of (Dunne et al. (2009, 2011)).

AFVs had a significant impact contributed significantly to practical reasoning, i.e., reasoning about what action is better for an agent to perform in a particular scenario. The authors in (Atkinson & Bench-Capon (2007)) justify the choice of an action through an argumentation scheme, which is subjected to a set of critical questions. In order for the argument scheme and critical question to be given correct interpretations, the authors use the Action-Based Alternating Transition System as the basis of their definitions. The contributions of AFVs is for the justification of an action, to show how preferences based upon specific values emerge through practical reasoning. The authors use values in the argumentation scheme to denote some descriptive social attitude or interest, which an agent (or a group of agents) wish to hold. Moreover, the values provide an explanation for moving from one state to another, after an action is performed. Therefore, values in this argumentation scheme obtain a qualitative, rather than a quantitative meaning. Two extensions of this study are (Atkinson & Bench-Capon (2007)) and (Atkinson & Bench-Capon (2018)). In the former, the agent must take into consideration the actions of another agent when it wants to perform an action. While in the latter, the agent must take in consideration the actions of all the agents that exist in a framework. An implementation of this argumentation scheme for formalizing the audit dialogue in which companies justify their compliance decisions to regulators can be found in (Burgemeestre et al. (2011)).

5.6 Dialogues and Argumentation for XAI

Explaining an opinion by developing an argumentation dialogue has its roots in Argumentation Dialogue Games (ADG) (Levin & Moore (1977)), which existed long before Dung presented the AAF (Dung (1995)). These dialogues occur between two parties which argue about the tenability of one or more claims or arguments, each trying to persuade the other participant to adopt their point of view. Hence, such dialogues are also called persuasion dialogues. Dung’s AAF (Dung (1995)) enhanced the area of ADG, and helped many scientists to implement the notions of ADG in an AF (Bench-Capon et al. (2000), Bench-Capon (1998), Hage et al. (1993)).

Nevertheless, the AAF of Dung helped only to some extent, because it could not capture all the aspects of an ADG. For instance, Dung’s AAF could not capture the notion of a clear reply structure, where each party waits for its turn in order to place a new argument. Henry Prakken identified this disadvantages and introduced a new AF which could capture all the aspects of an ADG (Prakken & Sartor (1998), Prakken (2005a)). More specifically, the author constructed an AF with a clear reply structure, where each dialogue move either attacks or surrenders, following a preceding move of the other participant, and allows for varying degrees of coherence and flexibility when it comes to maintaining the focus of a dialogue. Moreover, the framework can be implemented in various logics.

Subsequently, Verheij (2003) implemented the ideas of ADG and constructed two argument assistance tools, to guide the user in the computation of arguments. The author considers a context of argumentation in law. Moreover, the author uses defeasible reasoning, meaning that each argument no matter how commonly accepted it is, it can be questioned. Similar studies where defeasible argumentation is used are (Gordon (1993)) and (Loui & Norman (1995)).

McBurney and Parson in (McBurney & Parsons (2002)), offered a review on the protocols that ADG have in MAS, classifying them based on the task they intent to solve, which are: (i) Information Seeking Dialogues, where one participant seeks the answer to some question(s) from another participant, who is believed by the former to know the answer(s), (ii) Inquiry Dialogues, where the participants collaborate
to answer some question(s) whose answers are not known to any party, (iii) *Persuasion Dialogues*, which involve one participant seeking to persuade another to accept a proposition she does not currently endorse, (iv) *Negotiation Dialogues*, where the participants bargain over the division of some resource, (v) *Deliberation Dialogues*, where agents collaborate to decide what action or course of actions should be adopted in some situation, and (vi) *Eristic Dialogues*, where participants quarrel verbally as a substitute for physical fighting, aiming to vent perceived grievances. An extension of this study is (McBurney & Parsons (2009)), where the syntax and semantics in these protocols are analyzed, to help software engineering specification, design an implementation in MAS. In the latter study, McBurney comes to two important conclusions. Firstly, people or agents in a dialogue have an ostensible purpose, but their own goals or the goals of the other participants may not be consistent with this purpose. Secondly, both humans and agents involve mixtures of the dialogue protocols when they are in a dialogue. Analysis of protocols to purchase negotiations using ADG is also the topic of (McBurney et al. (2003)). Close to the aforementioned studies is (Atkinson et al. (2005b)), where a protocol for ADG is presented, which enables participants to rationally propose, attack, and defend, an action or course of actions.

Using Argumentation Based Dialogues (ABD) to explain an opinion is maybe the most natural method of providing an explanation (Kraus et al. (1998), García et al. (2007), Girle et al. (2003), Lúes-Argentina (2008), Lucero et al. (2009)). Dialogues are the most common way of displaying arguments, but it is not an easy task to define semantics that need to be followed by agents, in order to find the winning participant. Nevertheless, many attempts have been proposed, in order to make SAS and MAS more explainable, using ABD. Usually, ABD are performed on a strict set of rules, or otherwise known as protocols, which supervise the procedure of conversation by defining: whose turn is to speak; what knowledge can be used; when a conversation ends or begins; who the winner is; and what type of arguments must the agents use. But even these are not flawless, when faced with domain specific information. Studies, such as (Cogan et al. (2005)) indicate that some reconsideration should be applied on the protocols when the agent(s) are faced with domain specific knowledge.

One study that addresses such technicalities is (Panisson (2019)), in which the author presents MAS as an organization-oriented paradigm, where social relationships are considered. The author considers various organization models that share the characteristics, such as: (i) agents use a common language, (ii) agents are characterized by roles, (iii) explicit representation between roles exists, and (iv) activities can be either decomposed and asserted to individuals or can be solved as a whole. Furthermore, due to the nature of MAS the author defines preferences between the opinions of agents in a dialogue, by the level of authority. Social relations and Argumentation is also the topic of (Liao et al. (2018)), for action selection of a robotic platform.

The Hilton (1990) conversational model, which was extended by (Antaki & Leudar (1992)) from dialogues to arguments, shows that many statements made in explanations are actually argumentative claim supporters; that is, justifying that a particular cause holds (or does not hold) when a statement is made. Therefore, explanations are used to support claims, which are arguments. The authors extend the conversational model to a wider class of contrast cases, as well as explaining causes. Thus, explanations extend not just to the state of affairs external to the dialogue, but also to the internal attributes of the dialogue itself. These notions where supported by (Slugoski et al. (1993)).

In (Bex et al. 2012, Bex & Walton 2016), the authors consider a different approach for Argumentation and Explainability. They state that Argumentation and Explainability should consider two different aspects in an ABD. More specifically, Argumentation should only play the role of opposing the opinion of another participant, while Explainability should provide evidence to support an argument. The authors consider that this distinction can help an agent restrict the range of possible answers in an argumentation dialogue, because they will have to choose between a set of explanations or arguments in responses. Moreover, the authors demonstrate such concepts using the Argumentation Interchange Format (AIF) (Chesnuevar et al. 2006). A similar approach is found in (Letia & Groza 2012), where the authors propose an AF for MAS, in which the notion of arguments is separated from the explanation. The authors provide a formalization for the agents to understand evidence that is needed to support an argument, and
formalization to understand and explain why an event has occurred. The authors use their method in an ABD, where the aforementioned components are extracted from free text.

As mentioned, formalizing AFs to support dialectical explanation is not an easy task and this was understood from the beginning in the research of Explanation through Dialogues. Therefore, many studies propose different semantics on how to constrain Argumentation Theory into a dialogue. The study (García et al. (2013)), in the context of description logics the authors propose a method using Dispute Trees, where the topic of conversation is the root argument, and each level of the tree represents a step in the conversation. The authors also provide a method that justifies an argument $A$ through its supporting arguments, the arguments that support $\neg\neg(A)$, and the arguments that attack $A$.

Finally, a study that provides an explanation of non-acceptable arguments with Dispute Trees (Fan & Toni (2015b)). The core idea of the study is that given an AF $= (A, R)$, a non-acceptable argument $a \in A$ can become acceptable if a specific set of arguments $A \subseteq A$ and a set of attacks $R \subseteq R$ are removed from the framework. The authors call the set $A$ argument explanation of $a$, and the set $R$ attack explanation of $a$. The contribution of this paper is very important, due to the minimality of the definitions, meaning that only a specific set of arguments and attacks if removed will make an argument acceptable. Moreover, an agent which can use these semantics can understand which arguments must be attacked in order to make an argument acceptable, in a dialogue.

6 Argumentation and Explainable Systems

Argumentation from its beginning helped in the development of explainable systems, and today is becoming one of the most important reasoning methods to capture Explainability. We can see this in this section, if we consider how many AI systems choose Argumentation as their reasoning mechanism to enhance Explainability. We capture the topics of Argumentation in Medical Informatics, Law, the SW, Security, and some General Purpose AI systems (i.e., systems that can be implemented in various fields). Moreover, we include a subsection about Robotics, in which not all of the studies follow the strict rules of an AF, but they use arguments to gain the trust of the user, or to perform an argumentation dialogue, or even to help at cooperative decision making between humans and robots.

6.1 General Purpose Argumentation Systems

As General Purpose Argumentation Systems should be considered all the systems that use Argumentation to explain a result. These systems are not restricted to a specific scientific field. Systems like these could be implemented in various scientific fields only by changing the KB. One of the most well known frameworks based on SAF is ASPIC (Modgil & Prakken (2014), Dauphin & Cramer (2017)) which can provide explanations and arguments based on mathematical deduction.

Gorgias is an argumentation framework (Kakas & Moraitis (2003), Noel & Kakas (2009), Spanoudakis et al. (2016b)), where the arguments are represented in the form of rules; also it supports the notion of preference between logic rules and assumptions, when we have the case of missing information. Gorgias is implemented in Prolog and it can explain if an argument is acceptable using the Dispute Tree. Later versions of Gorgias, called Gorgias-B offer a friendlier user environment with a GUI in Java, where the user can give a set of arguments and facts in the form of logic rules, and make questions for the acceptability of any argument. Gorgias can support decision making in more advanced scenarios, such as choosing the best policy when we have an iterative set of restrictions. In detail, a primary decision based on some restrictions will lead to a specific set of new decisions that are governed again by restrictions, and iteratively the new choice will lead to a new set with restrictions, until we reach a final decision (Kakas et al. (2019)).

Two tools that use visualization to help empower the explanation capabilities of an AF can be found in (Betz et al. (2019), Green et al. (2019)). The first one is OpMap, a tool for visualizing large scale opinions
spaces into a geographical map. The goal of OpMap is to build a tool that can handle multi-opinion against or in favor of a topic, based on a SAF. OpMap first clusters the opinions using the GMap algorithm, which extends force directed algorithms and constructs visualizations resembling a geographic map. After the clusters are created, OpMap maps the opinions-arguments and the clusters into a 2D map. The clusters are considered as columns and each argument as a row. Then, the arguments are given the label of True, False or Judgement Suspension (i.e., Undefined). The second one is called AVIZE, and is a tool for constructing arguments in the domain of international politics. AVIZE tries to create a triplet \((\text{topic}, \text{claim}, \text{premise})\) for each argument and it represents this as a table with a column for each component of the triplet. Obviously, AVIZE is a tool that helps to understand the structure of an argument and make it easier for the user to find ways that can attack or support it. Moreover, AVIZE is a supervised tool where the user must evaluate the quality of the triplet, and it can be used in other domains.

An AF that evaluates alternatives and explains them with a scoring function, coupled also with a visual representation, is presented in (Baroni et al. (2015)). The framework is based on a QBAF and can support argument dialogues. The authors consider a scoring function \(SF\), which takes into consideration the base score of an argument \(a\), and the sum scores of its attackers and supporters. This will help the debater to use specific arguments based on the final score. For example, if we know that \(SF(a) > SF(b)\), it is wiser to use argument \(a\) instead of \(b\). The framework is also compared with an AAF in terms of finding the appropriate argument to win a conversation, and is projected to a visual representation using Visual Understanding Environment (VUE).

Another important aspect when we provide an explanation is the level of trust we have for an argument. ArgTrust (Sklar et al. (2015)) is maybe the most well-known system for quantifying the trust of an argument, to facilitate decision making. The system given an extended AF, and a set of weights for the relations between arguments, computes the trust level of the arguments in the framework by taking into consideration the negative and positive impact the other arguments have. Argument trust is also addressed in (Tang, Sklar & Parsons (2012)). Given a MAS, the authors construct an AF that can return the trust level of any argument. The framework is based on LBAF and the agent has to elaborate why to support an opinion, by displaying the premises it relies on. The framework offers an implementation in Java that infers the acceptability or non-acceptability of an argument, which is derived based on the beliefs of the agents and relations between the arguments.

### 6.2 Law

Argumentation and Law are strongly connected even outside the scope of Computer Science, because Law could be considered as a discipline in which a lawyer tries to obtain knowledge that will help him oppose each argument against his client. Henry Prakken was the first computer scientist that managed to define formalizations for AF used in Law. Among his many studies, the most relevant to this survey are (Prakken (2005a), Prakken et al. (2015), Prakken (2017)). In the first two studies, the author provides a formalization of a legal case into Argumentation Theory. The legal case is firstly given to the ASPIC framework that tries to produce defeasible rules, which are considered as arguments. The procedure of translating logical rules into arguments is identical to (Caminada et al. (2015)). After this, we can easily compile an AF. In the last study, the author proposes to represent legal cases as Dispute Trees, over description logics. Of similar nature is the study of (Bench-Capon (2020)), where the impact of AAF in AI and Law is discussed.

We can easily envision the dialogue that takes place during a trial, as an argumentation dialogue, where a lawyer provides arguments in favor of its client and the opposing lawyer against him, and this process goes on until one lawyer cannot issue any more arguments. This was the idea in (Al-Abdulkarim et al. (2016a, 2019, 2016b)), where the argumentation dialectical procedure is projected into a Dispute Tree. On the other hand, (Sklar & Azhar (2015)) consider legal reasoning as an interchange between two or more agents based on LBAF. The most important aspect of this study is the \textit{meta level argumentation semantics}. The authors provide a set of logic rules with preferences that can be implemented in a LBAF.

https://vue.tufts.edu
and change the procedure of how an admissible argument is derived. The preferences are called Social Biases and capture the notion of: (i) Argument from Authority, meaning that the decision of an agent might be stronger than others, (ii) Epistemic Entrenchment, when an agent is totally sure about its opinion and nothing will change it, (iii) Stereotyping, when an agent makes assumptions about the beliefs of another agent, and (iv) Defiance, when an agent constructs arguments from propositions which are against its beliefs.

Interesting is the study of Zhong et al. (2019), in which the authors present a framework that is meant to help judges to define a sentence. Their framework compares current cases with past performed cases and returns the verdicts that were taken for similar cases. Also, the framework provides arguments and explanations to the judge, by showcasing the Redundant Attributes. Redundant Attributes are the parts in the description of a case that may contribute to the verdict, such as: the type of crime, amount of stolen items, condition of the accusant, crime evidence, number of abettors, and if the accusant is also the moral instigator.

The implementation of Argumentation in the field of AI and Law has resulted into many software systems with important Explainability capabilities. For instance, TAXMAN (McCarty, 1976), which can develop argumentation dialogues in favor or against a side in a specific legal case (Eisner v Macomber Clark, 1919), HYPO (Rissland & Ashley, 1987), CATO (Aleven, 1997) and ANGELIC (Al-Abdulkarim et al., 2016c), which are used for legal case-base reasoning, and Gordon & Walton (2009), which uses the Argumentation schemes of Walton (2005), in order to construct and search for legal arguments.

6.3 Medical Informatics

Argumentation has emerged recently in the field of Medical Informatics, mostly to support the decision of AI systems. For instance, in Longo & Hederman (2013) Defeasible Reasoning and Argumentation Theory are used for Decision Making in a health care scenario, using data about Breast Cancer. The study shows how to translate clinical evidence in the form of arguments, add support and defeat relations, and apply Defeasible Reasoning. The authors represent arguments as rules, and take the clinical evidence as support. Using this representation, defeat can be derived through undercut and rebut attacks. Another similar approach is Spanoudakis et al. (2017), where the same notions are used to determine the level of access to a patient’s medical record. Moreover, the clinical evidence and medical record of an admissible argument are given as an explanation. Looking at Mozina et al. (2007), Chapman et al. (2019), Kokciyan et al. (2019), Kokciyan et al. (2018), Sassoon et al. (2019), we can find a decision making AF for patients that suffer from chronic diseases to help them decide how they can prevent worsening their health. More specifically, the framework involves sensors that record the health state. Then, if an anomaly in the blood pressure occurs, the framework can recommend a treatment, by computing an argument based on embedded data which was bounded with the input given by the anomaly in the health recorder. In Cyras et al. (2018), Cyras & Oliveira (2019), Oliveira et al. (2018), the authors translate clinical evidence into arguments for an ABA, and track patients health state to suggest a treatment in an emergency scenario. Due to the probability distribution which is part of the framework, an ABA which can handle uncertainty can tackle the demands of stochastic framework.

Argumentation and decision making are also presented in Qurat-ul-ain Shaheen & Bowles (2020). In this study, decision making through argumentation dialogues is used in order to recommend a treatment to patients with multi-morbidity (i.e., multiple chronic health conditions). The complexity for making a decision in this case is high, as a treatment for a chronic disease may affect another chronic disease. Nevertheless, the authors propose a novel approach to justify a decision with Satisfiability Modulo Theories solvers in an interactive way through argumentation-based dialogues. Moreover, the authors provide two different ways of explanation: the first one is called passive, where the patient accepts all the arguments which point to a specific treatment, while the other one is called active, where the patient can ask why a treatment is recommended.

Argumentation in Medical Informatics is also used to justify a decision. Donadello et al. (2019) use an OWL ontology with information about the dietary habits that a patient with chronic diseases should
The information is recorded into the KB of the framework by a domain expert. The framework keeps track of the daily meals of the patient and if it finds inconsistencies, it notifies the patient in the form of arguments. More specifically, predefined SPARQL templates exist that get for example what the user ate, and are addressed to the ontology. For instance, if a patient is diagnosed with diabetes, eating food with high sugar consistency will trigger an argument that explains why the user should not eat so much sugar. Very similar is the study of Grando et al. (2013); the authors also provide a medical purpose mechanism that receives data about the health anomaly that occurred. Then, it addresses a SPARQL query to an underlying OWL ontology and returns an explanation in the form of arguments and the supporting facts. The ontology is constructed by domain experts and the framework offers a GUI to receive data.

Two similar studies in argument classification for medical purposes are Mayer et al. (2018), Prentzas et al. (2019). In the former, the authors try to classify arguments, by finding evidence and claims from free text, which can be useful information to an expert, in order to lead him to a potential treatment. Moreover, the authors created their own annotated corpus with arguments called Random Clinical Trials. For the evaluation, they use three methods for classifying the arguments in the texts: (i) SubSet Tree Kernel (ii) SVM with Bag of Words features weighted by TF-IDF, and (iii) a Kernel mechanism that combines (i) and (ii). While in the latter, they propose a methodology for applying Argumentation on top of ML to build an XAI system for stroke prediction. The authors trained a Random Forest ML model on the Asymptomatic Carotid Stenosis and Risk of Stroke Study dataset (Nicolaides et al. (2010)). Then, the produced rules are extracted in the form of IF-THEN statements, and are given to Gorgias. Gorgias, provided with a set of facts and arguments represented as logical rules, generates an explanation through the Dispute Tree, if an argument is admissible. Zeng, Miao, Leung & Chin (2018) propose an explainable model based on Argumentation for detection of dementia. The framework uses a Convolutional Neural Network (CNN) to extract features from images, such as the size, the region, and the drawing test performance. The framework can then explain its decision through arguments using these features and the medical history. The explainable model is a combination of a graphical representation for modeling decision problems in various contexts, and a reasoning mechanism for context-based decision computed with an ABA.

6.4 Robotics

The recent evolution in the field of Robotics, which makes it easier for people to deploy a robotic assistant in a household environment, increased the urgency of making the robotic platforms decisions totally explainable. Argumentation has been used in several cases in the field of robotics, to explain decisions. The tasks of trust gaining, persuasion, and combined decision making between a human and a robot are the main reason for which a robotic platform will use an AF to explain its decisions.

Argumentation for shared decision making between a robotic platform and a human is the topic in Sklar & Azhar (2015), Azhar & Sklar (2016, 2017), Sklar & Azhar (2018). To the best of our knowledge, these were the first studies on argumentation-based dialogue games between a human and a robotic platform. The authors developed an LBAF along with a GUI that helps humans in cooperation with a robotic platform to reach a shared decision at each step of an activity, more specifically in a Treasure Hunt Game (Sklar & Azhar (2015)). Next, in Azhar & Sklar (2016, 2017), Sklar & Azhar (2018), the authors extent the Treasure Hunt Game with different methodologies for implementing multiple types of argumentation-based dialogues. The framework can explain which dialogues are appropriate given the beliefs of the participants and how multiple dialogues can occur simultaneously while containing consistency in the general dialogue. Interesting is the fact that they manage to define conditions for when the two participants are in a state of agreement, disagreement, one side lacks knowledge, and both lack knowledge. Additionally, the dialogue protocols are separated into three classes: (i) Persuasion, where one of the two participants tries to convince the other one about its beliefs, (ii) Information Seeking, when one participant tries to extract information from the other one, and (iii) Inquiry, when one participant asks information that the other participant does not have. The last dialogue protocol might seem odd, but the idea behind it is that even though one participant might not know the correct answer, he might give some secondary information which might be useful.
Next, we elaborate on two studies that can be considered as argumentation dialogue frameworks. In (Modgil S (2018)), a chatbot is used to collect arguments, counterarguments and supporting arguments from users over a large variety of topics. The robotic platform can then recall past received arguments to explain to the user its opinion on a topic. The biggest benefits of this system are: (i) The ability of capturing the probability of different people to provide same or similar arguments on a topic, (ii) The ability of initiating argumentation with zero knowledge about the topic, and (iii) The ability to identify and use a counter-argument for similar arguments. Similar arguments are considered those that use the same evidence to support or attack a claim.

On the other hand, the chatbot in (Wanner et al. (2017)) uses an OWL ontology to perform advanced dialogue planning, much like an AF which can help choose specific responses to elaborate and possibly justify a topic. More importantly, the study addresses the problem of recommendation with arguments. The idea behind it is the following: from the moment that we reached a point where the explanations exist for each system’s decision, we can trust it to give us recommendations. The robotic system receives verbal as well as visual signals, such as gestures. After textual parsing and visual inferencing occurs, the robotic system tries to understand: (i) The form of conversation it is having (i.e., if the user wants explanation, recommendation, or information about something), or (ii) If it needs to return information about the question using the Statistical Speech Model Vocapia\(^7\) from its internal KB. Very similar is the recent study of (Meditskos et al. (2019)), with one significant extension; the authors use web repositories such as DBpedia\(^8\) and WordNet\(^9\) to enrich the KB of the system.

Close to the two aforementioned studies are (Torres et al. (2019), Fischer et al. (2018)), in which the authors use argumentation in the form of recommendation. These studies are fully developed robotic platforms that can also perform other tasks, but it is interesting to see how the robotic platform uses external knowledge, information from its sensors, and information in its internal KB, to compute an argument. In (Torres et al. (2019)), the authors use preference in a non-monotonic KB with a closed word assumption. Also, the conflicting information which comes from non-monotonic logic is tackled with the principle of specificity according to which between two conflict propositions we shall always choose the more specific one (i.e., the one with more information). One example should make things clear, on how the recommendation in form of arguments is constructed:

When the user asks for a Coca-Cola, the system starts looking in the environment for objects that can be classified as a can or bottle of Coca-Cola, but it already knows that the user has diabetes and today he already had a lot of sugar, so it recommends bringing him tea instead.

In Fischer et al. (2018), the authors use a static ontology and external knowledge from WordNet and Wikidata\(^10\) to recommend tools to perform an action. The framework first tries to understand what activity the user wants to perform, for example “I want to cut this wooden block”, and then it annotates the action cut and the object on which he wants to perform the action on, in this case wooden block. Then, the framework is asked to give an argument recommendation on which tool is the most appropriate for the action from a collection of objects. The framework uses the internal knowledge from its ontology for actions and objects, Wikidata for hierarchical relations and WordNet for synonyms, to infer the desired tool. It can interact with the user via speech and even provide an explanation containing a log file with the internal steps it performed to reach this recommendation.

### 6.5 Semantic Web

In this subsection, we will describe some argumentation systems that use SW technologies, or try to solve problems that are related to the Web through Argumentation.

\(^7\)http://www.vocapia.com/
\(^8\)https://wordnet.princeton.edu
\(^9\)https://wordnet.princeton.edu
\(^10\)https://www.wikidata.org/wiki/Wikidata:Main_Page
One of the first studies in this domain concerns an OWL ontology, which is used to annotate arguments in natural language and to provide explanation on how the arguments were annotated (Rahwan et al. (2011)). The authors use the AIF to understand the structure of arguments in text. Other important aspects are that the authors manage to capture the support relations between arguments and that the ontology can automatically create new classes which take as instances arguments with specific structures. AIF is also used in the study (Indrie & Groza (2010)) to model interaction between arguments. Semantic MediaWiki (SMW) is used to translate arguments in specific templates for coherency in their structure. Next, in order to exploit the ontology of AIF in the SMW, the authors map the concepts and roles from the ontology to the internal structuring mechanisms available in SMW. This study offers a method of explaining how an argument was classified based on: (i) The structure of the argument, (ii) Patterns (for instance the opinion of experts), (iii) Relation of specific Wikipedia terms, (iv) Domain arguments that support a specific topic in the ontology KB, (v) Support level, and (vi) Context.

An approach to empower commonsense reasoning and make it more explainable with Argumentation is given in (Botschen et al. (2018)). The authors investigate whether external knowledge of event-based frames and fact-based entities can contribute to decompose an argument as stated in the Abstraction and Reasoning Corpus (ARC) task. In the ARC task the system must find the correct cause that derives a claim given some data. It is similar to finding the warrant of an argument in Hunter’s argumentation model (Besnard & Hunter (2001, 2009)). The outline of the study is that the authors use a Bi-LSTM trained on ARC KB to annotate the frames and entities. FrameNet is used to make semantic connections between frames and Wikidata to offer more information for the entities. The difference between frames and entities is that entities in a sentence could only have the role of subject and object, while frames can be any part of speech.

Online dialogues can be a rich source of argumentation dialogues and researchers are interested in using such repositories, to give explanations on how people consider an acceptable argument (Snaith et al. (2010), Reed et al. (2010)). These types of repositories are the most precious in understanding how Argumentation Theory works in real life, because they give data that are in the purest form. Using the Web as a source of arguments was originally introduced in (Bex et al. (2013)), where Bex envisioned the Argument of Web, a Web platform combining linked argument data with software tools that perform online debates. Inspired by the aforementioned study, a mechanism which attempts to find arguments that are against or in favor of a topic in a conversation was built (Boltužić & Snajder (2014, 2015)). The mechanism relies on an SVM trained on the custom made ComArg corpus (Boltužić & Snajder (2014)), a dataset of arguments supported by comments, using the data of ProCon and Idebate. Then, the system is evaluated on data from ProCon and Idebate. Additionally, the authors annotate the similarity of arguments using Bag of Words, Skip-Gram, and Semantic Textual Similarity methods. Identical is the study (Swanson et al. (2015)), where the authors train three different regressors: (i) Linear Least Squared Error, (ii) Ordinary Kriging, and (iii) an SVM on a dataset created from CreateDebate to identify arguments in online dialogues on the topic of gay marriage, gun control, and death penalty. A powerful tool for visualizing a BAF can be found in the web application of (Baroni, Borsato, Rago & Toni (2018)), where the user can build an argumentation graph and see the justification of an argument.

6.6 Security

Preserving the user’s privacy is maybe the most difficult task that computer science has to solve, as having users that do not trust the systems will make them reluctant to use digital products or services. Especially with the rise of the Internet of Things, more and more devices are connected to each other, and need to communicate and collaborate. Such a setting makes the devices even more vulnerable to an

\[1\] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki
\[2\] https://competitions.codalab.org/competitions/17327
\[3\] http://framenet.icsi.berkeley.edu/indrupal
\[4\] https://www.procon.org
\[5\] https://idebate.org
\[6\] https://www.createdebate.com
Argumentation and Explainable Artificial Intelligence - A survey

Argumentation has been used in recent years to provide security in such cases, mostly because Argumentation can use persuasion in order for an agent to understand if it receives an attack (Rowe et al. (2012), Murukannaiah et al. (2015), Santini & Yautsiukhin (2015), Panisson et al. (2018)).

In (Kökcüyan et al. (2017), the authors use an ABA, for agents to decide whether they should share the content of a file. The authors consider that when the content of file is shared by a user, other users who might get affected must agree in order for the file to be uploaded, otherwise privacy constraints might be violated. In order for this to be achieved, the authors model the users as agents in a social network and represent their users’ privacy constraints as semantic rules. The agents can argue between them on propositions that enable their privacy rules using assumptions from their personal ontology. Also, agents can ask for help from other agents, in order to enrich their ontology. This study offers a MAS with personal ontologies for each agent that contain domain knowledge, and semantic rules which describe privacy constraints such that the knowledge can be used to perform argumentation. Moreover, it offers an algorithm that allows agents to carry a dialogue such that each agent can attack the assumptions of others. Kökcüyan & Yolum (2017) is an extension in the context of the Internet of Things. On the same principles for accessing patient’s data is the study (Bassiliades et al. (2018)), which is based on the studies of (Spanoudakis et al. (2016a, 2007)).

Similar to the previous studies are (Fogues et al. (2017a,b), Shakarian et al. (2015)), where the users with the help of Argumentation find a sharing policy, when conflicting interests between many users exist. The authors in (Fogues et al. (2017a,b)) develop a computational model that understands how people decide the appropriate sharing policy in multi-user scenarios where arguments exist, and predicts an appropriate sharing policy for a given scenario. In (Shakarian et al. (2015)), the authors offer a framework for recommending sharing policies.

In (Karafili et al. (2017), Karafili, Wang, Kakas & Lupu (2018), Karafili et al. (2020), Karafili, Spanaki & Lupu (2018)), the authors propose a novel argumentation-based reasoner for analyzing cyber-attacks, to help the cyber security analysts understand from where a cyber attack came. The framework gives possible culprits of the attack and hints about missing evidence. The AF that is used is a SAF and the Argumentation based reasoner, which is part of their framework, is taking into consideration the logic rules with their preferences, as well as the social biases (or background knowledge) that may occur with the cyber attack. The running examples in the studies give a good understanding of the social biases: two countries that have some conflicting financial interests are more likely to perform a cyber attack against each other, therefore indicating a potential source of the cyber attack. The same principles were used in (Karafili et al. (n.d.)) for decision making over actions in drones.

The studies of Nunes et al. (Nunes, Shakarian, Simari & Ruef (2016), Nunes, Shakarian & Simari (2016)) are very interesting as they propose an argumentation model based on defeasible logic programming, designed to help the analyst find the source of a cyber-attack. These studies are based on (Nunes, Kulkarni, Shakarian, Ruef & Little (2016)), where the authors construct a dataset with real life cyber-attack scenarios from hackers, collected from the DEFCON competition. The studies of Nunes et al. are some of the few which are tested in real life cyber-attack cases. The experiments show that using argumentation can significantly reduce the number of potential culprits that the analyst must consider, and that the reduced set of culprits, used in conjunction with classification, leads to improved cyber-attribution decisions. In the same field are the studies of (Genitsaridi et al. (2013), Bikakis & Antoniou (2010)), which are based on the principles of defeasible logic programming and offer a high level authorization language for access control policies.

Finally, (Bandara et al. (2006, 2009)) use Argumentation and preference reasoning for firewall policy analysis, and automatically generate firewall policies from higher-level requirements. These studies managed to show that the non-monotonic reasoning with conflict rules that Argumentation offers, permits the analysis and generation of anomaly free firewall configuration rules. The modularity of this policy generator framework is that it allows for customisation according to characteristics of a network, and the usage of deductive and abductive reasoning offers explanatory power which can trace the source of an attack. Additionally, the authors state that their framework can be used to: (i) Review a firewall...
configuration by querying the formal model for reachable nodes, (ii) Analyse a firewall configuration in order to detect anomalies, and (iii) Generate a firewall configuration according to the characteristics of a specific network.

Table 1 displays all the aforementioned systems. We classified the systems based on which task they tackle: Decision Making, Justification, Explanation through Dialogue, or Argument Classification.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Decision Making</th>
<th>Justification</th>
<th>Explanation through Dialogue</th>
<th>Argument Classification</th>
</tr>
</thead>
</table>

7 Argumentation and Machine Learning for Explainability

ML could be considered as the field that brought XAI to the surface, due to the fact that the results returned by data driven models were considered as “black boxes” whose rationale is incomprehensible to most human users. The new regulation established by the EU that any system which can take a decision that can affect our life must have the capability to explain its decision (Regulation (2016)), led to the need of new methods that can provide Explainability even in this field. As we can see in this section, Argumentation can become the link between ML and XAI.

Cocarascu & Toni (2018) introduce a deep learning method for argument mining to extract attack and support relations, in order to explain how news headlines support tweets and whether reviews are deceptive by analyzing the influence these texts have on people. Thus, the authors elaborate on the level of trustworthiness, persuasion, and explainability of arguments that exist in these texts. Exploiting the knowledge relations that hold between arguments units, carries great potential of explaining why an argument holds (or does not hold) when presenting with supporting or attacking evidence. The method could be considered a pipeline classification problem with an LSTM trained over the argument corpus AIFdb to capture whether two different texts support, attack, or are neutral among each other. Furthermore, the framework obtains state of the art scores over small datasets such as the Hotel Dataset

http://corpora.aifdb.org/
Argumentation and Explainable Artificial Intelligence - A survey

(Ott et al. (2013)). Due to the fact that the method is based on mining a BAF, the authors can automatically map the attack and support relations on a BAF.

Case Based Reasoning (CBR) or Instance-based Learning is one of the most commonly used methods of ML, which was used from the early days of the research in ML. Therefore, there are many studies which use CBR and Argumentation to achieve Explainability. In (ˇCyras et al. (2016a,b)), the authors use CBR to classify arguments to set possible options, and when new information is inserted to the KB the class of the argument may change. Another study that uses CBR and Argumentation is (ˇCyras et al. (2019)), where the authors construct a framework that explains why certain legislation passes and other not, based on a set of features (see Section 5). A CBR model that classifies arguments based on precedents and features, for legal cases, is also presented in (Bex et al. (2011)). The authors use a framework that takes in consideration information from the KB in order to classify the argument. More specifically, given a verdict the system must classify the argument if the defendant is guilty or not. Hence, it searches for similar cases in its KB to make an inference.

One of the first studies that use Argumentation and ML together with external knowledge from domain experts is (Možina et al. (2007)). In this study, the authors try to tackle the fact that domain expert opinion may not be global for the domain as there might exist exceptions under specific circumstances. Therefore, the experts explain with examples why a specific argument is acceptable or not under specific circumstances. The examples work as templates for the characteristics of the domain. More specifically, if specific properties hold in the environment then the reasoning for the acceptability of an argument will be different if other properties would exist. A data driven model tries to learn this relation between properties and different type of reasoning, in order to help an AF decide over the acceptability of an argument.

Three studies that use ML and Argumentation to explain a claim while relying on an external KB can be found in (Samadi et al. (2016), Potash et al. (2017), Habernal & Gurevych (2016)). ClaimEval is presented in (Samadi et al. (2016)), as a mechanism that given a specific topic extracts a set of supporting and attacking arguments from various websites with the help of Bing19. ClaimEval relies on a Probabilistic Logic that allows it to state and incorporate different forms of already existing knowledge. The authors take into consideration the credibility of the source by mapping each source into a graph. Then, the authors propagate the credibility of the graphs of different sources based on some prior knowledge, which is defined as a set of rules to reach joint source credibility. Also, the authors use an SVM to evaluate if evidence is supporting or attacking for a specific topic and achieve state of the art results at this classification task. On the other hand, in (Potash et al. (2017)) the authors present various data driven models which are trained to find the most convincing argument for a topic. These models are very useful, because they can explain how the form of an argument should be, in order to be considered convincing for a topic. The models are evaluated on the argument convincingness dataset UKPConvArg (Habernal & Gurevych (2016)). Furthermore, the authors give four supervised models that achieve state of the art results on the same dataset: (i) An Bi-LSTM that receives as input the vector of the concatenation argument-topic pairs, using Glove embeddings20 (ii) A method with Bag of Words given the term-frequency representation of each pair, (iii) A method using Bag of Words and term-frequency of the triplet (argument, topic, most related wiki article with respect to wiki metric), and (iv) A probability distribution between arguments and Wikipedia articles. Also, the authors provide the largest dataset of annotated arguments in Wikipedia articles.21

Another study that uses external knowledge from ConceptNet22 and DBpedia, NLP methods, and KB features to predict the type of relations between arguments is (Kobbe et al. (2019), similar to (Cocarascu & Toni (2018)). The authors classify the relation between two arguments A and B using a pretrained Bi-LSTM. The Neural Network receives the vector representations of the words that each argument is composed of and returns a vector representation for A (denoted by $emb(A)$) and B (denoted by $emb(B)$). The relation between the vectors is $r(A, B) = emb(A) − emb(B)$, where the operation $−$ is performed.
element-wise. External knowledge can be used to enrich the obtained representation \( r(A, B) \) with relevant information for knowledge relation about concepts and entities mentioned in the two argumentative units. If \( v_K(A, B) \) is the new vector with external features, then the authors add element-wise to get a new vector for the relation \( r'(A, B) = r(A, B) \oplus v_K(A, B) \).

Argumentative discussion where agents recommend arguments to people to justify their opinion is addressed in (Rosenfeld & Kraus (2016a)). The authors train three different ML models, an SVM, a Decision Tree, and a Multi-Layered Neural Network, in three different scenarios to predict the most appropriate candidate argument that may justify the opinion of a person. The first scenario is a predefined conversation on topics, such as “Why should I buy the car x?”. Each classifier is trained with data collected from Amazon Mechanical Turk. The second scenario uses data from Penn Treebank Corpus, which contains real argument conversations, not annotated by workers as in the first case. The third scenario uses data from medical corpus and the classifiers are executed to give the pros and cons of each topic.

An interesting implementation of AFVs to enhance the Explainability on the decisions of data driven models is presented in (Garcez et al. (2005)). The authors establish a relationship between neural networks and argumentation networks, combining reasoning and learning in the same argumentation framework. The authors present a neural argumentation algorithm for translating argumentation networks into standard neural networks. The algorithm can translate acyclic and circular AFs into neural networks, and it enables the learning of arguments, as well as the parallel computation of arguments.

Cocarascu et al. (2018) present an architecture that combines artificial neural networks for feature selection and AAF, for effective predictions, explainable both logically and dialectically. More specifically, the authors train an auto-encoder to rank features from examples. The auto-encoder is trained on the Mushroom Dataset (Dheeru & Taniskidou (2017)), which contains types of mushrooms paired with a group of features for each one of them. Using the most important features, an AAF is produced to classify and explain if a mushroom is poisonous or not. The arguments in the AAF are composed of the set of important features and the class for the mushroom. The method produces a set of logical rules to explain the classification. The method, outperforms a Decision Tree model and the combination of an auto-encoder with an artificial neural network, when trained on the same dataset. An interesting implementation of Hunter’s argumentation model (Besnard & Hunter (2001, 2009)), can be found in (Mollas et al. (2020)). The authors use a feature importance technique, in order to extract untruthful parts in the explanation of a data driven model. Moreover, the authors use this methodology to find the less untruthful explanation among many explanations of various data driven models.

8 Discussion and Conclusion

In this survey, we elaborated over the topic of Argumentation combined with XAI. Our goal was to report the most important methods that appeared in the literature to achieve Explainability in AI Systems, as well as their implementations. For this reason, we presented how Argumentation enables Explainability when tackling problems, such as decision making, justification of an opinion, and argumentation through dialogues. Moreover, we give an extensive literature overview on how Argumentation can be implemented into an agent, in order to solve the aforementioned problems, and what principles they must follow, in order not to be considered biased. Also, we showed how Argumentation can be used to construct explainable systems in the application domains of Medical Informatics, Law, the SW, Security, Robotics, and some general purpose systems. In our last chapter, we showed ML models that use Argumentation Theory to unlock the black box of Explainability in ML. The main contribution of this survey is the extensive literature overview of theoretical studies that use Argumentation to enhance Explainability, the literature overview of Argumentation implementations to build explainable systems. In Section 7 it becomes clear that Argumentation can work as a link between ML and XAI.

This survey revealed that not many studies exist which address the topic of commonsense knowledge fused into Argumentation Theory. Fusing commonsense knowledge into Argumentation with strict definitions will empower the Explainability capability of an AF, because it will allow it to reason with methods closer to human thinking and therefore being more persuasive in a dialogue. The studies that
attempt to fuse commonsense knowledge into Argumentation Theory, consider commonsense knowledge as: commonly accepted knowledge on which they use preferences (ˇCyras (2016)), text analytics (Moens (2016), Zhang et al. (2017)), Event Calculus (Almpani & Stefaneas (2017)), and even knowledge from external Web Knowledge Graphs (Kobbe et al. (2019)). One study (Vassiliades et al. (2020)) gives a notion of how an argument with commonsense knowledge could be defined but it is at a preliminary level.

Furthermore, using Argumentation Theory to explain why an event started, or what led to a decision, is a reasoning capability that an AF can offer and it can enhance its Explainability power. Causality could be achieved by reasoning over each step that led to a decision, and explain why alternatives were left out. Nevertheless, we see that not many works exist that combine Argumentation and causality for this purpose, apart from (Collins et al. (2019)) where the authors use argumentation to explain planning.

For future work, we plan to focus on arguments with commonsense knowledge, an interesting area that has not yet received much attention. More specifically, we will extent our survey over the topic of multi-argumentation frameworks that include arguments with commonsense knowledge and the various types of attack relations between them, which can be used to model, among other things, exceptions to commonsense knowledge (Vassiliades et al. (2020)). Arguments that can use commonsense knowledge can enhance the Explainability capabilities of Argumentation. Moreover, a literature overview of arguments with commonsense knowledge may provide models to represent commonsense knowledge that can be used in other research areas, such as agent theory, robotics, and even data-driven models in natural language processing to help mine arguments from human text dialogues, which could subsequently be used in a human-machine dialogue.

Acknowledgment

This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No 188.

References


Bikakis, A. & Antoniou, G., 2010. ‘Defeasible contextual reasoning with arguments in ambient intelligence’, IEEE Transactions on Knowledge and Data Engineering 22(11), 1492–1506.


Dechter, R. et al., 2003. Constraint processing, Morgan Kaufmann.


Fitting, M., 1992. ‘The stable model semantics for logic programming’.


Rosenfeld, A. & Kraus, S., 2016a. ‘Providing arguments in discussions on the basis of the prediction of human argumentative behavior’, *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6(4), 1–33.


Sheh, R. K.-M., 2017. ” why did you do that?” explainable intelligent robots, in ‘Workshops at the Thirty-First AAAI Conference on Artificial Intelligence’.


