The Knowledge Engineering Review, Vol. 36, e5, 1 of 35.
doi:10.1017/S0269888921000011

Argumentation and explainable artificial intelligence: a survey

ALEXANDROS VASSILIADES[1,2], NICK BASSILIADES[1], and THEODORE PATKOS[2]

[1] *School of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; e-mail: valexande@csd.auth.gr, nbassili@csd.auth.gr*

[2] *Institute of Computer Science, Foundation for Research and Technology - Hellas, 70013, Heraklion, Greece; 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, toward 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 20 yr, 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 is presented in Figure 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 that is 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).

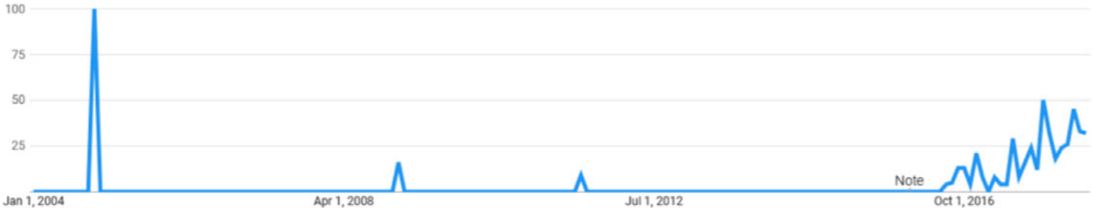

Figure 1. Google searches with XAI (Adadi & Berrada 2018)

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 (Core *et al*. 2005; Core *et al*. 2006; Madhikermi *et al*. 2019; Keneni *et al*. 2019), 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; Yang & Shafto 2017; Gunning & Aha 2019; Samek & Müller 2019; Fernandez *et al*. 2019); in some cases, visualization are also used (Choo & Liu 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 concludes 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 and 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:

1. We present an extensive literature review of how Argumentation enables XAI.
2. We show how Argumentation enables XAI, for solving problems in Decision-Making, Justification, and Dialogues.
3. We present how Argumentation has helped build explainable systems in the application domains of Medical Informatics, Law, the SW, Security, and Robotics.
4. 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.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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, is 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.

First, 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 who tries to locate which research studies are related to the solution of specific problem types.

Second, 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 and 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 $\mathcal{A}$ supports argument $\mathcal{B}$, then

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

the acceptance of $\mathcal{A}$ implies the acceptance of $\mathcal{B}$ and, as a consequence, the non-acceptance of $\mathcal{B}$ implies the non-acceptance of $\mathcal{A}$. Necessity support triggers the following constraint: if argument $\mathcal{A}$ supports argument $\mathcal{B}$, it means that $\mathcal{A}$ is necessary for $\mathcal{B}$. Hence, the acceptance of $\mathcal{B}$ implies the acceptance of $\mathcal{A}$ (conversely the non-acceptance of $\mathcal{A}$ implies the non-acceptance of $\mathcal{B}$). 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 an 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. First, 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 & Cohen 2003), or an Answer Set Programming (ASP) problem (Fitting 1992; Lifschitz 2019). Reduction-based techniques have the following 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 are fused in Longo (2016), Cocarascu and 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 and 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. Kakas and Michael (2020), 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 and 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 first 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 Debater[1], Debatepedia[2], Idebate[3], VBATES[4], and ProCon[5]. Moreover, Moens (2018) talks about studies where facial expressions are also inferred through a vision module to better understand the form of the argument. Another survey on AM is Lawrence and Reed (2020).

[1] https://www.research.ibm.com/artificial-intelligence/project-debater.

[2] http://www.debatepedia.org.

[3] https://idebate.org.

[4] http://vbate.idebate.org.

[5] https://www.procon.org.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Finally, the explanation of Case-Based Reasoning (CBR) systems is explored in the survey of Sørmo *et al.* (2005). Even though the authors do not include Argumentation in their survey, CBR works similarly to Case-Based Argumentation. Therefore, one could find interesting information about Explainability with CBR. Moreover, the authors extend the study of Wick and 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:

1. Transparency: Explain How the System Reached the Answer.
2. Justification: Explain Why the Answer is a Good Answer.
3. Relevance: Explain Why a Question Asked is Relevant.
4. Conceptualization: Clarify the Meaning of Concepts.
5. 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 and Berrada (2018), where the authors show how XAI methods have developed during the last 20 yr. 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 Možina *et al.* (2007), Došilović *et al.* (2018), Schoenborn and 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 and 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 (Tjoa & Guan 2019; Pocevičiūtė *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 2 yr ago by the European Union which has defined new regulations about this specific topic (Regulation 2016). Another, human-centric XAI survey is that of Páez (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.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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 every day problems. Yet, the demanding computational complexity limited the applicability of AFs for addressing real-world problems. Fortunately, during the last 20 yr, 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 *et al*. 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 Δ. 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 ϕ. The following properties hold: (i) $\Gamma \subset \Delta$, (ii) $\Gamma \vdash \phi$, (iii) $\Gamma \nvdash \bot$, and (iv) Γ is minimal with respect to set inclusion for which (i), (ii), and (iii) hold. Furthermore, there exist two types of attack, *undercut* and *rebut*.

1. For any propositions ϕ and ψ, ϕ attacks ψ when $\phi \equiv \neg\psi$ (the symbol $\neg$ denotes the strong negation).
2. Rebut: $\langle \Gamma_1, \phi_1 \rangle$ rebuts $\langle \Gamma_2, \phi_2 \rangle$, if ϕ_1 attacks ϕ_2.
3. Undercut: $\langle \Gamma_1, \phi_1 \rangle$ undercuts $\langle \Gamma_2, \phi_2 \rangle$, if ϕ_1 attacks some $\psi \in \Gamma_2$.
4. $\langle \Gamma_1, \phi_1 \rangle$ attacks $\langle \Gamma_2, \phi_2 \rangle$, if it rebuts or undercuts it.

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

1. *S conflict free, if* $\forall a, b \in \mathcal{S}$ *holds* $(a, b) \notin \mathcal{R}$ *(or* $(b, a) \notin \mathcal{R}$*)*.
2. *S defends an argument* $a \in \mathcal{A}$ *if* $\forall b \in \mathcal{A}$ *such that* $(b, a) \in \mathcal{R}$, $\exists c \in S$ *and* $(c, b) \in \mathcal{R}$.
3. *S is admissible if is conflict free, and* $\forall a \in S$, $\forall b \in A$, *such that* $(b, a) \in \mathcal{R}$, $\exists c \in S$ *holds* $(c, b) \in \mathcal{R}$.

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 $\mathcal{F}$, where $\mathcal{F} : 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$, such that for $S \subseteq \mathcal{A}$, $\mathcal{F}(S) = \{a \mid a \textit{ is defended by } S\}$. The fixpoint of a function $\mathcal{F}$ given a set S is a point where the input of the function is identical to the output, $\mathcal{F}(S) = S$.

1. Stable: *Let* $S \subseteq \mathcal{A}$, *S is a stable extension of* $(\mathcal{A}, \mathcal{R})$, *iff S is conflict free and* $\forall a \in A \setminus S$, $\exists b \in S$ such that $(b, a) \in \mathcal{R}$.
2. Preferred: *Let* $S \subseteq \mathcal{A}$, *S is a preferred extension of* $(\mathcal{A}, \mathcal{R})$, *iff S is maximal for the set inclusion among the admissible sets of* $\mathcal{A}$.
3. Complete: *Let* $S \subseteq \mathcal{A}$, *S is a complete extension of* $(\mathcal{A}, \mathcal{R})$, *iff S is conflict-free fixpoint of* $\mathcal{F}$.
4. Grounded: *Let* $S \subseteq \mathcal{A}$, *S is a grounded extension of* $(\mathcal{A}, \mathcal{R})$, *iff S is the minimal fixpoint of* $\mathcal{F}$.

Structured Argumentation Framework: SAF represents the arguments in the form of *logical rules* (Rule (1)), and it introduces the constraints of *preference between arguments*. First, we need to define some new concepts. A *theory* is a pair $(\mathcal{T}, \mathcal{P})$ 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, $\mathcal{T}$ is a set of ground literals, and $\mathcal{P}$ is a set of rules which follow the general form *label* : *claim* ← *premise* (Rule (1)).

$$r : a \leftarrow b_1, \ldots, b_n \tag{1}$$

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

Rule (1) 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 (1), we define the preference of *r* over *r*' by *prefer*(*r*,*r*'). The arguments in SAF have the same format similar to Rule (1). Therefore, when we say we prefer an argument *a* over *b*, we mean the relation *prefer*(*a*,*b*). If the preference rules are removed from the framework, then it is also called *Assumption-Based Argumentation Framework* (ABA) (Dung *et al*. 2009), 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:

1. an argument is labeled in if all of its attackers are labeled out and is called acceptable.
2. an argument is labeled out if at least one of its attackers is labeled in.
3. an argument is labeled undec, when we cannot label it neither in nor out.

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

Bipolar Argumentation Framework: A BAF is a triplet $\left(\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-\right)$, 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 $\left(\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-, \tau, \sigma\right)$, 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 τ 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 $b \nless a$. 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 $\left(\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-\right)$ is a BAF and $\mathcal{P}$ is a probability distribution function over the set $PD = \{(\mathcal{A}', \mathcal{R}'^+, \mathcal{R}'^-) \mid \mathcal{A}' \subseteq \mathcal{A} \wedge \mathcal{R}'^- \subseteq (\mathcal{A}' \times \mathcal{A}') \cap \mathcal{R}^- \wedge \mathcal{R}'^+ \subseteq (\mathcal{A}' \times \mathcal{A}') \cap \mathcal{R}^+\}$. The elements in $PD\left((\mathcal{A}, \mathcal{R}^+, \mathcal{R}^-, \mathcal{P})\right)$ 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 (Core *et al*. 2006). 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 (Regulation 2016). Therefore, in this section, we describe how Argumentation enables Explainability. We can see

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

that explaining a decision with argumentation is not something that emerged in recent years but existed in Argumentation Theory from its beginning (Pavese 2019).

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 (SAS) and multi-agent systems (MAS) 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 SAS and 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 and 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 is 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 $a =$ *'I will go for football after work'*, and $b =$ *'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 $prefer(b,a)$. We can see implementations of such theoretical frameworks in Panisson *et al.* (2014), Panisson and 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.* (2016b) 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 $\mathcal{X}$, a set of classes $\mathcal{C}$, and a set of hypothesis $\mathcal{H}$ which are governed by a pre-ordered relation $\leq$ that defines which hypothesis is stronger. Then, an example $x \in \mathcal{X}$ is classified in a class $c \in \mathcal{C}$ by the hypothesis $h \in \mathcal{H}$, 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 AFs 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 ease the exchange of arguments (Willmott *et al.* 2006; Panisson *et al.* 2015), allow the agents to perform negotiations (Pilotti *et al.* 2015), be more persuasive (Black & Atkinson 2011; Rosenfeld & Kraus 2016b), and to explain in more detail how they reached a decision (Laird & Nielsen 1994).

The recent study of Ciatto *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 $M : \mathcal{X} \rightarrow \mathcal{Y}$ maps an input set of objects $\mathcal{X}$ to an output set of actions $\mathcal{Y}$, 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.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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 and Atkinson (2009), Wyner *et al*. (2012b), 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*. (2012a), 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) use 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.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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 Šešelja and 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 $\left(\mathcal{A}, \mathcal{X}, \mathcal{R}, \mathcal{R}', \tilde{\mathcal{R}}\right)$ such that $\mathcal{X}$ is the set of topics that receive arguments from $\mathcal{A}$ as explanations, the relation $\mathcal{R}'$ defines that an argument $a \in \mathcal{A}$ is part of the explanation of an element from $x \in \mathcal{X}$, and finally $\tilde{\mathcal{R}}$ 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 $\mathcal{E}$ that contains all the arguments from $\mathcal{A}$ that are connected through $\mathcal{R}'$ with an element from $x \in \mathcal{X}$ and do not co-exist in $\tilde{\mathcal{R}}$. On the same, principles are the formalization of 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 Čyras *et al.* (2016a); Čyras *et al.* (2016b), 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 = book this hotel, O_2 = 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 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 Čyras *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 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 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 (Mercier & Sperber 2011). The contributions of

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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 Amgoud and 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. First, the authors consider beliefs and opinions as arguments. Second, pairs of options are compared using decision principles. The decision principles are: (i) *Unipolar* refers only to the arguments attacks or defenses; (ii) *Bipolar* takes into consideration both; and (iii) *Non-Polar* is 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 and 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 first 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 arguments 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 *et al*. (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 $(\mathcal{M}, \mathcal{AT}, \mathcal{T}, \mathcal{L}, \mathcal{U}, \mathcal{R})$ such that:

1. $\mathcal{M}$ is a finite set of *items*.
2. $\mathcal{AT}$ is a finite, non-empty set of *attributes* for the items.
3. $\mathcal{T}$ a set of *types* for the attributes.
4. the sets $\mathcal{M}$ and $\mathcal{AT}$ are pairwise disjoint.
5. $\mathcal{X} = \mathcal{M} \cup \mathcal{AT}$.
6. $\mathcal{L} \subseteq (\mathcal{M} \times \mathcal{AT})$ is a symmetrical binary relation.
7. $\mathcal{U}$ is a finite, non-empty set of *users*.
8. $\mathcal{R} : \mathcal{U} \times \mathcal{X} \rightarrow [-1, 1]$ 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 $\mathcal{M}$, and positive and negative ratings to the attributes related to an item from $\mathcal{M}$ 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,

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

defensive arguments from the AF, and external knowledge. One important study in this field is Čyras *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 and 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 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. 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 ϕ 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).

$$tr(A) = tr(\phi_1) \otimes \ldots \otimes tr(\phi_n) \tag{2}$$

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 Tang *et al.* (2012), where the authors also define trust of arguments in MAS. Cerutti *et al.* (2019) designed the *ArgSemSAT* system that can return the complete labelings of an AAF and is commonly used for the justification of the acceptability of an argument. ArgSemSAT is based on satisfiability solvers (SAT), and its biggest innovations are: (i) it can find a global labeling 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 labeling 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 Thimm and 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 *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 *BBC*, and the argument *A2* from *Fox News*.

$$A1 = \textit{The weather tomorrow will be shinny}$$

$$A2 = \textit{The weather tomorrow will be rainy}$$

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

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Bench-Capon (2003a); Bench-Capon (2002), where the author extends an AAF by adding values (AFV). Subsequently, AFVs were used to solve legal conflicts (Bench-Capon 2003b; Bench-Capon *et al*. 2005), to infer inconsistency between preferences of arguments (Amgoud & Cayrol 2002a), and to produce acceptable arguments (Amgoud & Cayrol 2002b).

The notions of AFVs were extended by Modgil in Hierarchical Argumentation Frameworks (Modgil 2006a, b, 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 $((\mathcal{A}_1, \mathcal{R}_1), \ldots, (\mathcal{A}_n, \mathcal{R}_n))$. $\mathcal{A}_i$ contains arguments whose value is a_i and $\mathcal{R}_i$ attacks which are related to the arguments of $\mathcal{A}_i$, 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, b) 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), Dunne *et al*. (2011).

AFVs had a significant impact contributed significantly to practical reasoning, that is, reasoning about what action is better for an agent to perform in a particular scenario. The authors in Atkinson and Bench-Capon (2007b) 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 are 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 and Bench-Capon (2007a) and (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 (Hage *et al*. 1993; Bench-Capon 1998; Bench-Capon *et al*. 2000).

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 2005b). More specifically, the author constructed an AF with a clear reply structure, where each dialogue moves 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 and Norman (1995).

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

McBurney and Parson in McBurney and Parsons (2002) study 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 and 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. First, 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. Second, 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; Girle *et al*. 2003; García *et al*. 2007; Luis-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 are 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 and 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 were supported by Slugoski *et al*. (1993).

In Bex *et al*. (2012), Bex and 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

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Format (AIF) (Chesnevar *et al*. 2006). A similar approach is found in Letia and 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. In 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 $not(A)$, 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 $AF = (\mathcal{A}, \mathcal{R})$, a non-acceptable argument $a \in \mathcal{A}$ can become acceptable if a specific set of arguments $A \subseteq \mathcal{A}$ and a set of attacks $R \subseteq \mathcal{R}$ 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 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; Noël & 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). Gorgias was also used in SAS (Spanoudakis & Moriaitis, 2009) for choosing the best policy on pricing products based on a set of restrictions, and affecting relations between products

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

from a KB. MASs (Bassiliades *et al.* 2018) have also used Gorgias for choosing the best policy in energy saving among conflict policies of agents and are translated to argument preferences.

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 (*topic*, *claim*, *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* $\mathcal{SF}$, 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 $\mathcal{SF}(a) > \mathcal{SF}(b)$, 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[6].

Another important aspect when we provide an explanation is the level of trust we have for an argument. *ArgTrust* (Sklar *et al.* 2016) 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 *et al.* (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 (2017), Prakken *et al.* (2015). In the first two studies, the author provides a formalization of a legal case into Argumentation Theory. The legal case is first 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 his 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

[6] https://vue.tufts.edu.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Al-Abdulkarim *et al.* (2016a, b, 2019), where the argumentation dialectical procedure is projected into a Dispute Tree. On the other hand, Sklar and 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 *meta level argumentation semantics*. The authors provide a set of logic rules with preferences that can be implemented in a LBAF 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 CBR, and Gordon and 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 and 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 Možina *et al.* (2007), Chapman *et al.* (2019), Kökciyan *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 Čyras *et al.* (2018), Čyras and 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 and 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

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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 have. 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 data set (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 *et al*. (2018) propose an explainable model based on Argumentation for detection of dementia. The framework uses a Convolutional Neural Network to extract features from images, such as the size, the region, and the drawing test performance. The framework can then to 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 and Azhar (2015, 2018), Azhar and Sklar (2016, 2017). 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 and Sklar (2016, 2017), Sklar and 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*.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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 *et al*. (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 the 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.

[7] http://www.vocapia.com/.

[8] https://wordnet.princeton.edu.

[9] https://wordnet.princeton.edu.

[10] https://www.wikidata.org/wiki/Wikidata:Main_Page.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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.

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*[11] (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[12]. 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; Besnard & Hunter 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[13] 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 & Šnajder. 2014b, 2015). The mechanism relies on an SVM trained on the custom made ComArg corpus (Boltužić & Šnajder 2014a), a data set of arguments supported by comments, using the data of ProCon[14] and Idebate[15]. 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 data set created from CreateDebate[16], 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 *et al*. (2018), where the user can build an argumentation graph and see the justification of an argument.

[11] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki.

[12] https://competitions.codalab.org/competitions/17327.

[13] http://framenet.icsi.berkeley.edu/fndrupal.

[14] https://www.procon.org.

[15] https://idebate.org.

[16] https://www.createdebate.com.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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 attack. 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 Kökciyan *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. Kökciyan and Yolum (2017) study is an extension in the context of the Internet of Things. The study (Bassiliades *et al.* 2018), based on the studies of Spanoudakis *et al.* (2016a, 2007), focuses also on the same principles for accessing patient's data.

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, 2018a, b, 2020), 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.* (2016b, c) 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 *et al.* (2016a), where the authors construct a data set with real-life cyber-attack scenarios from hackers, collected from the DEFCON[17] 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 and 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

[17] https://www.defcon.org/html/links/dc-ctf.html.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

the analysis and generation of anomaly free firewall configuration rules. The modularity of this policy generator framework is that it allows for customization 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 though Dialogue, or Argument Classification.

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 and 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[18], 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 data sets such as the Hotel Dataset (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.

CBR or Instance-based Learning is one of the most commonly used methods of ML, which was used from the the early days of the research in ML. Therefore, there are many studies which use CBR and Argumentation to achieve Explainability. In Čyras *et al.* (2016a); Čyras *et al.* (2016b), 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 Čyras *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.

[18] http://corpora.aifdb.org/.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Table 1. Overview of argumentation systems for XAI

Domain	Decision making	Justification	Explanation through dialogue	Argument classification
Law	Bench-Capon (2020)	Prakken (2017)	Sklar *et al.* (2013), Al-Abdulkarim *et al.* (2016a); Al-Abdulkarim *et al.* (2016b); Al-Abdulkarim *et al.* (2019), McCarty (1976)	Prakken *et al.* (2015, 2005a), Rissland and Ashley (1987), Aleven (1997), Al-Abdulkarim *et al.* (2016c), Gordon and Walton (2009)
Medical informatics	Longo and Hederman (2013), Spanoudakis *et al.* (2017), Chapman *et al.* (2019) Čyras and Oliveira (2019), Oliveira *et al.* (2018)	Grando *et al.* (2013), Donadello *et al.* (2019), Zeng *et al.* (2018)	Qurat-ul-ain Shaheen and Bowles (2020)	Prentzas *et al.* (2019), Mayer *et al.* (2018)
Robotics	Azhar and Sklar (2016); Azhar and Sklar (2017), Sklar and Azhar (2015)	Torres *et al.* (2019), Fischer *et al.* (2018)	Modgil *et al.* (2018), Wanner *et al.* (2017), Meditskos *et al.* (2019)	
SW	Botschen *et al.* (2018)	Baroni *et al.* (2018)	Boltužić and Šnajder (2014b), Bex *et al.* (2013), Boltužić and Šnajder (2015)	Swanson *et al.* (2015), Indrie and Groza (2010), Rahwan *et al.* (2011)
Security	Kökciyan *et al.* (2017), Kökciyan and Yolum (2017), Fogues *et al.* (2017a); Fogues *et al.* (2017b), Shakarian *et al.* (2015)	Karafili *et al.* (2017, 2018b, 2020, 2018a, n.d.), Nunes *et al.* (2016c); Nunes *et al.* (2016b); Nunes *et al.* (2016a), Genitsaridi *et al.* (2013), Bikakis and Antoniou (2010), Bandara *et al.* (2006); Bandara *et al.* (2009), Santini and Yautsiukhin (2015)	Rowe *et al.* (2012), Murukannaiah *et al.* (2015)	Panisson *et al.* (2018)
General purpose	Dauphin and Cramer (2017), Modgil and Prakken (2014), Green *et al.* (2019), Kakas *et al.* (2019), Bassiliades *et al.* (2018), Noël and Kakas (2009), Spanoudakis *et al.* (2016b)	Tang *et al.* (2012), Baroni *et al.* (2015), Spanoudakis and Moriaitis (2009)	Chesnevar *et al.* (2006), Betz *et al.* (2019)	Bex *et al.* (2010)

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 Bing[19]. 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 data set UKPConvArg (Habernal & Gurevych 2016). Furthermore, the authors give four supervised models that achieve state-of-the-art results on the same data set: (i) An Bi-LSTM that receives as input the vector of the concatenation argument-topic pairs, using Glove embeddings[20], (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 data set of annotated arguments in Wikipedia articles[21].

Another study that uses external knowledge from ConceptNet[22] and DBpedia, NLP methods, and KB features to predict the type of relations between arguments is Kobbe *et al*. (2019), similar to Cocarascu and 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 and 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[23], 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

[19] www.bing.com.

[20] https://nlp.stanford.edu/projects/glove/.

[21] https://github.com/UKPLab/.

[22] https://conceptnet.io.

[23] https://catalog.ldc.upenn.edu/LDC99T42.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

the Mushroom Data set (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 data set. 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 (Čyras 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 NLP to help mine arguments from human text dialogues, which could subsequently be used in a human–machine dialogue.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

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

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

Al-Abdulkarim, L., Atkinson, K. & Bench-Capon, T. 2016a. A methodology for designing systems to reason with legal cases using abstract dialectical frameworks. *Artificial Intelligence and Law* **24**(1), 1–49.

Al-Abdulkarim, L., Atkinson, K. & Bench-Capon, T. 2016b. *Statement Types in Legal Argument*. IOS Press.

Al-Abdulkarim, L., Atkinson, K. & Bench-Capon, T. J. 2016c. Angelic secrets: bridging from factors to facts in us trade secrets. In *JURIX*, 113–118.

Al-Abdulkarim, L., Atkinson, K., Bench-Capon, T., Whittle, S., Williams, R. & Wolfenden, C. 2019. Noise induced hearing loss: building an application using the angelic methodology. *Argument & Computation* **10**(1), 5–22.

Aleven, V. A. 1997. *Teaching Case-Based Argumentation Through a Model and Examples*. Citeseer.

Almpani, S. & Stefaneas, P. S. 2017. On proving and argumentation. In *AIC*, 72–84.

Amgoud, L. & Cayrol, C. 2002a. Inferring from inconsistency in preference-based argumentation frameworks. *Journal of Automated Reasoning* **29**(2), 125–169.

Amgoud, L. & Cayrol, C. 2002b. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence* **34**(1–3), 197–215.

Amgoud, L., Cayrol, C., Lagasquie-Schiex, M.-C. & Livet, P. 2008. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems* **23**(10), 1062–1093.

Amgoud, L. & Prade, H. 2006. Explaining qualitative decision under uncertainty by argumentation. In *Proceedings of the National Conference on Artificial Intelligence*, **21**, 219. AAAI Press, MIT Press, 1999.

Amgoud, L. & Prade, H. 2009. Using arguments for making and explaining decisions. *Artificial Intelligence* **173**(3–4), 413–436.

Amgoud, L. & Serrurier, M. 2007. Arguing and explaining classifications. In *International Workshop on Argumentation in Multi-Agent Systems*, 164–177, Springer.

Amgoud, L. & Serrurier, M. 2008. Agents that argue and explain classifications. *Autonomous Agents and Multi-Agent Systems* **16**(2), 187–209.

Anjomshoae, S., Najjar, A., Calvaresi, D. & Främling, K. 2019. Explainable agents and robots: results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

Antaki, C. & Leudar, I. 1992. Explaining in conversation: towards an argument model. *European Journal of Social Psychology* **22**(2), 181–194.

Arioua, A., Tamani, N., Croitoru, M. & Buche, P. 2014. Query failure explanation in inconsistent knowledge bases: a dialogical approach. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 119–133. Springer.

Atkinson, K. & Bench-Capon, T. 2007a. Action-based alternating transition systems for arguments about action. In *AAAI*, **7**, 24–29.

Atkinson, K. & Bench-Capon, T. 2007b. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence* **171**(10–15), 855–874.

Atkinson, K. & Bench-Capon, T. 2018. Taking account of the actions of others in value-based reasoning. *Artificial Intelligence* **254**, 1–20.

Atkinson, K., Bench-Capon, T. & Bollegala, D. 2020. Explanation in ai and law: past, present and future. *Artificial Intelligence*, 103387.

Atkinson, K., Bench-Capon, T. J. & McBurney, P. 2005a. Multi-agent argumentation for edemocracy. In *EUMAS*, 35–46.

Atkinson, K., Bench-Capon, T. & Mcburney, P. 2005b. A dialogue game protocol for multi-agent argument over proposals for action. *Autonomous Agents and Multi-Agent Systems* **11**(2), 153–171.

Atkinson, K. M., Bench-Capon, T. J., Cartwright, D. & Wyner, A. Z. 2011. Semantic models for policy deliberation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, 81–90.

Azhar, M. Q. & Sklar, E. I. 2016. Analysis of empirical results on argumentation-based dialogue to support shared decision making in a human-robot team. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 861–866. IEEE.

Azhar, M. Q. & Sklar, E. I. 2017. A study measuring the impact of shared decision making in a human-robot team. *The International Journal of Robotics Research* **36**(5–7), 461–482.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Bandara, A. K., Kakas, A. C., Lupu, E. C. & Russo, A. 2009. Using argumentation logic for firewall configuration management. In *2009 IFIP/IEEE International Symposium on Integrated Network Management*, 180–187. IEEE.

Bandara, A. K., Kakas, A., Lupu, E. C. & Russo, A. 2006. Using argumentation logic for firewall policy specification and analysis. In *International Workshop on Distributed Systems: Operations and Management*, 185–196. Springer.

Baroni, P., Borsato, S., Rago, A. & Toni, F. 2018. The "games of argumentation" web platform. *In COMMA*, 447–448.

Baroni, P., Caminada, M. & Giacomin, M. 2011. An introduction to argumentation semantics. *Knowledge Engineering Review* **26**(4), 365.

Baroni, P., Rago, A. & Toni, F. 2018. How many properties do we need for gradual argumentation? In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Baroni, P., Romano, M., Toni, F., Aurisicchio, M. & Bertanza, G. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* **6**(1), 24–49.

Bassiliades, N., Spanoudakis, N. I. & Kakas, A. C. 2018. Towards multipolicy argumentation. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, 1–10.

Bench-Capon, T. 2002. Value based argumentation frameworks. arXiv preprint cs/0207059.

Bench-Capon, T., Atkinson, K. & Chorley, A. 2005. Persuasion and value in legal argument. *Journal of Logic and Computation* **15**(6), 1075–1097.

Bench-Capon, T. J. 1998. Specification and implementation of toulmin dialogue game. In *Proceedings of JURIX*, **98**, 5–20.

Bench-Capon, T. J. 2003a. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* **13**(3), 429–448.

Bench-Capon, T. J. 2003b. Try to see it my way: modelling persuasion in legal discourse. *Artificial Intelligence and Law* **11**(4), 271–287.

Bench-Capon, T. J. 2020. Before and after dung: argumentation in ai and law. *Argument & Computation* (Preprint), 1–18.

Bench-Capon, T. J. M., Geldard, T. & Leng, P. H. 2000. A method for the computational modelling of dialectical argument with dialogue games. *Artificial Intelligence and Law* **8**(2–3), 233–254.

Besnard, P. & Hunter, A. 2001. A logic-based theory of deductive arguments. *Artificial Intelligence* **128**(1–2), 203–235.

Besnard, P. & Hunter, A. 2008. *Elements of Argumentation*, 47. MIT Press.

Besnard, P. & Hunter, A. 2009. Argumentation based on classical logic. In *Argumentation in Artificial Intelligence*, 133–152. Springer.

Betz, G., Hamann, M., Mchedlidze, T. & von Schmettow, S. 2019. Applying argumentation to structure and visualize multi-dimensional opinion spaces. *Argument & Computation* **10**(1), 23–40.

Bex, F., Bench-Capon, T. J. & Verheij, B. 2011. What makes a story plausible? the need for precedents. In *JURIX*, 23–32.

Bex, F., Budzynska, K. & Walton, D. 2012. Argumentation and explanation in the context of dialogue. *Explanation-aware Computing ExaCt 2012* **9**, 6.

Bex, F., Lawrence, J., Snaith, M. & Reed, C. 2013. Implementing the argument web. *Communications of the ACM* **56**(10), 66–73.

Bex, F., Prakken, H. & Reed, C. 2010. A formal analysis of the AIF in terms of the ASPIC framework. *In COMMA*, 99–110.

Bex, F. & Walton, D. 2016. Combining explanation and argumentation in dialogue. *Argument & Computation* **7**(1), 55–68.

Biere, A., Heule, M. & van Maaren, H. 2009. *Handbook of Satisfiability*, 185. IOS Press.

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

Bistarelli, S., Pirolandi, D. & Santini, F. 2009. Solving weighted argumentation frameworks with soft constraints. In *International Workshop on Constraint Solving and Constraint Logic Programming*, 1–18. Springer.

Black, E. & Atkinson, K. 2011. Choosing persuasive arguments for action. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume* **3**, 905–912. International Foundation for Autonomous Agents and Multiagent Systems.

Boltužić, F. & Šnajder, J. 2014a. Back up your stance: recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, 49–58. Association for Computational Linguistics. http://www.aclweb.org/anthology/W14-2107.

Boltužić, F. & Šnajder, J. 2014b. Back up your stance: recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, 49–58.

Boltužić, F. & Šnajder, J. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, 110–115.

Bonacina, M. P. 2017. Automated reasoning for explainable artificial intelligence. In *ARCADE@ CADE*, 24–28.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Bonzon, E., Delobelle, J., Konieczny, S. & Maudet, N. 2016. A comparative study of ranking-based semantics for abstract argumentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 30.

Botschen, T., Sorokin, D. & Gurevych, I. 2018. Frame-and entity-based knowledge for common-sense argumentative reasoning. In *Proceedings of the 5th Workshop on Argument Mining*, 90–96.

Burgemeestre, B., Hulstijn, J. & Tan, Y.-H. 2011. Value-based argumentation for justifying compliance. *Artificial Intelligence and Law* **19**(2–3), 149.

Buvac, S. & Mason, I. A. 1993. Propositional logic of context. In *AAAI*, 412–419.

Caminada, M. 2008. A gentle introduction to argumentation semantics. Lecture Material, Summer.

Caminada, M., Sá, S., Alcântara, J. & Dvořák, W. 2015. On the equivalence between logic programming semantics and argumentation semantics. *International Journal of Approximate Reasoning* **58**, 87–111.

Cartwright, D. & Atkinson, K. 2009. Using computational argumentation to support e-participation. *IEEE Intelligent Systems* **24**(5), 42–52.

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

Cayrol, C. & Lagasquie-Schiex, M.-C. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 378–389. Springer.

Cerutti, F., Giacomin, M. & Vallati, M. 2019. How we designed winning algorithms for abstract argumentation and which insight we attained. *Artificial Intelligence* **276**, 1–40.

Chapman, M., Balatsoukas, P., Ashworth, M., Curcin, V., Kökciyan, N., Essers, K., Sassoon, I., Modgil, S., Parsons, S. & Sklar, E. I. 2019. Computational argumentation-based clinical decision support. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2345–2347. International Foundation for Autonomous Agents and Multiagent Systems.

Charwat, G., Dvořák, W., Gaggl, S. A., Wallner, J. P. & Woltran, S. 2015. Methods for solving reasoning problems in abstract argumentation–a survey. *Artificial Intelligence* **220**, 28–63.

Chesnevar, C., McGinnis, J., Modgil, S., Rahwan, I., Reed, C., Simari, G., South, M., Vreeswijk, G. & Willmott, S. 2006. Towards an argument interchange format. *The Knowledge Engineering Review* **21**(4), 293–316.

Choo, J. & Liu, S. 2018. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications* **38**(4), 84–92.

Ciatto, G., Calvaresi, D., Schumacher, M. I. & Omicini, A. 2015. *An Abstract Framework for Agent-Based Explanations in AI*. Springer.

Clark, C. E. 1919. Eisner v Macomber and some income tax problems. *Yale LJ* **29**, 735.

Cocarascu, O., Čyras, K. & Toni, F. 2018. Explanatory predictions with artificial neural networks and argumentation. In *Proceedings of the 2nd Workshop on Explainable Artificial Intelligence (XAI 2018).*

Cocarascu, O. & Toni, F. 2016. Argumentation for machine learning: a survey. *In COMMA*, 219–230.

Cocarascu, O. & Toni, F. 2018. Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics* **44**(4), 833–858.

Cogan, E., Parsons, S. & McBurney, P. 2005. What kind of argument are we going to have today?. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, 544–551.

Cohen, A., Gottifredi, S., García, A. J. & Simari, G. R. 2014. A survey of different approaches to support in argumentation systems. *The Knowledge Engineering Review* **29**(5), 513–550.

Collenette, J., Atkinson, K. & Bench-Capon, T. 2020. An explainable approach to deducing outcomes in european court of human rights cases using ADFs. *Frontiers in Artificial Intelligence and Applications* **326**, 21–32.

Collins, A., Magazzeni, D. & Parsons, S. 2019. Towards an argumentation-based approach to explainable planning. In *ICAPS 2019 Workshop XAIP Program Chairs*.

Core, M. G., Lane, H. C., Van Lent, M., Gomboc, D., Solomon, S. & Rosenberg, M. 2006. Building explainable artificial intelligence systems. In *AAAI*, 1766–1773.

Core, M. G., Lane, H. C., Van Lent, M., Solomon, S., Gomboc, D. & Carpenter, P. 2005. Toward question answering for simulations. In *Proceedings of the IJCAI 2005 Workshop on Knowledge and Reasoning for Answering Questions (KRAQ05)*. Citeseer.

Correia, V. 2012. The ethics of argumentation. *Informal Logic* **32**(2), 222–241.

Coste-Marquis, S., Konieczny, S., Marquis, P. & Ouali, M. A. 2012a. Selecting extensions in weighted argumentation frameworks. *In COMMA*, **12**, 342–349.

Coste-Marquis, S., Konieczny, S., Marquis, P. & Ouali, M. A. 2012b. Weighted attacks in argumentation frameworks. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Čyras, K. 2016. Argumentation-based reasoning with preferences. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, 199–210. Springer.

Čyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., Greenberg, D. & Hapuarachchi, T. 2019. Explanations by arbitrated argumentative dispute. *Expert Systems with Applications* **127**, 141–156.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Čyras, K., Delaney, B., Prociuk, D., Toni, F., Chapman, M., Dominguez, J. & Curcin, V. 2018. Argumentation for explainable reasoning with conflicting medical recommendations. In *Proceedings of the Joint Proceedings of Reasoning with Ambiguous and Conflicting Evidence and Recommendations in Medicine (MedRACER 2018)*.

Čyras, K., Fan, X., Schulz, C. & Toni, F. 2017. Assumption-based argumentation: disputes, explanations, preferences. *Journal of Applied Logics-ifcolog Journal of Logics and their Applications* **4**(8), 2407–2456.

Čyras, K. & Oliveira, T. 2019. Resolving conflicts in clinical guidelines using argumentation. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 1731–1739. International Foundation for Autonomous Agents and Multiagent Systems.

Čyras, K., Satoh, K. & Toni, F. 2016a. Abstract argumentation for case-based reasoning. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Čyras, K., Satoh, K. & Toni, F. 2016b. Explanation for case-based reasoning via abstract argumentation. In *International Conference on the Principles of Argumentation*.

Czubaroff, J. 2007. Justice and argument: toward development of a dialogical argumentation theory. *Argumentation and Advocacy* **44**(1), 18–35.

Das, A. & Rad, P. 2020. Opportunities and challenges in explainable artificial intelligence (xai): a survey. arXiv preprint arXiv:2006.11371.

Dauphin, J. & Cramer, M. 2017. Aspic-end: structured argumentation with explanations and natural deduction. In *International Workshop on Theorie and Applications of Formal Argumentation*, 51–66. Springer.

Dechter, R. & Cohen, D. 2003. *Constraint Processing*. Morgan Kaufmann.

Deeks, A. 2019. The judicial demand for explainable artificial intelligence. *Columbia Law Review* **119**(7), 1829–1850.

Dheeru, D. & Taniskidou, E. K. 2017. UCI machine learning repository: mushroom data set.

Donadello, I., Dragoni, M. & Eccher, C. 2019. Persuasive explanation of reasoning inferences on dietary data. In *Contributo in Atti di Convegno (Proceeding)*.

Došilović, F. K., Brčić, M. & Hlupić, N. 2018. Explainable artificial intelligence: a survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. IEEE.

Doutre, S. & Mailly, J.-G. 2018. Constraints and changes: a survey of abstract argumentation dynamics. *Argument & Computation* **9**(3), 223–248.

Dung, P. M. 1995. An argumentation-theoretic foundation for logic programming. *The Journal of Logic Programming* **22**(2), 151–177.

Dung, P. M. 2016. An axiomatic analysis of structured argumentation with priorities. *Artificial Intelligence* **231**, 107–150.

Dung, P. M., Kowalski, R. A. & Toni, F. 2009. Assumption-based argumentation. In *Argumentation in Artificial Intelligence*, 199–218. Springer.

Dung, P. M. & Son, T. C. 1995. Nonmonotonic inheritance, argumentation and logic programming. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, 316–329. Springer.

Dunne, P. E., Hunter, A., McBurney, P., Parsons, S. & Wooldridge, M. 2011. Weighted argument systems: basic definitions, algorithms, and complexity results. *Artificial Intelligence* **175**(2), 457–486.

Dunne, P. E., Hunter, A., McBurney, P., Parsons, S. & Wooldridge, M. J. 2009. Inconsistency tolerance in weighted argument systems. In *AAMAS (2)*, 851–858.

Fan, X. & Toni, F. 2014. On computing explanations in abstract argumentation. In *ECAI*, 1005–1006.

Fan, X. & Toni, F. 2015a. On computing explanations in argumentation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Fan, X. & Toni, F. 2015b. On explanations for non-acceptable arguments. In *International Workshop on Theory and Applications of Formal Argumentation*, 112–127. Springer.

Fan, X., Toni, F., Mocanu, A. & Williams, M. 2014. Dialogical two-agent decision making with assumption-based argumentation. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, 533–540.

Fazzinga, B., Flesca, S. & Furfaro, F. 2018. Probabilistic bipolar abstract argumentation frameworks: complexity results. *In IJCAI*, 1803–1809.

Fernandez, A., Herrera, F., Cordon, O., del Jesus, M. J. & Marcelloni, F. 2019. Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to?. *IEEE Computational Intelligence Magazine* **14**(1), 69–81.

Fischer, L., Hasler, S., Deigmöller, J., Schnürer, T., Redert, M., Pluntke, U., Nagel, K., Senzel, C., Ploennigs, J., Richter, A. & Eggert, J. 2018. Which tool to use? grounded reasoning in everyday environments with assistant robots. In *CogRob@ KR*, 3–10.

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

Fogues, R. L., Murukannaiah, P. K., Such, J. M. & Singh, M. P. 2017a. Sharing policies in multiuser privacy scenarios: incorporating context, preferences, and arguments in decision making. *ACM Transactions on Computer-Human Interaction (TOCHI)* **24**(1), 1–29.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Fogues, R. L., Murukannaiah, P. K., Such, J. M. & Singh, M. P. 2017b. Sosharp: recommending sharing policies in multiuser privacy scenarios. *IEEE Internet Computing* **21**(6), 28–36.

Friedrich, G. & Zanker, M. 2011. A taxonomy for generating explanations in recommender systems. *AI Magazine* **32**(3), 90–98.

Garcez, A. S., Gabbay, D. M. & Lamb, L. C. 2005. Value-based argumentation frameworks as neural-symbolic learning systems. *Journal of Logic and Computation* **15**(6), 1041–1058.

García, A., Chesñevar, C., Rotstein, N. & Simari, G. 2007. An abstract presentation of dialectical explanations in defeasible argumentation. In *ArgNMR07*, 17–32.

García, A. J., Chesñevar, C. I., Rotstein, N. D. & Simari, G. R. 2013. Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. *Expert Systems with Applications* **40**(8), 3233–3247.

Genitsaridi, I., Bikakis, A. & Antoniou, G. 2013. Deal: a distributed authorization language for ambient intelligence. In *Pervasive and Ubiquitous Technology Innovations for Ambient Intelligence Environments*, 188–204. IGI Global.

Girle, R., Hitchcock, D., McBurney, P. & Verheij, B. 2003. Decision support for practical reasoning. In *Argumentation Machines*, 55–83. Springer.

Gordon, T. F. 1993. The pleadings game. *Artificial Intelligence and Law* **2**(4), 239–292.

Gordon, T. F. & Walton, D. 2009. Legal reasoning with argumentation schemes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 137–146.

Grando, M. A., Moss, L., Sleeman, D. & Kinsella, J. 2013. Argumentation-logic for creating and explaining medical hypotheses. *Artificial Intelligence in Medicine* **58**(1), 1–13.

Green, N. L., Branon, M. & Roosje, L. 2019. Argument schemes and visualization software for critical thinking about international politics. *Argument & Computation* **10**(1), 41–53.

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

Gunning, D. & Aha, D. W. 2019. Darpa's explainable artificial intelligence program. *AI Magazine* **40**(2), 44–58.

Habernal, I. & Gurevych, I. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1214–1223.

Hage, J. C., Leenes, R. & Lodder, A. R. 1993. Hard cases: a procedural approach. *Artificial Intelligence and Law* **2**(2), 113–167.

Hilton, D. J. 1990. Conversational processes and causal explanation. *Psychological Bulletin* **107**(1), 65.

Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R. & Zatloukal, K. 2017. Towards the augmented pathologist: challenges of explainable-ai in digital pathology. arXiv preprint arXiv:1712.06657.

Indrie, S. M. & Groza, A. 2010. Enacting argumentative web in semantic wikipedia. In *9th RoEduNet* IEEE International Conference, 163–168. IEEE.

Josephson, J. R. & Josephson, S. G. 1996. *Abductive Inference: Computation, Philosophy, Technology*. Cambridge University Press.

Kakas, A. C., Moraitis, P. & Spanoudakis, N. I. 2019. Gorgias: applying argumentation. *Argument & Computation* **10**(1), 55–81.

Kakas, A. & Michael, L. 2020. Abduction and argumentation for explainable machine learning: a position survey. arXiv preprint arXiv:2010.12896.

Kakas, A. & Moraitis, P. 2003. Argumentation based decision making for autonomous agents. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, 883–890.

Kakas, A. & Moraitis, P. 2006. Adaptive agent negotiation via argumentation. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, 384–391.

Karafili, E., Kakas, A. C., Spanoudakis, N. I. & Lupu, E. C. 2017. Argumentation-based security for social good. In *2017* AAAI Fall Symposium Series.

Karafili, E., Lupu, E. C., Arunkumar, S. & Bertino, E. n.d.. Policy analysis for drone systems: an argumentation-based approach.

Karafili, E., Spanaki, K. & Lupu, E. C. 2018a. An argumentation reasoning approach for data processing. *Computers in Industry* **94**, 52–61.

Karafili, E., Wang, L., Kakas, A. C. & Lupu, E. 2018b. Helping forensic analysts to attribute cyber-attacks: an argumentation-based reasoner. In *International Conference on Principles and Practice of Multi-Agent Systems*, 510–518. Springer.

Karafili, E., Wang, L. & Lupu, E. C. 2020. An argumentation-based reasoner to assist digital investigation and attribution of cyber-attacks. *Forensic Science International: Digital Investigation* **32**, 300925.

Kemke, C. 2006. An architectural framework for natural language interfaces to agent systems. *In Computational Intelligence*, 371–376.

Keneni, B. M., Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaientz, J. D. & Marinier, R. P. 2019. Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access* **7**, 17001–17016.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Kobbe, J., Opitz, J., Becker, M., Hulpus, I., Stuckenschmidt, H. & Frank, A. 2019. Exploiting background knowledge for argumentative relation classification. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Kökciyan, N., Chapman, M., Balatsoukas, P., Sassoon, I., Essers, K., Ashworth, M., Curcin, V., Modgil, S., Parsons, S. & Sklar, E. I. 2019. A collaborative decision support tool for managing chronic conditions. *In MedInfo*, 644–648.

Kokciyan, N., Sassoon, I., Young, A. P., Chapman, M., Porat, T., Ashworth, C., Modgil, S., Parsons, S. & Sklar, E. 2018. Towards an argumentation system for supporting patients in self-managing their chronic conditions. In *AAAI*.

Kökciyan, N., Yaglikci, N. & Yolum, P. 2017. An argumentation approach for resolving privacy disputes in online social networks. *ACM Transactions on Internet Technology (TOIT)* **17**(3), 1–22.

Kökciyan, N. & Yolum, P. 2017. Context-based reasoning on privacy in internet of things. In *IJCAI*, 4738–4744.

Koshiyama, A., Kazim, E. & Engin, Z. 2019. Xai: digital ethics. In *HeXAI Workshop*.

Kraus, S., Sycara, K. & Evenchik, A. 1998. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence* **104**(1–2), 1–69.

Labrie, N. & Schulz, P. J. 2014. Does argumentation matter? a systematic literature review on the role of argumentation in doctor–patient communication. *Health Communication* **29**(10), 996–1008.

Laird, J. E. & Nielsen, E. 1994. Coordinated behavior of computer generated forces in TacAir-Soar. *AD-A280 063* **1001**, 57.

Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J. & Séroussi, B. 2019. Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. *Artificial Intelligence in Medicine* **94**, 42–53.

Langley, P. 2019. Explainable, normative, and justified agency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 9775–9779.

Lawrence, J. & Reed, C. 2020. Argument mining: a survey. *Computational Linguistics* **45**(4), 765–818.

Letia, I. A. & Groza, A. 2012. Interleaved argumentation and explanation in dialog. In *The 12th workshop on Computational Models of Natural Argument*, 44.

Levin, J. A. & Moore, J. A. 1977. Dialogue-games: metacommunication structures for natural language interaction. *Cognitive Science* **1**(4), 395–420.

Liao, B., Anderson, M. & Anderson, S. L. 2018. Representation, justification and explanation in a value driven agent: an argumentation-based approach. arXiv preprint arXiv:1812.05362.

Lifschitz, V. 2019. *Answer Set Programming*. Springer International Publishing.

Lippi, M. & Torroni, P. 2016. Argumentation mining: state of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)* **16**(2), 1–25.

Liu, X., Eshghi, A., Swietojanski, P. & Rieser, V. 2019. Benchmarking natural language understanding services for building conversational agents. arXiv preprint arXiv:1903.05566.

Lombrozo, T. 2006. The structure and function of explanations. *Trends in Cognitive Sciences* **10**(10), 464–470.

Longo, L. 2016. Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. In *Machine Learning for Health Informatics*, 183–208. Springer.

Longo, L. & Hederman, L. 2013. Argumentation theory for decision support in health-care: a comparison with machine learning. In *International Conference on Brain and Health Informatics*, 168–180, Springer.

Loui, R. P. & Norman, J. 1995. Rationales and argument moves. *Artificial Intelligence and Law* **3**(3), 159–189.

Lucero, M. J. G., Chesnevar, C. I. & Simari, G. R. 2009. On the accrual of arguments in defeasible logic programming. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Luis-Argentina, S. 2008. Decision rules and arguments in defeasible decision making. In *Computational Models of Argument: Proceedings of COMMA 2008*, **172**, 171.

Madhikermi, M., Malhi, A. K. & Främling, K. 2019. Explainable artificial intelligence based heat recycler fault detection in air handling unit. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 110–125. Springer.

Malle, B. F. 2006. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT Press.

Mayer, T., Cabrio, E., Lippi, M., Torroni, P. & Villata, S. 2018. Argument mining on clinical trials. *In COMMA*, 137–148.

McBurney, P. & Parsons, S. 2002. Dialogue games in multi-agent systems. *Informal Logic* **22**(3).

McBurney, P. & Parsons, S. 2009. Dialogue games for agent argumentation. In *Argumentation in Artificial Intelligence*, 261–280, Springer.

Mcburney, P., Van Eijk, R. M., Parsons, S. & Amgoud, L. 2003. A dialogue game protocol for agent purchase negotiations. *Autonomous Agents and Multi-Agent Systems* **7**(3), 235–273.

McCarty, L. T. 1976. Reflections on taxman: an experiment in artificial intelligence and legal reasoning. *Harvard Law Review* **90**, 837.

Meditskos, G., Kontopoulos, E., Vrochidis, S. & Kompatsiaris, I. 2019. Converness: ontology-driven conversational awareness and context understanding in multimodal dialogue systems. Expert Systems, e12378.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Melo, V. S., Panisson, A. R. & Bordini, R. H. 2016. Argumentation-based reasoning using preferences over sources of information. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 1337–1338.

Mercier, H. & Sperber, D. 2011. *Why do Humans Reason? Arguments for an Argumentative Theory*. Cambridge University Press.

Miller, T. 2019. Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence* **267**, 1–38.

Modgil, S. 2006a. Hierarchical argumentation. In *European Workshop on Logics in Artificial Intelligence*, 319–332. Springer.

Modgil, S. 2006b. Value based argumentation in hierarchical argumentation frameworks. *COMMA* **144**, 297–308.

Modgil, S. 2009. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence* **173**(9–10), 901–934.

Modgil, S., Budzynska, K. & Lawrence, J. 2018. Argument harvesting using chatbots. In *Computational Models of Argument: Proceedings of COMMA 2018*, 305, 149.

Modgil, S. & Prakken, H. 2014. The aspic+ framework for structured argumentation: a tutorial. *Argument & Computation* **5**(1), 31–62.

Modgil, S., Toni, F., Bex, F., Bratko, I., Chesñevar, C. I., Dvořák, W., Falappa, M. A., Fan, X., Gaggl, S. A., García, A. J., González, M. P., Gordon, T. F., Leite, J., Možina, M., Reed, C., Simari, G. R., Szeider, S., Torroni, P. & Woltran, S. 2013. The added value of argumentation. In *Agreement Technologies*, 357–403. Springer.

Moens, M.-F. 2016. Argumentation mining: how can a machine acquire world and common sense knowledge?. *In COMMA*, 4.

Moens, M.-F. 2018. Argumentation mining: how can a machine acquire common sense and world knowledge?. *Argument & Computation* **9**(1), 1–14.

Mollas, I., Bassiliades, N. & Tsoumakas, G. 2020. Altruist: argumentative explanations through local interpretations of predictive models. arXiv preprint arXiv:2010.07650.

Mosca, F., Sarkadi, S., Such, J. M. & McBurney, P. 2020. Agent expri: licence to explain. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 21–38. Springer.

Možina, M., Žabkar, J. & Bratko, I. 2007. Argument based machine learning. *Artificial Intelligence* **171**(10–15), 922–937.

Murukannaiah, P. K., Kalia, A. K., Telangy, P. R. & Singh, M. P. 2015. Resolving goal conflicts via argumentation-based analysis of competing hypotheses. In *2015* IEEE *23rd International Requirements Engineering Conference (RE)*, 156–165. IEEE.

Nicolaides, A. N., Kakkos, S. K., Kyriacou, E., Griffin, M., Sabetai, M., Thomas, D. J., Tegos, T., Geroulakos, G., Labropoulos, N., Doré, C. J., Morris, T. P., Naylor, R. & Abbott, A. L. 2010. Asymptomatic internal carotid artery stenosis and cerebrovascular risk stratification. *Journal of Vascular Surgery* **52**(6), 1486–1496.

Noël, V. & Kakas, A. 2009. Gorgias-c: extending argumentation with constraint solving. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, 535–541. Springer.

Nunes, E., Kulkarni, N., Shakarian, P., Ruef, A. & Little, J. 2016a. Cyber-deception and attribution in capture-the-flag exercises. In *Cyber Deception*, 149–165. Springer.

Nunes, E., Shakarian, P. & Simari, G. I. 2016b. Toward argumentation-based cyber attribution. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 177–184. AI Access Foundation.

Nunes, E., Shakarian, P., Simari, G. I. & Ruef, A. 2016c. Argumentation models for cyber attribution. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 837–844. IEEE.

Nute, D. 2001. Defeasible logic. In *International Conference on Applications of Prolog*, 151–169. Springer.

Oliveira, T., Dauphin, J., Satoh, K., Tsumoto, S. & Novais, P. 2018. Argumentation with goals for clinical decision support in multimorbidity. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*.

Ott, M., Cardie, C. & Hancock, J. T. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 497–501.

Páez, A. 2019. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines* **29**(3), 441–459.

Panisson, A. R. 2019. Towards an organisation-centred semantics for argumentation-based dialogues. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 491–496. IEEE.

Panisson, A. R., Ali, A., McBurney, P. & Bordini, R. H. 2018. Argumentation schemes for data access control. *In COMMA*, 361–368.

Panisson, A. R. & Bordini, R. H. 2016. Knowledge representation for argumentation in agent-oriented programming languages. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, 13–18. IEEE.

Panisson, A. R., Meneguzzi, F., Vieira, R. & Bordini, R. H. 2014. An approach for argumentation-based reasoning using defeasible logic in multi-agent programming languages. In *11th International Workshop on Argumentation in Multiagent Systems*, 1–15.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Panisson, A. R., Meneguzzi, F., Vieira, R. & Bordini, R. H. 2015. Towards practical argumentation-based dialogues in multi-agent systems. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, **2**, 151–158. IEEE.

Pavese, C. 2019. The semantics and pragmatics of argumentation. Academia.

Pilotti, P., Casali, A. & Chesñevar, C. 2015. A belief revision approach for argumentation-based negotiation agents. *International Journal of Applied Mathematics and Computer Science* **25**(3), 455–470.

Pocevičiūtė, M., Eilertsen, G. & Lundström, C. 2020. Survey of XAI in digital pathology. In *Artificial Intelligence and Machine Learning for Digital Pathology*, 56–88, Springer.

Potash, P., Bhattacharya, R. & Rumshisky, A. 2017. Length, interchangeability, and external knowledge: observations from predicting argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume* **1:** *Long Papers)*, 342–351.

Prakken, H. 2005a. Ai & law, logic and argument schemes. *Argumentation* **19**(3), 303–320.

Prakken, H. 2005b. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation* **15**(6), 1009–1040.

Prakken, H. 2017. *Logics of Argumentation and the Law*. Cambridge University Press.

Prakken, H. & Sartor, G. 1998. Modelling reasoning with precedents in a formal dialogue game. In *Judicial Applications of Artificial Intelligence*, 127–183. Springer.

Prakken, H., Wyner, A., Bench-Capon, T. & Atkinson, K. 2015. A formalization of argumentation schemes for legal case-based reasoning in aspic+. *Journal of Logic and Computation* **25**(5), 1141–1166.

Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A. & Pattichis, C. 2019. Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 817–821. IEEE.

Qurat-ul-ain Shaheen, A. T. & Bowles, J. K. 2020. Dialogue games for explaining medication choices. In *Rules and Reasoning: 4th International Joint Conference, RuleML+ RR 2020, Oslo, Norway, June 29–July 1, 2020, Proceedings*, 97. Springer Nature.

Rago, A., Cocarascu, O. & Toni, F. 2018. Argumentation-based recommendations: fantastic explanations and how to find them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

Rahwan, I., Banihashemi, B., Reed, C., Walton, D. & Abdallah, S. 2011. Representing and classifying arguments on the semantic web. *The Knowledge Engineering Review* **26**(4), 487–511.

Reed, C., Wells, S., Budzynska, K. & Devereux, J. 2010. Building arguments with argumentation: the role of illocutionary force in computational models of argument. *In COMMA*, 415–426.

Regulation, P. 2016. *Regulation (eu) 2016/679 of the European Parliament and of the Council. REGULATION (EU)*, 679.

Ripley, M. L. 2005. Arguing for the ethics of an ad: an application of multi-modal argumentation theory.

Rissland, E. L. & Ashley, K. D. 1987. A case-based system for trade secrets law. In *Proceedings of the 1st International Conference on Artificial Intelligence and Law*, 60–66.

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.

Rosenfeld, A. & Kraus, S. 2016b. Strategical argumentative agent for human persuasion. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, 320–328. IOS Press.

Rowe, J., Levitt, K., Parsons, S., Sklar, E., Applebaum, A. & Jalal, S. 2012. Argumentation logic to assist in security administration. In *Proceedings of the 2012 New Security Paradigms Workshop*, 43–52.

Sakama, C. 2018. Abduction in argumentation frameworks. *Journal of Applied Non-Classical Logics* **28**(2–3), 218–239.

Samadi, M., Talukdar, P., Veloso, M. & Blum, M. 2016. Claimeval: integrated and flexible framework for claim evaluation using credibility of sources. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Samek, W. & Müller, K.-R. 2019. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 5–22. Springer.

Santini, F. & Yautsiukhin, A. 2015. Quantitative analysis of network security with abstract argumentation. In *Data Privacy Management, and Security Assurance*, 30–46. Springer.

Sassoon, I., Kökciyan, N., Sklar, E. & Parsons, S. 2019. Explainable argumentation for wellness consultation. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 186–202. Springer.

Schoenborn, J. M. & Althoff, K.-D. 2019. Recent trends in XAI: a broad overview on current approaches, methodologies and interactions. In *ICCBR Workshops*, 51–60.

Schreier, M., Groeben, N. & Christmann, U. 1995. 'that's not fair! argumentational integrity as an ethics of argumentative communication. *Argumentation* **9**(2), 267–289.

Schulz, C. & Toni, F. 2016. Justifying answer sets using argumentation. *Theory and Practice of Logic Programming* **16**(1), 59–110.

Šešelja, D. & Straßer, C. 2013. Abstract argumentation and explanation applied to scientific debates. *Synthese* **190**(12), 2195–2217.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Shakarian, P., Simari, G. I., Moores, G. & Parsons, S. 2015. Cyber attribution: an argumentation-based approach. In *Cyber Warfare*, 151–171. Springer.

Sheh, R. K.-M. 2017. "Why did you do that?" explainable intelligent robots. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Sklar, E. I. & Azhar, M. Q. 2015. Argumentation-based dialogue games for shared control in human-robot systems. *Journal of Human-Robot Interaction* **4**(3), 120–148.

Sklar, E. I. & Azhar, M. Q. 2018. Explanation through argumentation. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, 277–285.

Sklar, E. I., Parsons, S., Li, Z., Salvit, J., Perumal, S., Wall, H. & Mangels, J. 2016. Evaluation of a trust-modulated argumentation-based interactive decision-making tool. *Autonomous Agents and Multi-Agent Systems* **30**(1), 136–173.

Sklar, E., Parsons, S. & Singh, M. P. 2013. Towards an argumentation-based model of social interaction. In *Proceedings of the Workshop on Argumentation in Multiagent Systems (ArgMAS) at the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Slugoski, B. R., Lalljee, M., Lamb, R. & Ginsburg, G. P. 1993. Attribution in conversational context: effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology* **23**(3), 219–238.

Smullyan, R. M. 1995. *First-Order Logic*. Courier Corporation.

Snaith, M., Lawrence, J. & Reed, C. 2010. Mixed initiative argument in public deliberation. *Online Deliberation*, 2.

Sørmo, F., Cassens, J. & Aamodt, A. 2005. Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review* **24**(2), 109–143.

Spanoudakis, G., Kloukinas, C. & Androutsopoulos, K. 2007. Towards security monitoring patterns. In *Proceedings of the 2007 ACM Symposium on Applied Computing*, 1518–1525.

Spanoudakis, N. I., Constantinou, E., Koumi, A. & Kakas, A. C. 2017. Modeling data access legislation with gorgias. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 317–327. Springer.

Spanoudakis, N. I., Kakas, A. C. & Moraitis, P. 2016a. Applications of argumentation: the soda methodology. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, 1722–1723.

Spanoudakis, N. I., Kakas, A. C. & Moraitis, P. 2016b. Gorgias-b: argumentation in practice. *In COMMA*, 477–478.

Spanoudakis, N. & Moriaitis, P. 2009. Engineering an agent-based system for product pricing automation. *Engineering Intelligent Systems* **17**(2), 139.

Swanson, R., Ecker, B. & Walker, M. 2015. Argument mining: extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 217–226.

Tang, Y., Cai, K., McBurney, P., Sklar, E. & Parsons, S. 2012. Using argumentation to reason about trust and belief. *Journal of Logic and Computation* **22**(5), 979–1018.

Tang, Y., Sklar, E. & Parsons, S. 2012. An argumentation engine: argtrust. In *Ninth International Workshop on Argumentation in Multiagent Systems*.

Thimm, M. & Kersting, K. 2017. Towards argumentation-based classification. In Logical Foundations of Uncertainty and Machine Learning, Workshop at IJCAI, 17.

Tjoa, E. & Guan, C. 2019. A survey on explainable artificial intelligence (XAI): towards medical XAI. arXiv preprint arXiv:1907.07374.

Torres, I., Hernández, N., Rodrguez, A., Fuentes, G. & Pineda, L. A. 2019. Reasoning with preferences in service robots. *Journal of Intelligent & Fuzzy Systems* **36**(5), 5105–5114.

Toulmin, S. 1958. *The Uses of Argument.* Cambridge University Press.

Vassiliades, A., Patkos, T., Bikakis, A., Flouris, G., Bassiliades, N. & Plexousakis, D. 2020. Preliminary notions of arguments from commonsense knowledge. In *11th Hellenic Conference on Artificial Intelligence*, 211–214.

Verheij, B. 2003. Artificial argument assistants for defeasible argumentation. *Artificial Intelligence* **150**(1–2), 291–324.

Waltl, B. & Vogl, R. 2018. Explainable artificial intelligence the new frontier in legal informatics. *Jusletter IT* **4**, 1–10.

Walton, D. 2005. *Argumentation Methods for Artificial Intelligence in Law*. Springer Science & Business Media.

Wanner, L., André, E., Blat, J., Dasiopoulou, S., Farrùs, M., Fraga, T., Kamateri, E., Lingenfelser, F., Llorach, G., Martínez, O., Meditskos, G., Mille, S., Minker, W., Pragst, L., Schiller, D., Stam, A., Stellingwerff, L., Sukno, F., Vieru, B. & Vrochidis, S. 2017. Kristina: a knowledge-based virtual conversation agent. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, 284–295. Springer.

Wardeh, M., Wyner, A., Atkinson, K. & Bench-Capon, T. 2013. Argumentation based tools for policy-making. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, 249–250.

Wick, M. R. & Thompson, W. B. 1992. Reconstructive expert system explanation. *Artificial Intelligence* **54**(1–2), 33–70.

Willmott, S., Vreeswijk, G., Chesnevar, C., South, M., McGinnis, J., Modgil, S., Rahwan, I., Reed, C. & Simari, G. 2006. Towards an argument interchange format for multiagent systems. In *3rd International Workshop on Argumentation in Multi-Agent Systems, ArgMAS-06*, 17–34.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press

Wooldridge, M. 2009. *An Introduction to Multiagent Systems*. John Wiley & Sons.

Wooldridge, M. & Van Der Hoek, W. 2005. On obligations and normative ability: towards a logical analysis of the social contract. *Journal of Applied Logic* **3**(3–4), 396–420.

Wyner, A. Z., Atkinson, K. & Bench-Capon, T. 2012a. Model based critique of policy proposals. In *International Conference on Electronic Participation*, 120–131, Springer.

Wyner, A. Z., Atkinson, K. & Bench-Capon, T. J. 2012b. Opinion gathering using a multi-agent systems approach to policy selection. In *AAMAS*, 1171–1172.

Yang, S. C.-H. & Shafto, P. 2017. Explainable artificial intelligence via bayesian teaching. In *NIPS 2017 Workshop on Teaching Machines, Robots, and Humans*.

Zeng, Z., Fan, X., Miao, C., Leung, C., Jih, C. J. & Soon, O. Y. 2018. Context-based and explainable decision making with argumentation. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1114–1122. International Foundation for Autonomous Agents and Multiagent Systems.

Zeng, Z., Miao, C., Leung, C. & Chin, J. J. 2018. Building more explainable artificial intelligence with argumentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, S., Rudinger, R., Duh, K. & Van Durme, B. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics* **5**, 379–395.

Zhong, Q., Fan, X., Luo, X. & Toni, F. 2019. An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications* **117**, 42–61.

Zhong, Q., Fan, X., Toni, F. & Luo, X. 2014. Explaining best decisions via argumentation. In *ECSI*, 224–237.

https://doi.org/10.1017/S0269888921000011 Published online by Cambridge University Press