Argumentative XAI: A Survey

Kristijonas Čyras[1*], **Antonio Rago**[2], **Emanuele Albini**[2], **Pietro Baroni**[3] and **Francesca Toni**[2]
[1]Ericsson Research, Sweden
[2]Department of Computing, Imperial College London, UK
[3]Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
kristijonas.cyras@ericsson.com, {a.rago, emanuele, ft}@imperial.ac.uk, pietro.baroni@unibs.it

Abstract

Explainable AI (XAI) has been investigated for decades and, together with AI itself, has witnessed unprecedented growth in recent years. Among various approaches to XAI, argumentative models have been advocated in both the AI and social science literature, as their dialectical nature appears to match some basic desirable features of the explanation activity. In this survey we overview XAI approaches built using methods from the field of *computational argumentation*, leveraging its wide array of reasoning abstractions and explanation delivery methods. We overview the literature focusing on different types of explanation (intrinsic and post-hoc), different models that argumentation-based explanations are deployed with, different forms of delivery, and different argumentation frameworks they use. We also lay out a roadmap for future work.

1 Introduction

Explainable AI (XAI) has attracted a considerable amount of attention in recent years, due mostly to its role in providing a bridge between applications of AI and humans who develop or use them. Several approaches to support XAI have been proposed (see e.g. some recent overviews [Adadi and Berrada, 2018; Guidotti *et al.*, 2019]) and the crucial role of XAI in human-machine settings emphasised [Rosenfeld and Richardson, 2019]. Whereas several recent efforts in XAI are focused on explaining machine learning models [Adadi and Berrada, 2018], XAI has also been a recurrent concern in other AI settings, e.g. expert systems [Swartout *et al.*, 1991], answer set programming [Fandinno and Schulz, 2019] and planning [Chakraborti *et al.*, 2020].

We provide a comprehensive survey of literature in XAI viewing explanations as *argumentative* (independently of the underlying methods to be explained). Argumentative explanations are advocated in the social sciences (e.g. [Antaki and Leudar, 1992]), focusing on the human perspective, and argumentation's potential advantages for XAI have been pointed out by several, e.g. [Moulin *et al.*, 2002; Bex and Walton, 2016; Sklar and Azhar, 2018].

Many methods for generating explanations in XAI can be seen as argumentative. Indeed, attribution methods, including model-agnostic [Lundberg and Lee, 2017] and model-specific [Shih *et al.*, 2018] approaches, link inputs to outputs via (weighted) positive and negative relations, and contrastive explanations identify reasons pro and con outputs [Chakraborti *et al.*, 2020; Miller, 2019]. In this survey we focus instead on overtly argumentative approaches, with an emphasis on the several existing XAI solutions using forms of *computational argumentation* (see [Baroni *et al.*, 2018] for a recent overview of this field of symbolic AI).

The application of computational argumentation to XAI is supported by its strong theoretical and algorithmic foundations and the flexibility it affords, particularly in the wide variety of *argumentation frameworks* (AFs) on offer in the literature. These AFs include ways to specify *arguments* and *dialectical relations* between them, as well as *semantics* to evaluate the dialectical *acceptability* or *strength* of arguments, while differing (sometimes substantially) in how they define these components. When AFs are used to obtain explanations, (weighted) arguments and dialectical relations may suitably represent anything from input data, e.g. categorical data or pixels in an image, to knowledge, e.g. rules, to internal components of the method being explained, e.g. filters in convolutional neural networks, to problem formalisations, e.g. planning, scheduling or decision making models, to outputs, e.g. classifications, recommendations, or logical inference. This flexibility and wide-ranging applicability has led to a multitude of methods for *AF-based explanations*, providing the motivation and need for this survey.

After giving some background (§2), our contributions are:

- we overview the literature on AF-based explanations, cataloguing representative approaches according to what they explain and by means of which AF (§3);
- we overview the prevalent forms which AF-based explanations take after being drawn from AFs (§4);
- we lay out a roadmap for future work, covering: the need to focus on properties of AF-based explanations, computational aspects, and further applications and other potential developments of AF-based explanations (§5).

Note that we ignore, throughout, argumentative explanations based on informal notions or conceptual models lacking one or more aspects of AFs (notably, notions of semantics).

*Contact Author

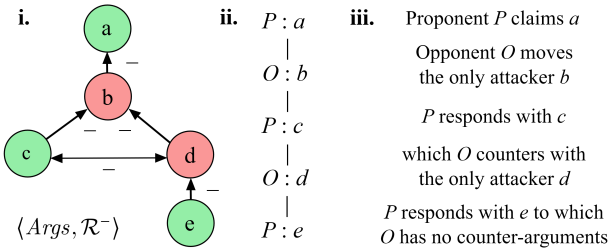

Figure 1: (i) An AA framework $\langle Args, \mathcal{R}^- \rangle$ visualised as a graph, with nodes as arguments (labelled a, b, c, d, e) and directed edges as attacks (labelled $-$). Here, the extension $\{a, b\}$ is not conflict-free, but $\{a, c\}$ is. Arguments (not) in the grounded extension [Dung, 1995] are coloured green (resp. red). (ii) A *grounded dispute tree* (DT) [Dung *et al.*, 2007] between the proponent P and opponent O, for the *topic* argument a. (Intuitively: a is defended from b by c; e defends c from d.) (iii) A simple *dialogical* explanation where the proponent P wins (drawn from the DT in (ii)).

2 Argumentation Frameworks

In this section we give a brief, high-level overview of the argumentation frameworks (AFs) *used so far to support XAI* in the literature, focusing on how these various AFs understand *arguments*, *dialectical relations* and *semantics*.

A first category of approaches sees arguments as *abstract* entities (in some given set $Args$), starting from **Abstract Argumentation (AA)** [Dung, 1995]. Here, a single dialectical relation of *attack* ($\mathcal{R}^- \subseteq Args \times Args$) is considered and semantics are given in terms of mappings from AA frameworks $\langle Args, \mathcal{R}^- \rangle$ to sets of so-called *extensions*, each a set of arguments $\mathcal{E} \subseteq Args$ fulfilling some specified, dialectically meaningful constraints (e.g. *conflict-freeness*, amounting to absence of internal conflicts: $\{\alpha, \beta \in \mathcal{E} | (\alpha, \beta) \in \mathcal{R}^-\} = \varnothing$). Figure 1i illustrates AA frameworks, conflict-freeness and the widely used *grounded extension semantics*. Some of the extension-based semantics can be equivalently understood in terms of *dispute trees* **(DTs)** [Dung *et al.*, 2007] (as illustrated in Figure 1ii), which form the basis of several approaches to argumentation-based XAI (e.g. by providing content for the simple dialogical explanations illustrated in Figure 1iii).

The category of AFs with arguments as abstract entities counts several other approaches. In **Bipolar Argumentation (BA)** [Cayrol and Lagasquie-Schiex, 2005] an additional dialectical relation of *support* ($\mathcal{R}^+ \subseteq Args \times Args$) is considered, and a variety of semantics in terms of extensions given (reflecting, in particular, different interpretations of support [Cohen *et al.*, 2014]). BA frameworks where the attack relation is empty are also called **Support Argumentation (SA)** frameworks. BA (and thus AA) frameworks can be also equipped with *gradual semantics*, in terms of mappings from frameworks to value assignments for arguments, representing their *dialectical strength*. These values may belong to any given (ordered) set (e.g. $[0, 1]$) and may be influenced by other factors, e.g. in **Quantitative Bipolar Argumentation (QBA)** [Baroni *et al.*, 2019], by values ascribing to arguments their initial, *intrinsic strength*. Some AFs consider further dialectical relations, e.g. **Tripolar Argumentation (TA)** [Rago *et al.*, 2018] uses a *neutralising* relation (intuitively pushing arguments' strengths towards the neutral value) and **Generalised Argumentation (GA)** [Gabbay, 2016] can in principle use any number of dialectical relations. Further, among the approaches with abstract arguments, **Abstract Dialectical Frameworks (ADFs)** [Brewka and Woltran, 2010] allow for generalised notions of dialectical relations and semantics, specified in terms of user-given *acceptance conditions*.

A second category of AFs focuses on *structured* arguments instead. Among those used for explanation: **ABA** [Bondarenko *et al.*, 1997] sees arguments as deductions from assumptions using rules, whereas **ASPIC+** [Modgil and Prakken, 2014] and **DeLP** [García *et al.*, 2013] draw arguments from strict and defeasible rules. These approaches, like AA, consider only attacks between arguments. Further, ABA and ASPIC+, like AA, use extension-based semantics, whereas DeLP uses *dialectical trees*, similar in spirit to DTs.

Finally, several approaches relevant to this paper use instances of AA for specific choices of (structured) arguments. These may be: (1) case-based reasoning (CBR)-inspired, with arguments as past or new cases (where features support a given or open label), as in [Cyras *et al.*, 2019a; Cocarascu *et al.*, 2020]; (2) deductions in abductive logic programming (ALP), as in [Wakaki *et al.*, 2009]; (3) logical deductions (LD), possibly built from rules, as in [Arioua *et al.*, 2015; Brarda *et al.*, 2019; Sendi *et al.*, 2019]; or built from argument schemes (AS), as in [Sassoon *et al.*, 2019]. We refer to AA instances using arguments of these forms, resp., as **AA-CBR**, **AA-ALP**, **AA-LD**, and **AA-AS** frameworks.

3 Types of Argumentative Explanations

Here, we review the literature for models explained using argumentative explanations *built from AFs*, referred to as *AF-based explanations*. We divide them into those which are:

- *intrinsic*, i.e. defined for an interpretable model that is natively built on argumentative foundations, covered in §3.1;
- *post-hoc*, i.e. obtained by means of some argumentative post-processing activity on a non-argumentative model; we further divide the post-hoc AF-based explanations depending on whether they provide a *complete* (§3.2) or *approximate* (§3.3) representation of the explained model.

Note that this three-fold distinction among types of AF-based explanation is not crisp, and, as we will see, some of the approaches we survey may be deemed to be hybrids.

We use the term 'model' in a very general sense, in the spirit of [Geffner, 2018], to stand for a variety of systems, amounting, in our review, to the following categories: recommender systems, classifiers and probabilistic methods, decision-making and knowledge-based systems, planners and schedulers, as well as tools for logic programming.

Note also that our focus in this section is on explaining models *other than* the argumentation process itself, while we will overview later, in §4, the forms of (intrinsic) AF-based explanations adopted by some approaches to explain argumentation/argumentative agents. Table 1 provides an overview of the surveyed approaches, distinguished according to the category of model for which they provide intrinsic, complete or approximate types of AF-based explanations, and indicating the form of AF the explanations are obtained from. Due to space limitations we do not provide an exhaus-

	Intrinsic	Compl.	Approx.
Recommender Systems			
[Briguez *et al.*, 2014]	DeLP		
[Rodríguez *et al.*, 2017]	DeLP		
[Rago *et al.*, 2018]			TA
[Cocarascu *et al.*, 2019]	QBA		
[Rago *et al.*, 2020]			BA
Classification			
[Cyras *et al.*, 2019a]	AA-CBR		
[Sendi *et al.*, 2019]			AA-LD
[Cocarascu *et al.*, 2020]	AA-CBR		
[Dejl *et al.*, 2021]			GA
Probabilistic Methods			
[Timmer *et al.*, 2017]			SA
[Albini *et al.*, 2020a]		QBA	
Decision Making			
[Amgoud and Prade, 2009]	AA		
[Zeng *et al.*, 2018]		ABA	
[Brarda *et al.*, 2019]	AA-LD		
[Zhong *et al.*, 2019]		ABA	
Knowledge-Based Systems			
[Arioua *et al.*, 2015]		AA-LD	
[Kökciyan *et al.*, 2020]	AA, ASPIC+		
Planning & Scheduling			
[Fan, 2018]		ABA	
[Cyras *et al.*, 2019b]		AA	
[Collins *et al.*, 2019]			ASPIC+
[Oren *et al.*, 2020]	ASPIC+		
Logic Programming			
[Wakaki *et al.*, 2009]		AA-ALP	
[Schulz and Toni, 2016]		ABA	
[Rolf *et al.*, 2019]		ADF	

Table 1: Overview of AF-based explanation approaches (divided wrt class of model, category of explanation, and form of AF used).

tive catalogue of approaches to AF-based explanations in the literature, but exemplify, for each of our chosen categories of models, some representative works.

3.1 Intrinsic Approaches

One application where intrinsic AF-based explanations are popular is **recommender systems** (RSs). Various RSs have been built with DeLP as the main recommendation and explanation engine. One is that of [Briguez *et al.*, 2014] for the movie domain (see Figure 2), handling incomplete and contradictory information and using a comparison criterion to solve conflicting situations. Another is introduced by [Rodríguez *et al.*, 2017], deploying DeLP to provide a hybrid RS in an educational setting, using argumentation to differentiate between different techniques for generating recommendations. Other approaches deploy other forms of AFs. For example, [Cocarascu *et al.*, 2019] provide argumentation-based review aggregations for movie recommendations and conversational explanations extracted from QBA frameworks, in turn extracted from reviews by natural language processing.

[Cyras *et al.*, 2019a] propose a case-based reasoning (CBR)-inspired but AA-driven **classification** method for

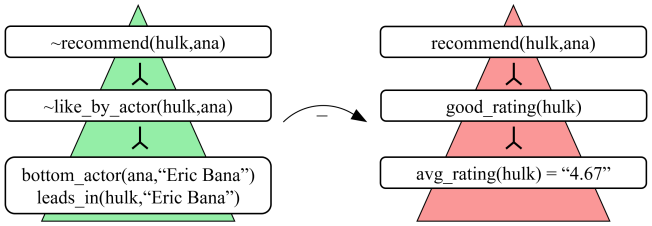

Figure 2: Intrinsic AF-based explanation for the RS of [Briguez *et al.*, 2014] explaining why the movie *Hulk* was not recommended to the user *Ana*. The undefeated argument against this recommendation (left) attacks (labelled −) the argument for the recommendation (right), which is thus defeated. DeLP structured arguments are constructed of statements linked by defeasible rules (indicated by ⩚).

classifying and explaining legislation outcomes. [Cocarascu *et al.*, 2020] use a similarly-inspired method for classifying and explaining with a variety of (binary) classification tasks. In all these settings, classification results from the semantics of AA frameworks constructed from cases and explanations are based on DTs. Whereas in [Cyras *et al.*, 2019a] cases amount to categorical data, in [Cocarascu *et al.*, 2020] they may be unstructured (e.g. textual).

[Amgoud and Prade, 2009] define a method for argumentative **decision making**, where AA is used to evaluate the acceptance of arguments for and against each potential decision, and decisions are selected based on their arguments' acceptance, with the AA frameworks constituting the explanations. The AA-based, multi-criteria decision support system of [Brarda *et al.*, 2019] operates similarly but using an instance of AA, whereby arguments are obtained from conditional preference rules. [Kökciyan *et al.*, 2020] use AA and argument schemes mappable to ASPIC+ to produce domain-specific explanations for **knowledge-based systems**. To conclude this section, [Oren *et al.*, 2020] devise an explainable system for argument-based **planning**, using a variant of ASPIC+ to translate domain rules from a knowledge base to arguments which feed into explanations in dialogical format.

3.2 Complete Post-Hoc Approaches

Complete post-hoc AF-based explanations are also deployed in a variety of settings, for several AFs. With regards to **probabilistic methods**, [Albini *et al.*, 2020a] represent web graphs, with webpages as nodes and (hyper)links as edges, with equivalent QBA frameworks in order to explain PageRank scores [Page *et al.*, 1998] argumentatively (see Figure 3).

A popular area for post-hoc AF-based explanations is **decision making**, where ABA frameworks have been shown to represent and explain decision models. [Zeng *et al.*, 2018] model decision problems as decision graphs with context before mapping them into ABA, where 'best' decisions correspond to a specific type of (admissible) extensions. [Zhong *et al.*, 2019] show that decision problems can be mapped to equivalent ABA frameworks, such that minimally redundant decisions correspond to (admissible) extensions.

In **knowledge-based systems**, [Arioua *et al.*, 2015] provide explanations for the purpose of query answering in Datalog under inconsistency-tolerant semantics. Here, entailment of a query from a knowledge base is mapped in a

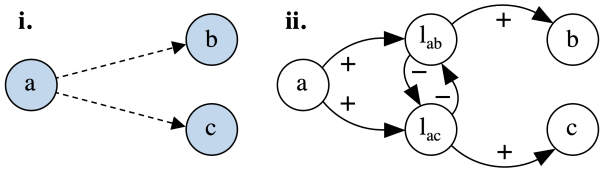

Figure 3: Complete post-hoc explanations for PageRank [Albini *et al.*, 2020a]. (i) Graphical representation of three webpages as nodes, with edges showing (hyper)links between them. (ii) Extracted QBA framework, where arguments amount to webpages (a, b, c) and links l_{xy} (from x to y), the strength of an argument (omitted) is computed by the PageRank algorithm, and attacks (labelled −) represent negative effects while supports (labelled +) represent positive effects on this strength. It can be seen that a's outgoing links 'compete', thus explaining how its PageRank score gets distributed to b and c.

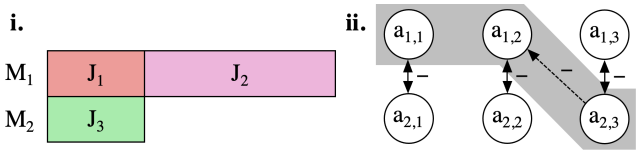

Figure 4: Complete post-hoc AF-based explanations for scheduling [Cyras *et al.*, 2019b]. (i) An inefficient schedule with jobs J_1, J_2 assigned to machine M_1 and J_3 to M_2 is in 1-1 correspondence with (ii) the non-conflict-free extension (in grey) in the corresponding AA framework where argument $a_{i,j}$ represents assignment of J_j to M_i and attacks capture scheduling constraints. The (dashed) attack underlies an explanation as to why the schedule is inefficient: jobs J_2 and J_3 can be swapped between machines M_2 and M_1.

sound and complete fashion onto argument acceptance in AA frameworks constructed from the rules of the knowledge base. Intuitively, argumentative explanations for query entailment/failure result from *defence trees* (similar to DTs).

For **planning**, [Fan, 2018] generates an ABA counterpart of planning problems such that extensions of the former correspond to solutions to the latter. Also, **scheduling** has been targeted by [Cyras *et al.*, 2019b], where a makespan scheduling problem with user decisions is translated into AA to extract sound and complete explanations as to why a schedule is (not) feasible, efficient and/or satisfies user decisions (see Figure 4). Here, explanations are actionable in suggesting actions for fixing problems, if any exist.

Several complete post-hoc approaches have been proposed for **logic programming**. For example, [Wakaki *et al.*, 2009] show that exact mappings exist between answer sets and AA-ALP, whereas [Schulz and Toni, 2016] explain (non-)membership in answer sets in terms of ABA, based on sound and complete mappings between answer sets and extensions. A similar approach is taken by [Rolf *et al.*, 2019], who, however, use extensions for ADFs.

3.3 Approximate Post-Hoc Approaches

Approximate post-hoc AF-based explanations rely upon incomplete mappings between the model to be explained and the AF from which explanations are drawn. For RSs, [Rago *et al.*, 2018] use TA frameworks extracted from a hybrid RS as the basis for explanations. This RS exploits collaborative

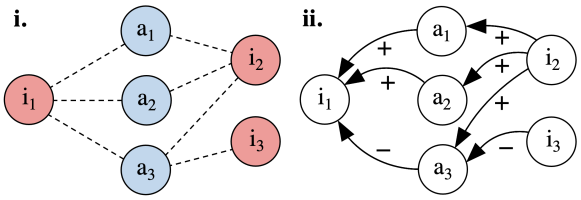

Figure 5: Approximate post-hoc AF-based explanations for RSs [Rago *et al.*, 2020]. (i) Connections between aspects and items underpinning the RS (predicted ratings omitted). (ii) Example BA framework extracted to explain the recommendation of i_1, where attacks (labelled −) represent negative effects and supports (labelled +) represent positive effects on predicted ratings.

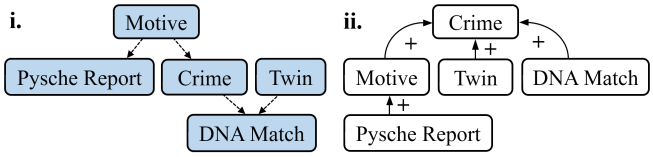

Figure 6: Approximate post-hoc AF-based explanations for Bayesian networks [Timmer *et al.*, 2017]. (i) Bayesian network with conditional dependencies (conditional probabilities ignored). (ii) Extracted SA framework, where the support relation is directly derived from each variable's Markov blanket.

filtering and connections between items and their aspects to compute predicted ratings for items propagated through the connections. The TA framework approximates this propagation, categorising connections as dialectical relations if they lead to satisfaction of specified properties of TA semantics. A similar RS with a greater explanatory repertoire, but supported by BA, is given by [Rago *et al.*, 2020] (see Figure 5).

Another category of model explained by approximated AF-based explanations are **classification** methods. [Sendi *et al.*, 2019] extract rules from neural networks in an ensemble and deploy AA to explain their outputs. Neural methods are also targeted by the general formalism of deep argumentative explanations [Dejl *et al.*, 2021], which map neural networks to GA frameworks to explain their outputs in terms of inputs and, affording a *deep* nature, intermediate components.

Approximated AF-based explanations have also been extracted from **probabilistic methods**, including Bayesian networks: for example, [Timmer *et al.*, 2017] model explanations as SA frameworks to show the interplay between variables in Bayesian networks (see Figure 6).

Finally, to conclude our overview of approaches for obtaining AF-based explanations, [Collins *et al.*, 2019] take preliminary steps towards extracting approximate post-hoc AF-based explanations in **planning**: causal relationships are first extracted from a plan and then abstracted into ASPIC+.

4 Forms of AF-Based Explanations

In §3 we have discussed and seen examples of three types of AF-based explanations. Here, we discuss various forms of argumentation-based structures that are used to formally define explanations in stand-alone AFs as well as AF-based agents. These could be (and often are) used in turn to support

AF-based explanations in their settings of use. We summarise the most representative such works in Table 2 and illustrate the various forms of explanations by referring to Figures 1–6.

	Form of Explanation (AF)
AFs	
[Wakaki *et al.*, 2009]	Dialogue (AA, AA-ALP)
[Modgil and Caminada, 2009]	Dialogue (AA)
[Seselja and Straßer, 2013]	Sub-graph (variant of AA)
[Grando *et al.*, 2013]	Sub-graph (ASPIC+)
[García *et al.*, 2013]	Sub-graph (DeLP)
[Booth *et al.*, 2014]	Change, Dialogue (AA, AA-ALP)
[Fan and Toni, 2015b]	Extensions, DTs (AA, ABA)
[Fan and Toni, 2015a]	Change (AA)
[Cyras *et al.*, 2019a]	DTs (AA-CBR)
[Arioua *et al.*, 2017]	Dialogue (AA-LD)
[Sakama, 2018]	Change (AA)
[Zeng *et al.*, 2019]	DTs (BA)
[Saribatur *et al.*, 2020]	Change (AA)
[Liao and van der Torre, 2020]	Extensions (AA)
AF-Based Agents	
[Gao *et al.*, 2016]	Extensions (AA)
[Sklar and Azhar, 2018]	Dialogue (AA-LD)
[Madumal *et al.*, 2019]	Dialogue (AA-LD)
[Raymond *et al.*, 2020]	Dialogue (AA-LD)
[Sassoon *et al.*, 2019]	Dialogue (AA-AS)
[Kökciyan *et al.*, 2020]	Structure (AA-LD, ASPIC+)
[Espinoza *et al.*, 2020]	Sub-graph, Extensions (AA-LD)

Table 2: Forms of explanations drawn from various forms of AF (stand-alone or in multi-agent contexts).

A common approach to explaining argument acceptability (i.e. membership in extensions) in AFs essentially amounts to traversing the AFs to show any arguments and elements of the dialectical relations relevant to determining a given argument's acceptance status. When the AFs can be understood as graphs (see Figure 1i), this process amounts to identifying **sub-graphs** in the AFs, while also requiring satisfaction of some formal property (such as being a tree rooted in the argument whose acceptability status needs explaining, alternating attacking and defending arguments and with unattacked arguments as leaves). Formally [Cocarascu *et al.*, 2018], given a property P of graphs and an AA framework $\mathcal{G} = \langle Args, \mathcal{R}^- \rangle$, an explanation for the acceptance status of some *topic argument* $a \in Args$ can be defined as a sub-graph $\mathcal{G}' = \langle Args', \mathcal{R}^{-\prime} \rangle$ of $\mathcal{G}$ such that $a \in Args'$ and $\mathcal{G}'$ satisfies P.

A popular form of explanation as sub-graph is given by **dispute trees** (DTs, see §2 and Figure 1ii). Importantly, DTs often carry theoretical guarantees towards desirable properties of explanations [Modgil and Caminada, 2009; Fan and Toni, 2015b; Cyras *et al.*, 2019a], such as existence, correctness (the explanation actually establishes the acceptance status of the topic argument) and relevance (all arguments in the explanation play a role towards acceptance).

AF-based explanations as sub-graphs can also take forms of (sets of) paths, cycles or branches [Seselja and Straßer, 2013; García *et al.*, 2013; Timmer *et al.*, 2017; Cocarascu *et al.*, 2018; Cyras *et al.*, 2019b; Espinoza *et al.*, 2020]. They are used with various AFs, e.g. BA [Cocarascu *et al.*, 2018; Rago *et al.*, 2018], QBA [Cocarascu *et al.*, 2019] and DeLP [García *et al.*, 2013] (see Figures 2–6 for examples).

Sub-graphs, especially DTs, act as proofs for argument acceptance [Modgil and Caminada, 2009]. In a similar vain, in structured argumentation the (logic- or argument scheme-based) **structure** of arguments and dialectical relations can act as an explanation [Moulin *et al.*, 2002; Grando *et al.*, 2013; Naveed *et al.*, 2018; Kökciyan *et al.*, 2020].

Whether comprising structured or abstract arguments, AFs naturally give rise to another form of AF-based explanation, namely a **dialogue** (**game**), whereby 'participants engage in structured, rule-guided, goal-oriented exchange' [Sklar and Azhar, 2018] of arguments in order to establish or explain argument acceptability. Dialogues can be constructed from DTs (see Figure 1iii) or AFs in general [Wakaki *et al.*, 2009; Booth *et al.*, 2014; Arioua *et al.*, 2017; Madumal *et al.*, 2019; Raymond *et al.*, 2020; Walton, 2004], typically as formal games between two parties, *proponent* and *opponent*, with the aim of 'winning' the game regarding the topic argument.

Explanations can also be defined as **extensions** (see §2), keeping the relationships implicit [Fan and Toni, 2015b; Gao *et al.*, 2016; Zeng *et al.*, 2018; Espinoza *et al.*, 2020; Liao and van der Torre, 2020]. For instance, [Fan and Toni, 2015b] define explanations as *related admissible* extensions in AA frameworks, which are conflict-free extensions defending against all attackers and in which one (topic) argument is defended by all other arguments. For instance, in Figure 1i, related admissible extensions $\{a, c\}$ and $\{a, c, e\}$ are explanations of (the acceptability of) the topic argument a.

Another form of AF-based explanation amounts to indicating which modifications to the AF could change some topic argument's acceptability status. Such an explanation works via a **change** of AFs, often given some 'universal' space of AF modifications [Wakaki *et al.*, 2009; Booth *et al.*, 2014; Sakama, 2018]. Specifically, addition or removal of arguments and/or (elements of) relations that change the acceptability status of the topic argument is a form of explanation [Fan and Toni, 2015a; Sakama, 2018; Saribatur *et al.*, 2020]: in Figure 1i, $\{c\}$ can explain that removing c from $\langle Args, \mathcal{R}^- \rangle$ would make a non-acceptable (under, say, grounded extension semantics); likewise, $(c, b) \in \mathcal{R}^-$ can explain that removing the attack would make a non-acceptable.

Whichever the form of AF-based explanations, much work is often needed to generate explanations presentable to the (human) users: this is the focus of several works, e.g. via natural language [Wyner and Strass, 2017; Madumal *et al.*, 2019; Zhong *et al.*, 2019; Cyras *et al.*, 2019b; Cocarascu *et al.*, 2019; Raymond *et al.*, 2020; Espinoza *et al.*, 2020] or visualisations [Grando *et al.*, 2013; Rago *et al.*, 2018; Cyras *et al.*, 2019a; Zhong *et al.*, 2019]. A popular format of presentation is in terms of *conversational* explanations [Rago *et al.*, 2018; Cocarascu *et al.*, 2019; Rago *et al.*, 2020; Sklar and Azhar, 2018; Sassoon *et al.*, 2019; Madumal *et al.*, 2019; Kökciyan *et al.*, 2020], often arising naturally from argument structure, DTs, sub-graphs and dialogue games (see Figure 1iii).

5 A Roadmap for Argumentative XAI

Here we identify some gaps in the state-of-the-art on argumentation-based XAI and discuss opportunities for further research, focusing on three avenues: the need to devote more attention to *properties* of AF-based explanations; computational aspects of AF-based explanations; and broadening both applications and the scope of AF-based explanations.

Properties. Properties of AFs have been well studied, e.g. [Dung *et al.*, 2007; Baroni *et al.*, 2019], but properties of AF-based explanations less so. Notable exceptions include forms of *fidelity*, amounting to a sound and complete mapping from the system being explained and the generated AF-based explanations [Cyras *et al.*, 2019b; Fan, 2018], and properties of extension-based *explanation semantics* [Liao and van der Torre, 2020]. Other desirable properties from the broader XAI landscape [Sokol and Flach, 2020] have been mostly neglected, though some user-acceptance aspects such as *cognitive tractability* [Cyras *et al.*, 2019b] as well as transparency and trust [Rago *et al.*, 2020] have been considered for AF-based explanations. For some properties, experiments with human users may be needed, as in much of the XAI literature [Adadi and Berrada, 2018], and creativity in the actual format of AF-based explanations shown to humans required.

Computational aspects. To effectively support XAI solutions, AF-based explanations need to be efficiently computable. In the case of *intrinsic AF-based explanations* this requires efficient systems for the relevant reasoning tasks and a good understanding of the computational complexity thereof. For illustration, the approaches of [Cyras *et al.*, 2019a; Cocarascu *et al.*, 2020] rely upon the tractable membership reasoning task for the grounded extension for AA. In the case of *post-hoc (complete or approximate)* AF-based explanations, a further hurdle is the extraction of AFs from the models in need of explanation, prior to the extraction of the AF-based explanations themselves. For illustration, the (complete) approach of [Cyras *et al.*, 2019b] proves 'soundness and completeness' tractability.

For all types of AF-based explanations, further consideration must be given to the extraction task, of explanations of various formats from AFs. For illustration, the AF-based explanations for the approaches of [Cyras *et al.*, 2019a; Cocarascu *et al.*, 2020] rely upon DTs that can be extracted efficiently from AFs, given the grounded extension. Further, [Saribatur *et al.*, 2020] give complexity results for extracting certain sets of arguments as explanations in AA. In general, however, computational issues in AF-based explanations require a more systematic investigation both in terms of underpinning reasoning tasks and explanation extraction.

Extending applications and the scope of AF-based explanations. While already having a variety of instantiations and covering a wide range of application contexts, AF-based explanations have a wide potential of further development.

Concerning applications, arguably the strongest demand for XAI solutions is currently driven by applications of machine learning (ML). In this context, it is interesting to note that in a loose sense some forms of ML have dialectical roots: supervised ML uses positive and negative examples of concepts to be learnt, and reinforcement learning uses positive and negative rewards. Further, several of the existing XAI solutions for ML, albeit not explicitly argumentative in the sense of this survey, are argumentative in spirit, as discussed in §1 (e.g. SHAP [Lundberg and Lee, 2017] can be seen as identifying reasons for and against outputs). However, AF-based explanations have been only sparingly deployed in ML-driven (classification and probabilistic) settings (see Table 1). We envisage a fruitful interplay, where the explanation needs of ML, while benefiting from the potential of argumentation techniques, also stimulate further research in computational argumentation. Specifically, at the ML end, the analysis of dialectics is a crucial, yet often ignored, underpinning of XAI for ML: it would then be interesting to explore whether an understanding of some existing methods in terms of AF-based explanations could pave the way to new developments. This in turn could lead, on the computational argumentation side, to novel forms of explanation-oriented AFs.

As a first step, it would be interesting to see whether existing approaches on logic-based explanations, either model-agnostic [Ignatiev *et al.*, 2019; Darwiche and Hirth, 2020] or model-specific [Shih *et al.*, 2018], could be understood as AF-based explanations, potentially relying upon existing logic-based AFs such as [Besnard and Hunter, 2001], or ADFs/AFs with structured arguments (see §2). Connections with the widely used *counterfactual explanations (CFs)* (e.g. see [Sokol and Flach, 2020]) represent another stimulating investigation topic. CFs identify, as explanations for models' outputs, hypothetical changes in the inputs that would change these outputs. As such, they show, again, some dialectical flavour and call for the study of forms of AF-based explanations able to provide CF functionalities. For instance, *relation-based CFs* [Albini *et al.*, 2020b] may be interpretable in terms of AF-based explanations for suitable AFs (accommodating different types of support to match the underpinning relations). Given that CFs are based on 'changes', the corresponding form of AF-based explanations discussed in §4 could support this kind of development also.

6 Conclusions

We have given a comprehensive survey of the active research area of argumentative XAI, focusing on explanations built using argumentation frameworks (AFs) from computational argumentation. We have shown how varied AF-based explanations are, in terms of the models they explain, the AFs they deploy, and the forms they take. We have also discussed how AF-based explanations may be defined as integral components of systems that are (argumentatively) explainable by design (we called these explanations 'intrinsic'), but, in several settings, may be provided for existing systems in need of explaining (we called these explanations 'post-hoc') as a result of a marriage between symbolic representations (the AFs) and various forms of (e.g. symbolic or statistical/probabilistic) models. Finally, we have set out a roadmap for future developments of AF-based explanations and their use, which we hope will be beneficial to the AI research community at large, from experts in symbolic AI (and computational argumentation in particular) to application experts in need of customisable, powerful XAI solutions.

References

[Adadi and Berrada, 2018] A. Adadi and M. Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[Albini *et al.*, 2020a] E. Albini, P. Baroni, A. Rago, and F. Toni. Pagerank as an argumentation semantics. In *COMMA*, 2020.

[Albini *et al.*, 2020b] E. Albini, A. Rago, P. Baroni, and F. Toni. Relation-based counterfactual explanations for bayesian network classifiers. In *IJCAI*, 2020.

[Amgoud and Prade, 2009] L. Amgoud and H. Prade. Using arguments for making and explaining decisions. *Artif. Intell.*, 173(3-4):413–436, 2009.

[Antaki and Leudar, 1992] C. Antaki and I. Leudar. Explaining in conversation: Towards an argument model. *Europ. J. of Social Psychology*, 22:181–194, 1992.

[Arioua *et al.*, 2015] A. Arioua, N. Tamani, and M. Croitoru. Query Answering Explanation in Inconsistent Datalog +/- Knowledge Bases. In *DEXA*, 2015.

[Arioua *et al.*, 2017] A. Arioua, P. Buche, and M. Croitoru. Explanatory dialogues with argumentative faculties over inconsistent knowledge bases. *Exp. Syst. Appl.*, 80:244–262, 2017.

[Baroni *et al.*, 2018] P. Baroni, A. Rago, and F. Toni. How many properties do we need for gradual argumentation? In *AAAI*, 2018.

[Baroni *et al.*, 2019] P. Baroni, A. Rago, and F. Toni. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *Int. J. Approx. Reason.*, 105:252–286, 2019.

[Besnard and Hunter, 2001] P. Besnard and A. Hunter. A logic-based theory of deductive arguments. *Artif. Intell.*, 128(1-2):203–235, 2001.

[Bex and Walton, 2016] F. Bex and D. Walton. Combining explanation and argumentation in dialogue. *Arg. & Comput.*, 7(1):55–68, 2016.

[Bondarenko *et al.*, 1997] A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.*, 93:63–101, 1997.

[Booth *et al.*, 2014] R. Booth, D. M. Gabbay, S. Kaci, T. Rienstra, and L. van der Torre. Abduction and Dialogical Proof in Argumentation and Logic Programming. In *ECAI*, 2014.

[Brarda *et al.*, 2019] M. E. Buron Brarda, L. H. Tamargo, and A. J. García. An approach to enhance argument-based multi-criteria decision systems with conditional preferences and explainable answers. *Exp. Syst. Appl.*, 126:171–186, 2019.

[Brewka and Woltran, 2010] G. Brewka and S. Woltran. Abstract dialectical frameworks. In *KR*, 2010.

[Briguez *et al.*, 2014] C. E. Briguez, Maximiliano C. Budán, Cristhian A. D. Deagustini, Ana G. Maguitman, Marcela Capobianco, and Guillermo R. Simari. Argument-based mixed recommenders and their application to movie suggestion. *Exp. Syst. Appl.*, 41(14):6467–6482, 2014.

[Cayrol and Lagasquie-Schiex, 2005] C. Cayrol and M. C. Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *ESQARU*, 2005.

[Chakraborti *et al.*, 2020] T. Chakraborti, S. Sreedharan, and S. Kambhampati. The Emerging Landscape of Explainable Automated Planning & Decision Making. In *IJCAI*, 2020.

[Cocarascu *et al.*, 2018] O. Cocarascu, K. Cyras, A. Rago, and F. Toni. Explaining with Argumentation Frameworks Mined from Data. In *DEXAHAI Workshop*, 2018.

[Cocarascu *et al.*, 2019] O. Cocarascu, A. Rago, and F. Toni. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *AAMAS*, 2019.

[Cocarascu *et al.*, 2020] O. Cocarascu, A. Stylianou, K. Cyras, and F. Toni. Data-empowered argumentation for dialectically explainable predictions. In *ECAI*, 2020.

[Cohen *et al.*, 2014] A. Cohen, S. Gottifredi, A. J. García, and G. R. Simari. A survey of different approaches to support in argumentation systems. *Knowl. Eng. Rev.*, 29(5):513–550, 2014.

[Collins *et al.*, 2019] A. Collins, D. Magazzeni, and S. Parsons. Towards an Argumentation-Based Approach to Explainable Planning. In *XAIP*, 2019.

[Cyras *et al.*, 2019a] K. Cyras, D. Birch, Y. Guo, F. Toni, R. Dulay, S. Turvey, D. Greenberg, and T. Hapuarachchi. Explanations by arbitrated argumentative dispute. *Exp. Syst. Appl.*, 127:141–156, 2019.

[Cyras *et al.*, 2019b] K. Cyras, D. Letsios, R. Misener, and F. Toni. Argumentation for explainable scheduling. In *AAAI*, 2019.

[Darwiche and Hirth, 2020] A. Darwiche and A. Hirth. On the reasons behind decisions. In *ECAI*, 2020.

[Dejl *et al.*, 2021] A. Dejl, P. He, P. Mangal, H. Mohsin, B. Surdu, E. Voinea, E. Albini, P. Lertvittayakumjorn, A. Rago, and F. Toni. Argflow: A toolkit for deep argumentative explanations for neural networks. In *AAMAS*, 2021.

[Dung *et al.*, 2007] P. M. Dung, P. Mancarella, and F. Toni. Computing Ideal Sceptical Argumentation. *Artif. Intell.*, 171(10-15):642–674, 2007.

[Dung, 1995] P. M. Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artif. Intell.*, 77(2):321–358, 1995.

[Espinoza *et al.*, 2020] M. Morveli Espinoza, C. A. Tacla, and H. M. R. Jasinski. An argumentation-based approach for explaining goals selection in intelligent agents. In *BRACIS*, 2020.

[Fan and Toni, 2015a] X. Fan and F. Toni. On Computing Explanations for Non-Acceptable Arguments. In *TAFA*, 2015.

[Fan and Toni, 2015b] X. Fan and F. Toni. On computing explanations in argumentation. In *AAAI*, 2015.

[Fan, 2018] X. Fan. On generating explainable plans with assumption-based argumentation. In *PRIMA*, 2018.

[Fandinno and Schulz, 2019] J. Fandinno and C. Schulz. Answering the "Why" in Answer Set Programming - A Survey of Explanation Approaches. *Theory Pract. Log. Program.*, 19(2):114–203, 2019.

[Gabbay, 2016] D. M. Gabbay. Logical foundations for bipolar and tripolar argumentation networks: preliminary results. *J. Log. Comput.*, 26(1):247–292, 2016.

[Gao *et al.*, 2016] Y. Gao, F. Toni, H. Wang, and F. Xu. Argumentation-based multi-agent decision making with privacy preserved. In *AAMAS*, 2016.

[García *et al.*, 2013] A. J. García, C. I. Chesñevar, N. D. Rotstein, and G. R. Simari. Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. *Exp. Syst. Appl.*, 40(8):3233–3247, 2013.

[Geffner, 2018] H. Geffner. Model-free, Model-based, and General Intelligence. In *IJCAI*, 2018.

[Grando *et al.*, 2013] M. A. Grando, L. Moss, D. H. Sleeman, and J. Kinsella. Argumentation-logic for creating and explaining medical hypotheses. *Artif. Intell. Medicine*, 58(1):1–13, 2013.

[Guidotti *et al.*, 2019] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

[Ignatiev *et al.*, 2019] A. Ignatiev, N. Narodytska, and J. Marques-Silva. On relating explanations and adversarial examples. In *NeurIPS*, 2019.

[Kökciyan *et al.*, 2020] N. Kökciyan, S. Parsons, I. Sassoon, E. Sklar, and S. Modgil. An argumentation-based approach to generate domain-specific explanations. In *EUMAS-AT*, 2020.

[Liao and van der Torre, 2020] B. Liao and L. van der Torre. Explanation semantics for abstract argumentation. In *COMMA*, 2020.

[Lundberg and Lee, 2017] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

[Madumal *et al.*, 2019] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere. A grounded interaction protocol for explainable artificial intelligence. In *AAMAS*, 2019.

[Miller, 2019] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

[Modgil and Caminada, 2009] S. Modgil and M. Caminada. Proof Theories and Algorithms for Abstract Argumentation Frameworks. In *Argumentation in Artificial Intelligence*, chapter 6, pages 105–129. Springer, 2009.

[Modgil and Prakken, 2014] S. Modgil and H. Prakken. The *ASPIC*$^+$ framework for structured argumentation: a tutorial. *Arg. & Comp.*, 5(1):31–62, 2014.

[Moulin *et al.*, 2002] B. Moulin, H. Irandoust, M. Bélanger, and G. Desbordes. Explanation and Argumentation Capabilities: Towards the Creation of More Persuasive Agents. *Artif. Intell. Rev.*, 17(3):169–222, 2002.

[Naveed *et al.*, 2018] S. Naveed, T.Donkers, and J. Ziegler. Argumentation-based explanations in recommender systems: Conceptual framework and empirical results. In *UMAP*, 2018.

[Oren *et al.*, 2020] N. Oren, K. van Deemter, and W. Weber Vasconcelos. Argument-based plan explanation. In *Knowledge Engineering Tools and Techniques for AI Planning*, pages 173–188. Springer, 2020.

[Page *et al.*, 1998] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *WWW: Internet and Web Inf. Syst.*, 54(1999-66):1–17, 1998.

[Rago *et al.*, 2018] A. Rago, O. Cocarascu, and F. Toni. Argumentation-based recommendations: Fantastic explanations and how to find them. In *IJCAI*, 2018.

[Rago *et al.*, 2020] A. Rago, O. Cocarascu, C. Bechlivanidis, and F. Toni. Argumentation as a framework for interactive explanations for recommendations. In *KR*, 2020.

[Raymond *et al.*, 2020] A. Raymond, H. Gunes, and A. Prorok. Culture-Based Explainable Human-Agent Deconfliction. In *AAMAS*, 2020.

[Rodríguez *et al.*, 2017] P. Rodríguez, S. Heras, J. Palanca, J. M. Poveda, N. D. Duque, and V. Julián. An educational recommender system based on argumentation theory. *AI Commun.*, 30(1):19–36, 2017.

[Rolf *et al.*, 2019] L. Rolf, G. Kern-Isberner, and G. Brewka. Argumentation-based explanations for answer sets using ADF. In *LPNMR*, 2019.

[Rosenfeld and Richardson, 2019] A. Rosenfeld and A. Richardson. Explainability in Human-Agent Systems. *Autonomous Agents and Multi-Agent Syst.*, 33:673–705, 2019.

[Sakama, 2018] C. Sakama. Abduction in Argumentation Frameworks. *J. Applied Non-Class. Log.*, 28(2-3):218–239, 2018.

[Saribatur *et al.*, 2020] Z. G. Saribatur, J. P. Wallner, and S. Woltran. Explaining Non-Acceptability in Abstract Argumentation. In *ECAI*, 2020.

[Sassoon *et al.*, 2019] I. Sassoon, N. Kökciyan, E. Sklar, and S. Parsons. Explainable argumentation for wellness consultation. In *EXTRAAMAS*, 2019.

[Schulz and Toni, 2016] C. Schulz and F. Toni. Justifying Answer Sets Using Argumentation. *Theory Pract. Log. Program.*, 16(1):59–110, 2016.

[Sendi *et al.*, 2019] N. Sendi, N. Abchiche-Mimouni, and F. Zehraoui. A new Transparent Ensemble Method based on Deep learning. *Procedia Comput. Sci.*, 159:271–280, 2019.

[Seselja and Straßer, 2013] Dunja Seselja and Christian Straßer. Abstract argumentation and explanation applied to scientific debates. *Synth. Commun.*, 190(12):2195–2217, 2013.

[Shih *et al.*, 2018] A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In *IJCAI*, 2018.

[Sklar and Azhar, 2018] E. I. Sklar and M. Q. Azhar. Explanation through argumentation. In *HAI '18*, 2018.

[Sokol and Flach, 2020] K. Sokol and P. A. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *FAT* '20*, 2020.

[Swartout *et al.*, 1991] W. R. Swartout, C. Paris, and J. D. Moore. Explanations in Knowledge Systems: Design for Explainable Expert Systems. *IEEE Expert*, 6(3):58–64, 1991.

[Timmer *et al.*, 2017] S. T. Timmer, J. C. Meyer, H. Prakken, S. Renooij, and B. Verheij. A two-phase method for extracting explanatory arguments from bayesian networks. *Int. J. Approx. Reason.*, 80:475–494, 2017.

[Wakaki *et al.*, 2009] T. Wakaki, K. Nitta, and H. Sawamura. Computing Abductive Argumentation in Answer Set Programming. In *ArgMAS*, 2009.

[Walton, 2004] D. Walton. A New Dialectical Theory of Explanation. *Philos. Explor.*, 7(1):71–89, 2004.

[Wyner and Strass, 2017] A. Z. Wyner and H. Strass. dARe - using argumentation to explain conclusions from a controlled natural language knowledge base. In *IEA/AIE*, 2017.

[Zeng *et al.*, 2018] Z. Zeng, X. Fan, C. Miao, C. Leung, J. J. Chin, and Y. S. Ong. Context-based and explainable decision making with argumentation. In *AAMAS*, 2018.

[Zeng *et al.*, 2019] Z. Zeng, C. Miao, C. Leung, Z. Shen, and J. J. Chin. Computing argumentative explanations in bipolar argumentation frameworks. In *AAAI*, 2019.

[Zhong *et al.*, 2019] Q. Zhong, X. Fan, X. Luo, and F. Toni. An explainable multi-attribute decision model based on argumentation. *Exp. Syst. Appl.*, 117:42–61, 2019.