Argumentative XAI: A Survey

Kristijonas Čyras\(^1\)\(^\ast\), 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 with which argumentation-based explanations are deployed, 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 great amount of attention in recent years, due mostly to its role in bridging 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 has been 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 [Swarfout 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 modelspecific [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).

We ignore argumentative explanations based on informal notions or models lacking aspects of AFs (notably, semantics) and application domains of argumentative XAI, covered in a recent, orthogonal survey [Vassiliades et al., 2021].
Figure 1: (i) An AA framework \(\{\text{Args}, \mathcal{R}\}^\sim\) 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 \(\text{Args}\)), starting from Abstract Argumentation (AA) [Dung, 1995]. Here, a single dialectical relation of attack (\(\mathcal{R}^\sim \subseteq \text{Args} \times \text{Args}\)) is considered and semantics are given in terms of mappings from AA frameworks \(\{\text{Args}, \mathcal{R}\}^\sim\) to sets of so-called extensions, each a set of arguments \(\mathcal{E} \subseteq \text{Args}\) fulfilling some specified, dialectically meaningful constraints (e.g. conflict-freeness, amounting to absence of internal conflicts: \(\{\alpha, \beta\} \in \mathcal{E} \Rightarrow (\alpha, \beta) \in \mathcal{R}^\sim\) = \(\varnothing\)). 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) [Gayrol and Lages, 2005] an additional dialectical relation of support (\(\mathcal{R}^+ \subseteq \text{Args} \times \text{Args}\)) is considered, and a variety of semantics in terms of extensions are 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 [Bordonaro 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; Branda 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 models that are natively using argumentative techniques, covered in §3.1;
- **post-hoc**, i.e. obtained from non-argumentative models; 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 from which the explanations are obtained. We exemplify each of our chosen categories of models with selected 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.

[Cýras et al., 2019a] propose a case-based reasoning (CBR)-inspired but AA-driven classification method for 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 [Cýras et al., 2019a] cases amount to categorical data, in [Cocarascu et al., 2020a] 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. [Kökcıyan 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 sound and complete fashion onto argument acceptance in AAB 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., 2020] use TA frameworks extracted from a hybrid RS as the basis for explanations. This RS exploits collaborative 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 [Deij 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 discussed examples of three types of AF-based explanation. 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.
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 acceptability 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, given a property $P$ of graphs and an AA framework $G = \langle \text{Args}, R \rangle$, an explanation for the acceptability status of some topic argument $a \in \text{Args}$ can be defined as a sub-graph $G' = \langle \text{Args}', R' \rangle$ of $G$ such that $a \in \text{Args}'$ and $G'$ satisfies $P$ [Cocarascu et al., 2018].

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; Čyras et al., 2019a], such as existence, correctness (the explanation actually establishes the acceptability 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; Čyras 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]. Similarly, 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; Kőkciyan 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 1ii) 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 extensions (see §2), keeping the relationships implicit [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 (conflict-free, defending against all attackers and in where one (topic) argument is defended by all other arguments) in AA frameworks. For illustration, in Figure 1i, related admissible extensions $\{a, c\}$ and $\{a, c, c\}$ are explanations of (the acceptability of) the topic argument $a$.

Another form of AF-based explanation amounts to indicating a change of the AF that would change some topic argument’s acceptability status, 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 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 $\text{Args}, R^{-}$ would make $a$ non-acceptable (under, say, grounded extension semantics); likewise, $(c, b) \in R^{-}$ 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; Čyras 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; Čyras et al., 2019a; Zhong et al., 2019]. Popular are conversational explanations [Cocarascu et al., 2019; Rago et al., 2020; Sklar and Azhar, 2018; Sassoon et al., 2019; Madumal et al., 2019; Kőkciyan et al., 2020], often arising naturally from argument structure, 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.

---

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

<table>
<thead>
<tr>
<th>Form of Explanation (AF)</th>
<th>AF-Based Agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue (AA, AA-ALP)</td>
<td>[Wakaki et al., 2009]</td>
</tr>
<tr>
<td>Dialogue (AA)</td>
<td>[Modgil and Caminada, 2009]</td>
</tr>
<tr>
<td>Sub-graph (variant of AA)</td>
<td>[Seselja and Straßer, 2013]</td>
</tr>
<tr>
<td>Sub-graph (ASPIC+)</td>
<td>[Grando et al., 2013]</td>
</tr>
<tr>
<td>Sub-graph (DeLP)</td>
<td>[Booth et al., 2014]</td>
</tr>
<tr>
<td>Change, Dialogue (AA, AA-ALP)</td>
<td>[Fan and Toni, 2015b]</td>
</tr>
<tr>
<td>Extensions, DTs (AA, ABA)</td>
<td>[Fan and Toni, 2015a]</td>
</tr>
<tr>
<td>Change (AA)</td>
<td>[Čyras et al., 2019a]</td>
</tr>
<tr>
<td>Dialogue (AA-LD)</td>
<td>[Arioua et al., 2017]</td>
</tr>
<tr>
<td>Change (AA)</td>
<td>[Sakama, 2018]</td>
</tr>
<tr>
<td>DTs (AA-CBR)</td>
<td>[Zeng et al., 2019]</td>
</tr>
<tr>
<td>Change (AA)</td>
<td>[Saribatur et al., 2020]</td>
</tr>
<tr>
<td>Extensions (AA)</td>
<td>[Liao and van der Torre, 2020]</td>
</tr>
</tbody>
</table>

| [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) |
| [Kőkciyan et al., 2020]  | Structure (AA-LD, ASPIC+) |
| [Espinoza et al., 2020]  | Sub-graph, Extensions (AA-LD) |
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 [Čyras 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 [Čyras 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 [Čyras 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 [Čyras 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 [Čyras 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 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.

Acknowledgements
This research was funded in part by the Royal Academy of Engineering, UK, and by J.P. Morgan.


