A Little of That Human Touch: Achieving Human-Centric Explainable AI via Argumentation[*]

Antonio Rago
Department of Computing, Imperial College London
a.rago@imperial.ac.uk

Abstract

As data-driven AI models achieve unprecedented feats across previously unthinkable tasks, the diminishing levels of interpretability of their increasingly complex architectures can often be sidelined in place of performance. If we are to comprehend and trust these AI models as they advance, it is clear that symbolic methods, given their unparalleled strengths in knowledge representation and reasoning, can play an important role in explaining AI models. In this paper, I discuss some of the ways in which one branch of such methods, computational argumentation, given its human-like nature, can be used to tackle this problem. I first outline a general paradigm for this area of *explainable AI*, before detailing a prominent methodology therein which we have pioneered. I then illustrate how this approach has been put into practice with diverse AI models and types of explanations, before looking ahead to challenges, future work and the outlook in this field.

1 Introduction

Explainable AI (XAI) is a field of research dedicated to methods for explaining the outputs of AI models, which may be deployed in everything from everyday tasks, e.g. explaining a movie recommendation [Rago *et al.*, 2018b], to endeavours on which millions of lives depend, e.g. using XAI for drug discovery [Wong *et al.*, 2024]. These methods are usually designed to target metrics, which can be roughly categorised as either *machine-centric* or *human-centric*. The former set of metrics, most often evaluated empirically using datasets, concern the AI model only, e.g. *faithfulness* [Jacovi and Goldberg, 2020], i.e. how closely the explanations align with the AI model. Meanwhile, the latter set of metrics, e.g. whether users *comprehend* or *trust* the AI model and/or the explanation, are more elusive and often subjective. Indeed, how explainability facilitates trust in users is far from trivial [Ferrario and Loi, 2022]. Consequently, user studies with XAI methods are lacking in the literature [Keane *et al.*, 2021].

Concurrently, a prominent trend of late has been the use of techniques from symbolic AI to explain the outputs of data-driven models which lack interpretability [Ferreira *et al.*, 2022], notably targeting trust [Marques-Silva and Ignatiev, 2022]. Symbolic methods are arguably second-to-none in representing and reasoning with the knowledge behind a decision, giving a variety of tools to tackle this problem. One such research area is that of *computational argumentation*, as introduced in the seminal [Dung, 1995], a branch of logic which excels in uncertainty management and conflict resolution. This has led to its successful application in diverse domains from law [Sartor *et al.*, 2022] to medicine [Sassoon *et al.*, 2021]. Another of argumentation's strengths is its human-like nature: it has been argued that all human reasoning [Mercier and Sperber, 2011] and the majority of statements in explanation [Antaki and Leudar, 1992] are argumentative. This affords great potential for producing explanations which perform not only in machine-centric metrics, but also in the more elusive human-centric metrics.

In this paper, I first introduce a family of argumentation formalisms which have proved popular in explaining AI models, examining their particular intricacies which render them suitable for this task (§2). Next, I outline one methodology for extracting argumentative representations which harbour the relevant explanatory information from AI models (§3), before covering a set of instantiations of this methodology (§4). I then showcase some of the explanations generated by this methodology, considering their format and interactivity (§5). Finally, I discuss existing challenges, future work and the outlook for this fruitful avenue of research (§6).

2 Gradual Argumentation for Explanation

Argumentation (see [Atkinson *et al.*, 2017] for an overview) has long been known to excel in representing knowledge and resolving conflicts therein. *Abstract argumentation frameworks* (AFs) [Dung, 1995] represent *arguments* as abstract entities in a graph with a relation of *attack* showing which arguments are in conflict. However, it has been suggested that a number of applications, notably those where human cognition is concerned [Benferhat *et al.*, 2002], call for an additional relation that is diametrically opposed to attack. This motivated the introduction of a *support* relation in *bipolar argumentation frameworks* (BAFs) [Amgoud *et al.*, 2008] and has since been verified in user studies as aligning with human

---

[*] Adapted from the Bruce Springsteen single "Human Touch".

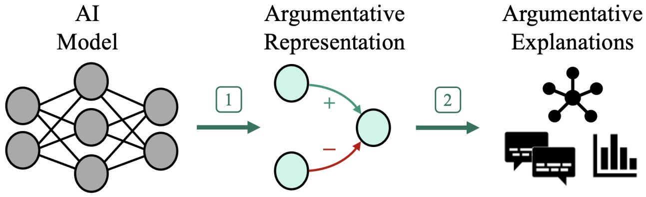


Figure 1: A general paradigm for extracting argumentative explanations from AI models: Step 1 concerns forging an argumentative representation from the AI model (covered in §3 and §4) and Step 2 concerns generating argumentative explanations (covered in §5).

reasoning and, notably, being distinct from the notion of defence (attacking an argument's attacker) [Polberg and Hunter, 2018].

Gradual semantics have also been introduced to incorporate uncertainty into the evaluation of arguments in both AFs [Besnard and Hunter, 2001] and BAFs [Amgoud *et al.*, 2008]. Further, intrinsic strengths, i.e. a quantitative evaluation of arguments before the effects of other arguments are considered, allow for additional information, e.g. social media votes in AFs [Leite and Martins, 2011], to be embodied in arguments. BAFs with an intrinsic strength, or *quantitative BAFs* (QBAFs) [Baroni *et al.*, 2019], are applicable in a number of settings, e.g. law, social media, engineering and e-democracy [Rago *et al.*, 2018a]. This is complemented by the fact that there is a rich variety of gradual semantics for QBAFs in the literature, e.g. [Rago *et al.*, 2016; Potyka, 2018; Amgoud and Ben-Naim, 2018], offering different behaviours for different contexts. Indeed, there is a direct mapping of a particular semantics to *multi-layer perceptrons* (MLPs) [Potyka, 2021]. These semantics' behaviours are typically characterised by theoretical properties (see [Amgoud and Ben-Naim, 2018; Baroni *et al.*, 2019] for overviews), many of which are intuitive from an explanatory viewpoint. It is for these reasons that I posit that QBAFs are particularly amenable to abstracting away the explanatory knowledge to be delivered to humans, with their theoretical properties forming the building blocks of explanations, as we will see in §3.

This family of argumentation formlisms is just one of a whole host which have been used to explain AI models (see [Cyras *et al.*, 2021; Vassiliades *et al.*, 2021; Guo *et al.*, 2023] for recent overviews). However, the nexus of this paper will be on the use of (the various derivatives of) AFs, under gradual semantics, to explain the outputs of AI models. In particular, I will focus on those that follow the general paradigm shown in Figure 1. Here, an argumentative representation, e.g. a QBAF, is first extracted from an AI model, harbouring the explanatory information used to generate various forms of (thus argumentative) explanation to be delivered to users. Such a modular approach allows for uniform explanations to be created across different AI models and settings, which may bring benefits from a regulatory viewpoint, or simply for delivering consistent explanations to humans that they are comfortable with.

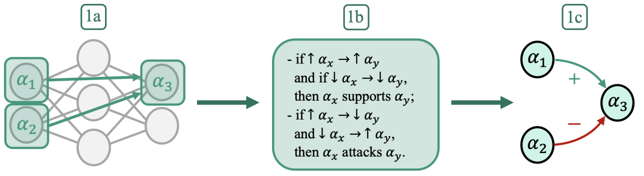


Figure 2: The methodology for forging argumentative representations from AI models [Rago *et al.*, 2022], where Step 1a consists of identifying potential arguments and potential relations in an AI model; Step 1b concerns the definition of an explanation mould, i.e. a set of relation characterisations obtained by inverting properties of gradual semantics; and Step 1c comprises the generation of arguments and argumentative relations by checking potential relations' satisfaction of the relation characterisations.

3 Forging Argumentative Representations

A methodology which we have pioneered, and forms the basis of the core ideas through a number of the approaches discussed in §4, is the *forging* of argumentative representations from AI models [Rago *et al.*, 2022]. This process begins with the identification of potential arguments and relations in an AI model, limiting the explanation to those which are relevant, as is formalised as a principle for explanations in [Cyras *et al.*, 2022]. We then determine which *potential arguments* and *potential relations* could be instantiated as part of the argumentation representation harbouring the explanatory information, thus ensuring the explanation is selective, as recommended by [Miller, 2019]. To do so, we reverse the usual process of gradual evaluation of arguments based on their relations with other arguments. Instead, we interpret some quantitative evaluation of the components in the AI model which are potential arguments, e.g. an activation value in a neural network. We then define an *explanation mould*, i.e. a set of *relation characterisations* such that if two potential arguments with a potential relation between them satisfy such constraints, we categorise the potential relation as an argumentative relation, e.g. an attack or support. These relation characterisations may thus be obtained by reinterpreting properties of gradual semantics. Given an input, we can then forge an argumentative representation which explains the AI model, based on predefined behaviours tailored to the user and setting. This process is illustrated in Figure 2.

In [Rago *et al.*, 2022], we deployed this methodology to explain *structural causal models* [Pearl, 1999], using an explanation mould based on the reinterpretation of the property of *bi-variate reinforcement* [Amgoud and Ben-Naim, 2018]. Our theoretical analysis demonstrated advantages of the resulting argumentative representation from both explanatory and argumentative viewpoints. Then, in [Rago *et al.*, 2023b], we forged argumentative representations to explain classifiers in general, demonstrating empirically its advantages over *SHAP* [Lundberg and Lee, 2017] in certain conditions. Both explanations here were *input-output*, i.e. those where only the inputs and outputs of the AI model are required. However, *mechanistic* explanations, which consider the AI model's internal functionality, are not beyond the potential of the methodology, as we will see in §4.

4 Argumentative Representations of AI Models

The general methodology of forging argumentative explanations has been deployed for different types of AI models in various settings. Neural networks of various types and architectures have also been shown to be amenable to argumentative explanations via the forging process. We pioneered this approach in [Dejl *et al.*, 2021], selecting (single or groups of) neurons to be potential arguments in the forging process, allowing for different architectures, e.g. MLPs or *convolutional neural networks* (CNNs), in various tasks, e.g. text or image classification. This work was extended in [Sukpanichnant *et al.*, 2021], where we interpreted an existing explanation method as a gradual semantics for the QBAFs representing neural networks. Meanwhile, the authors of [Ayoobi *et al.*, 2023] took a different approach, including the step of sparsifying an MLP before translating it to a gradual argumentation framework, which outperformed the baselines in faithfulness.

Another fruitful domain for this methodology is in recommender systems. We defined an explainable recommender system with a purpose-built graphical structure, from which argumentation frameworks with three relations (attacks, supports and neutralisers) can be extracted, consisting of the reasoning for the recommendation [Rago *et al.*, 2018b]. A similar process is used in [Cocarascu *et al.*, 2019], in which we deploy a variation of the forging methodology for aggregating movie reviews. Here, arguments are generated from entities in an ontology with the *part of* relation, instances of which serve as potential relations since more specific entities can be seen as attacking or supporting more general entities that they are part of. We use the intrinsic strength in a QBAF to represent the sentiment from reviews on each argument, extracted via NLP techniques, which are then used to determine the argumentative relations, as illustrated in Figure 3. The method gives an automated way of extracting argumentative explanations which are faithful to the review aggregation.

Other types of model have been explained with roughly the same methodology, e.g. Bayesian [Albini *et al.*, 2023] and tree-based [Potyka *et al.*, 2023] classifiers, with notable benefits in the explanations' faithfulness.

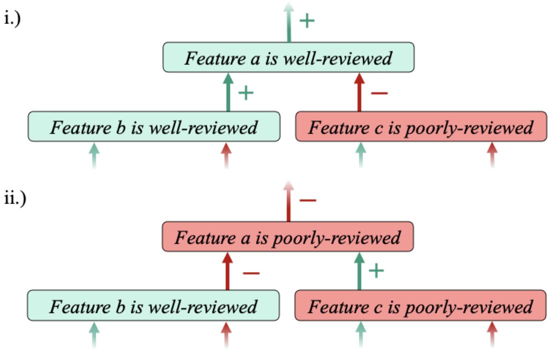


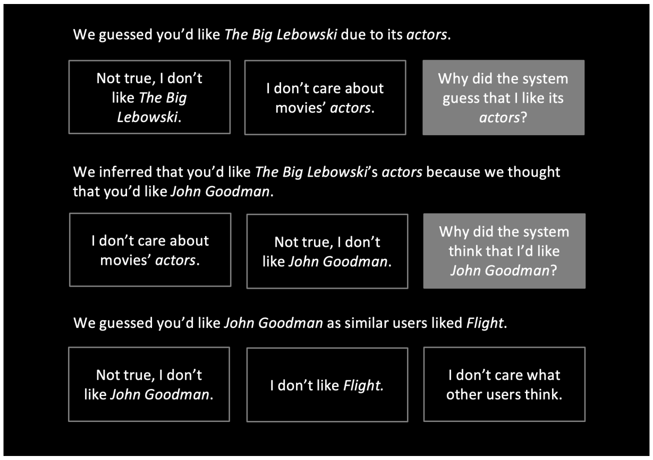
Figure 3: In the review aggregation setting, instances of the *part of* relation may be considered potential relations, e.g. features b and c are part of feature a here. Relations of attack and support are then characterised based the general sentiment of reviews on the feature, e.g. in case i here, since feature a is well-reviewed, the well-reviewed feature b supports this argument, while the poorly-reviewed feature c attacks it. However, if the sentiment on feature a is generally negative, the incoming and outgoing relations are inverted, as in case ii.

5 Generating Argumentative Explanations

In §4 we saw how argumentative representations may be extracted from AI models, but such argumentation frameworks alone are not sufficient explanations for humans, as demonstrated by the works on explaining argumentation frameworks themselves, e.g. [Borg and Bex, 2021]. These argumentation representations instead provide the means for generating explanations in a variety of different forms. Indeed, it has been shown that it is not only the content of an explanation which is crucial for performing in human-centric metrics such as comprehensibility and trust, but also the explanation's format [Bertrand *et al.*, 2023]. Also important is an explanation's level of interactivity, a capability which is in line with the movement towards human-like, social explanations, as advocated by [Miller, 2019].

With the content of an explanation having been determined by the forging process, we now consider the effect of format. We demonstrated argumentation's capability for generating explanations of different formats in recommender systems, but also for supporting interactivity via human feedback, ensuring that this additional information affects the recommendations in an intuitive manner [Rago *et al.*, 2020; Rago *et al.*, 2021]. In [Rago *et al.*, 2021], we undertook user studies which examined the effect of three different formats of argumentative explanation, namely textual, tabular and conversational, for recommendations in the movie domain. We found that users' comprehensibility of and trust in the recommender system improved after receiving explanations of any of the formats. Further, we found that humans did indeed have preferences over the format of explanation, but also that these preferences were diverse, a finding which corroborated that in [Rago *et al.*, 2020]. Given that this is the case for something as seemingly innocuous as a movie recommendation, I would posit that this effect may be even more pronounced in high-stakes settings. This highlights the need for our methodology, supporting a range of explanations via a modular approach.

Argumentation itself has long been known to be an effective means for supporting dialogues, e.g. in *persuasion* [Hunter, 2018] or *inquiry* [Black and Hunter, 2007]. A recent contribution of ours in this area was a general framework for interactivity in explainable AI (and beyond) [Rago *et al.*, 2023a]. Here, we frame the process of explanation between agents, e.g. a machine and a human, as a conflict resolution problem, using QBAFs once again. Given the fact that a number of AI models can be represented as (possibly restricted forms of) QBAFs, as we have seen in §4, we provide the groundwork for a comprehensive line of research into whether they can also represent humans' reasoning processes, and thus machine-human interactions. We also demonstrate the potential for representing cognitive biases, which are known to be a crucial component in XAI [Bertrand

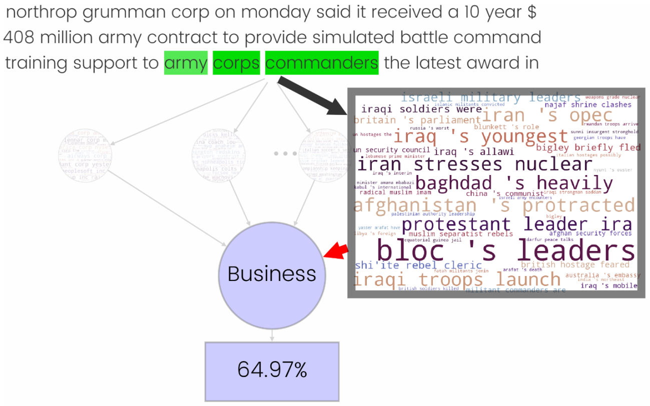


We guessed you'd like *The Big Lebowski* due to its *actors*.

We inferred that you'd like *The Big Lebowski's actors* because we thought that you'd like *John Goodman*.

We guessed you'd like *John Goodman* as similar users liked *Flight*.

Figure 4: An argumentative explanation in a conversational format for the recommender systems from [Rago *et al.*, 2020], demonstrating how user feedback can be incorporated via interactivity. As options are selected by the user (shown in grey), the recommender system steps through the argumentative reasoning to formulate its response, where any user feedback is guaranteed to affect the recommendations intuitively.



northrop grumman corp on monday said it received a 10 year $ 408 million army contract to provide simulated battle command training support to army corps commanders the latest award in

Figure 5: An argumentative explanation in a graphical format for the text classifier in [Dejl *et al.*, 2021], demonstrating how interactivity can be used to ensure explanations are selective, relevant and cognitively manageable. Here, the words in the input highlighted in green support the argument representing the CNN's filter, which has been selected by the user and roughly represents militaristic terms, that attacks the classification *Business*.

et al., 2022]. Finally, we show that greedily selecting the strongest, most relevant reasoning is not always the most effective in explanation, compared with taking an argumentative strategy in the selection.

Interactivity can also be simply providing the means for users to select the content of information shown in an explanation, e.g. such that it is cognitively manageable or tailored to that individual user's preferences. We demonstrated argumentation's potential for this in [Dejl *et al.*, 2021; Sukpanichnant *et al.*, 2021], where arguments in explanations showing the words or pixels which activated a filter in a CNN are highlighted when selected, as demonstrated in Figure 5. In [Sukpanichnant *et al.*, 2021] we also provided a comparison with the method of [Olah *et al.*, 2018] within image classification, noting that, in our approach, the positions of the extracted arguments in the image were less influential.

6 Challenges, Future Work and Outlook

We have seen that gradual argumentation provides a comprehensive repertoire for explaining the outputs of data-driven AI models. I introduced the methodology of forging an argumentative representation based on the explanatory content required for an AI model, task and application, before generating different forms of argumentative explanation depending on users' explanatory requirements. I then demonstrated the scope of this methodology across a subset of the domains in which it has been deployed. However, many challenges remain in this field, each bringing opportunities for future work.

One of the most prominent challenges is the widespread deployment of argumentative systems in real-world applications. As XAI proliferates, it is imperative that this wave of interest is ridden by argumentative XAI researchers, especially in applications such as engineering, which is fertile but somewhat untouched ground for AI models and their explanations. For this to take place, argumentative XAI methods need to prove they can navigate the bottlenecks of knowledge acquisition and scale, and their professed advantages in aligning with human behaviour and achieving user-centric metrics need to be validated in user studies, as in [Vesic *et al.*, 2022]. This is especially the case for interactivity: as AI models become ever more intertwined with our daily lives, this offers an excellent opportunity for argumentative XAI to take advantage of its strengths in representing these interactions, via frameworks such as that we introduced [Rago *et al.*, 2023a].

The representation of state-of-the-art AI models in argumentative forms is also an outstanding task. For example, most of the argumentative representations of neural networks in the literature concern either MLPs or CNNs. Such an abstraction for the *transformer* architecture [Vaswani *et al.*, 2017] driving *large language models* (LLMs) [Min *et al.*, 2024] is lacking, as far as I am aware. A simpler route may be to supplement the LLM with an explainable, argumentative wrapper, given the recent benefits which have been seen in training LLMs with argumentative knowledge [Furman *et al.*, 2023]. A separate line of research which looks fruitful is the use of argumentation to explain tree-based classifiers, e.g. as in [Potyka *et al.*, 2023], particularly given their widespread use in financial domains.

Other research directions which present opportunities for argumentative explanations include the proposed realignment of the XAI paradigm towards *evaluative AI* [Miller, 2023], wherein humans are given multiple options from an AI model, rather than a single output, with positive and negative reasoning for each. BAFs or QBAFs, with their attack and support relations, may prove to be effective here. Similarly, argumentation seems a natural fit for *contestable AI*, which is required by law in some jurisdictions, e.g. GDPR, article 22(3)[1], states that a data subject shall have "at least the right [. . .] to contest the decision".

---

[1] https://gdpr-text.com/read/article-22/.

Acknowledgements

The author was partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme, and by ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934). Any views or opinions expressed herein are solely those of the author.

References

[Albini *et al.*, 2023] Emanuele Albini, Antonio Rago, Pietro Baroni, and Francesca Toni. Achieving descriptive accuracy in explanations via argumentation: The case of probabilistic classifiers. *Frontiers Artif. Intell.*, 6, 2023.

[Amgoud and Ben-Naim, 2018] Leila Amgoud and Jonathan Ben-Naim. Evaluation of arguments in weighted bipolar graphs. *Int. J. Approx. Reason.*, 99:39–55, 2018.

[Amgoud *et al.*, 2008] Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex, and P. Livet. On bipolarity in argumentation frameworks. *Int. J. Intell. Syst.*, 23(10):1062–1093, 2008.

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

[Atkinson *et al.*, 2017] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Magazine*, 38(3):25–36, 2017.

[Ayoobi *et al.*, 2023] Hamed Ayoobi, Nico Potyka, and Francesca Toni. Sparx: Sparse argumentative explanations for neural networks. In *ECAI*, pages 149–156, 2023.

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

[Benferhat *et al.*, 2002] Salem Benferhat, Didier Dubois, Souhila Kaci, and Henri Prade. Bipolar representation and fusion of preferences on the possibilistic logic framework. In *KR*, pages 421–448, 2002.

[Bertrand *et al.*, 2022] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *AIES*, pages 78–91, 2022.

[Bertrand *et al.*, 2023] Astrid Bertrand, James R. Eagan, and Winston Maxwell. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *FAccT*, pages 943–958, 2023.

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

[Black and Hunter, 2007] Elizabeth Black and Anthony Hunter. A generative inquiry dialogue system. In *AAMAS*, page 241, 2007.

[Borg and Bex, 2021] AnneMarie Borg and Floris Bex. A basic framework for explanations in argumentation. *IEEE Intell. Syst.*, 36(2):25–35, 2021.

[Cocarascu *et al.*, 2019] Oana Cocarascu, Antonio Rago, and Francesca Toni. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *AAMAS*, pages 1261–1269, 2019.

[Cyras *et al.*, 2021] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. In *IJCAI*, pages 4392–4399, 2021.

[Cyras *et al.*, 2022] Kristijonas Cyras, Timotheus Kampik, and Qingtao Weng. Dispute trees as explanations in quantitative (bipolar) argumentation. In *ArgXAI*, 2022.

[Dejl *et al.*, 2021] Adam Dejl, Chloe He, Pranav Mangal, Hasan Mohsin, Bogdan Surdu, Eduard Voinea, Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, and Francesca Toni. Argflow: A toolkit for deep argumentative explanations for neural networks. In *AAMAS*, pages 1761–1763, 2021.

[Dung, 1995] Phan Minh 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.

[Ferrario and Loi, 2022] Andrea Ferrario and Michele Loi. How explainability contributes to trust in AI. In *FAccT*, pages 1457–1466, 2022.

[Ferreira *et al.*, 2022] João Ferreira, Manuel de Sousa Ribeiro, Ricardo Gonçalves, and João Leite. Looking inside the black-box: Logic-based explanations for neural networks. In *KR*, 2022.

[Furman *et al.*, 2023] Damián Ariel Furman, Pablo Torres, José A. Rodríguez, Diego Letzen, Maria Vanina Martinez, and Laura Alonso Alemany. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *EMNLP*, pages 2942–2956, 2023.

[Guo *et al.*, 2023] Yihang Guo, Tianyuan Yu, Liang Bai, Jun Tang, Yirun Ruan, and Yun Zhou. Argumentative explanation for deep learning: A survey. In *ICUS*, pages 1738–1743, 2023.

[Hunter, 2018] Anthony Hunter. Towards a framework for computational persuasion with applications in behaviour change. *Argument Comput.*, 9(1):15–40, 2018.

[Jacovi and Goldberg, 2020] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *ACL*, pages 4198–4205, 2020.

[Keane *et al.*, 2021] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In *IJCAI*, pages 4466–4474, 2021.

[Leite and Martins, 2011] João Leite and João G. Martins. Social abstract argumentation. In *IJCAI*, pages 2287–2292, 2011.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.

[Marques-Silva and Ignatiev, 2022] João Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through formal XAI. In *AAAI*, pages 12342–12350, 2022.

[Mercier and Sperber, 2011] Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.

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

[Miller, 2023] Tim Miller. Explainable AI is dead, long live explainable AI!: Hypothesis-driven decision support using evaluative AI. In *FAccT*, pages 333–342, 2023.

[Min *et al.*, 2024] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2):30:1–30:40, 2024.

[Olah *et al.*, 2018] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018.

[Pearl, 1999] Judea Pearl. Reasoning with cause and effect. In *IJCAI*, pages 1437–1449, 1999.

[Polberg and Hunter, 2018] Sylwia Polberg and Anthony Hunter. Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *Int. J. Approx. Reason.*, 93:487–543, 2018.

[Potyka *et al.*, 2023] Nico Potyka, Xiang Yin, and Francesca Toni. Explaining random forests using bipolar argumentation and markov networks. In *AAAI*, pages 9453–9460, 2023.

[Potyka, 2018] Nico Potyka. Continuous dynamical systems for weighted bipolar argumentation. In *KR*, pages 148–157, 2018.

[Potyka, 2021] Nico Potyka. Interpreting neural networks as quantitative argumentation frameworks. In *AAAI*, pages 6463–6470, 2021.

[Rago *et al.*, 2016] Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. Discontinuity-free decision support with quantitative argumentation debates. In *KR*, pages 63–73, 2016.

[Rago *et al.*, 2018a] Antonio Rago, Pietro Baroni, and Francesca Toni. On instantiating generalised properties of gradual argumentation frameworks. In *SUM*, pages 243–259, 2018.

[Rago *et al.*, 2018b] Antonio Rago, Oana Cocarascu, and Francesca Toni. Argumentation-based recommendations: Fantastic explanations and how to find them. In *IJCAI*, pages 1949–1955, 2018.

[Rago *et al.*, 2020] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, and Francesca Toni. Argumentation as a framework for interactive explanations for recommendations. In *KR*, pages 805–815, 2020.

[Rago *et al.*, 2021] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David A. Lagnado, and Francesca Toni. Argumentative explanations for interactive recommendations. *Artif. Intell.*, 296:103506, 2021.

[Rago *et al.*, 2022] Antonio Rago, Pietro Baroni, and Francesca Toni. Explaining causal models with argumentation: the case of bi-variate reinforcement. In *KR*, 2022.

[Rago *et al.*, 2023a] Antonio Rago, Hengzhi Li, and Francesca Toni. Interactive explanations by conflict resolution via argumentative exchanges. In *KR*, pages 582–592, 2023.

[Rago *et al.*, 2023b] Antonio Rago, Fabrizio Russo, Emanuele Albini, Francesca Toni, and Pietro Baroni. Explaining classifiers' outputs with causal models and argumentation. *FLAP*, 10(3):421–509, 2023.

[Sartor *et al.*, 2022] Giovanni Sartor, Michal Araszkiewicz, Katie Atkinson, Floris Bex, Tom M. van Engers, Enrico Francesconi, Henry Prakken, Giovanni Sileno, Frank Schilder, Adam Wyner, and Trevor J. M. Bench-Capon. Thirty years of artificial intelligence and law: the second decade. *Artif. Intell. Law*, 30(4):521–557, 2022.

[Sassoon *et al.*, 2021] Isabel Sassoon, Nadin Kökciyan, Sanjay Modgil, and Simon Parsons. Argumentation schemes for clinical decision support. *Argument Comput.*, 12(3):329–355, 2021.

[Sukpanichnant *et al.*, 2021] Purin Sukpanichnant, Antonio Rago, Piyawat Lertvittayakumjorn, and Francesca Toni. Neural qbafs: Explaining neural networks under lrp-based argumentation frameworks. In *AIxIA*, pages 429–444, 2021.

[Vassiliades *et al.*, 2021] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence: a survey. *Knowl. Eng. Rev.*, 36:e5, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Vesic *et al.*, 2022] Srdjan Vesic, Bruno Yun, and Predrag Teovanovic. Graphical representation enhances human compliance with principles for graded argumentation semantics. In *AAMAS*, pages 1319–1327, 2022.

[Wong *et al.*, 2024] Felix Wong, Erica J. Zheng, Jacqueline A. Valeri, Nina M. Donghia, Melis N. Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, Abigail L. Manson, Jens Friedrichs, Ralf Helbig, Behnoush Hajian, Dawid K. Fiejtek, Florence F. Wagner, Holly H. Soutter, Ashlee M. Earl, Jonathan M. Stokes, Lars D. Renner, and James J. Collins. Discovery of a structural class of antibiotics with explainable deep learning. *Nat.*, 626(7997):177–185, 2024.