Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning
Main Track

Interactive Explanations by Conflict Resolution via Argumentative Exchanges

Antonio Rago , **Hengzhi Li** and **Francesca Toni**
Department of Computing, Imperial College London, UK
{a.rago, hengzhi.li21, ft}@imperial.ac.uk

Abstract

As the field of explainable AI (XAI) is maturing, calls for interactive explanations for (the outputs of) AI models are growing, but the state-of-the-art predominantly focuses on *static explanations*. In this paper, we focus instead on interactive explanations framed as *conflict resolution* between agents (i.e. AI models and/or humans) by leveraging on computational argumentation. Specifically, we define *Argumentative eXchanges (AXs)* for dynamically sharing, in multi-agent systems, information harboured in individual agents' *quantitative bipolar argumentation frameworks* towards resolving conflicts amongst the agents. We then deploy AXs in the XAI setting in which a machine and a human interact about the machine's predictions. We identify and assess several theoretical properties characterising AXs that are suitable for XAI. Finally, we instantiate AXs for XAI by defining various agent behaviours, e.g. capturing counterfactual patterns of reasoning in machines and highlighting the effects of cognitive biases in humans. We show experimentally (in a simulated environment) the comparative advantages of these behaviours in terms of conflict resolution, and show that the strongest argument may not always be the most effective.

1 Introduction

The need for interactivity in explanations of the outputs of AI models has long been called for (Cawsey 1991), and the recent wave of explainable AI (XAI) has given rise to renewed urgency in the matter. In (Miller 2019), it is stated that explanations need to be social, and thus for machines to truly explain themselves, they must be interactive, so that XAI is not just "more AI", but a human-machine interaction problem. Some have started exploring explanations as dialogues (Lakkaraju et al. 2022) , while several are exploring forms of interactive machine learning for model debugging (Teso et al. 2023). It has also been claimed that it is our responsibility to create machines which can argue with humans (Hirsch et al. 2018). However, despite the widespread acknowledgement of the need for interactivity, typical approaches to XAI deliver *"static" explanations*, whether they be based on feature attribution (e.g. as in (Lundberg and Lee 2017)), counterfactuals (e.g. as in (Wachter, Mittelstadt, and Russell 2017)) or other factors such as prime implicants (e.g. as in (Shih, Choi, and Darwiche 2018; Ignatiev, Narodytska, and Marques-Silva 2019)). These explanations typically focus exclusively on aspects of the input deemed responsible (in different ways, according to the method used) for the outputs of the explained AI model, and offer little opportunity for interaction. For illustration, consider a recommender system providing positive and negative evidence drawn from input features as an explanation for a movie recommendation to a user: this form of explanation is static in that it does not support interactions between the system and the user, e.g. if the latter disagrees with the role of the input features in the explanation towards the recommendation, or with the system's recommendation itself.

A parallel research direction focuses on *argumentative explanations* for AI models of various types (see (Cyras et al. 2021; Vassiliades, Bassiliades, and Patkos 2021) for recent overviews), often motivated by the appeal of argumentation in explanations amongst humans, e.g. as in (Antaki and Leudar 1992), within the broader view that XAI should take findings from the social sciences into account (Miller 2019). Argumentative explanations in XAI employ *computational argumentation* (see (Atkinson et al. 2017; Baroni et al. 2018) for overviews), leveraging upon (existing or novel) argumentation frameworks, semantics and properties.

Argumentative explanations seem well suited to support interactivity when the mechanics of AI models can be abstracted away argumentatively (e.g. as for some recommender systems (Rago, Cocarascu, and Toni 2018) or neural networks (Albini et al. 2020; Potyka 2021)). For illustration, consider the case of a movie review aggregation system, as in (Cocarascu, Rago, and Toni 2019), and assume that its recommendation of a movie x and its reasoning therefor can be represented by the *bipolar argumentation framework* (BAF) (Cayrol and Lagasquie-Schiex 2005) $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ with *arguments* $\mathcal{X} = \{e, m_1, m_2\}$, *attacks* $\mathcal{A} = \varnothing$ and *supports* $\mathcal{S} = \{(m_1, e), (m_2, m_1)\}$ (see left of Figure 1 for a graphical visualisation). Then, by supporting e, m_1 (statically) conveys shallow evidence for the output (i.e. movie x being recommended). Argumentative explanations may go beyond the shallow nature of state-of-the-art explanations by facilitating dynamic, interactive explanations, e.g. by allowing a human explainee who does not agree with the machine's output or the evidence it provides (in other words, there is a *conflict* between the machine and the human) to provide feedback (in Figure 1, by introducing attacks (h_1, e) or (h_2, m_1)), while also allowing for the system to provide additional information (in Figure 1, by introducing the support (m_2, m_1)). The

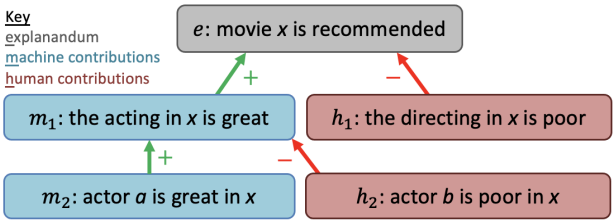

Figure 1: An argumentative explanation for a review aggregation system, amounting to the interactions between a machine and a human sharing their reasoning following a recommendation for x.

resulting interactive explanations can be seen as a *conflict resolution* process, e.g. as in (Raymond, Gunes, and Prorok 2020). Existing approaches focus on specific settings. Also, although the need for studying properties of explanations is well-acknowledged (e.g. see (Sokol and Flach 2020; Amgoud and Ben-Naim 2022)), to the best of our knowledge properties of interactive explanations, e.g. relating to how well they represent and resolve any conflicts, have been neglected to date. We fill these gaps by providing a general argumentative framework for interactive explanations as conflict resolution, as well as properties and instantiations thereof, backed by simulated experiments. Specifically:

- We define *Argumentative eXchanges* (AXs, §4), in which agents, whose reasoning is represented as *quantitative bipolar argumentation frameworks* (QBAFs) under gradual semantics (Baroni, Rago, and Toni 2018), contribute attacks/supports between arguments, to interactively obtain BAFs as in Figure 1 towards resolving conflicts on the agents' *stances* on explananda. We use QBAFs, which are BAFs where arguments are equipped with intrinsic strengths, as they are well suited to modelling private viewpoints, public conflicts, and resolutions, as well as cognitive *biases*, which are important in XAI (Bertrand et al. 2022). We use gradual semantics to capture individual *evaluations* of stance, taking biases into account.
- We identify and assess several properties (§5) which AXs may satisfy to be rendered suitable in an XAI setting. These properties concern, amongst others, the representation and possible resolution of conflicts within interactive explanations drawn from AXs.
- We instantiate AXs to the standard XAI setting of two agents, a machine and a human, and define a catalogue of agent behaviours for this setting (§6). We experiment in a simulated environment (§7) with the behaviours, exploring five hypotheses about conflict resolution and the accuracy of contributed arguments towards it, noting that the strongest argument is not always the most effective.

2 Related Work

There is a vast literature on *multi-agent argumentation*, e.g. recently, (Raymond, Gunes, and Prorok 2020) define an argumentation-based human-agent architecture integrating regulatory compliance, suitable for human-agent path deconfliction and based on abstract argumentation (Dung 1995); (Panisson, McBurney, and Bordini 2021) develop a multi-agent frameworks whereby agents can exchange information to jointly reason with argument schemes and critical questions; and (de Tarlé, Bonzon, and Maudet 2022) let agents debate using a shared abstract argumentation framework. These works mostly focus on narrow settings using structured and abstract argumentation under extension-based semantics, and mostly ignore the XAI angle ((Raymond, Gunes, and Prorok 2020; Calegari et al. 2022) are exceptions). Instead, with XAI as our core drive, we focus on (quantitative) bipolar argumentation under gradual semantics, motivated by their usefulness in several XAI approaches (e.g. in (Cocarascu, Rago, and Toni 2019; Albini et al. 2020; Potyka 2021; Rago, Baroni, and Toni 2022)). Other works consider (Q)BAFs in multi-agent argumentation, e.g. (Kontarinis and Toni 2015), but not for XAI. We adapt some aspects of these works on multi-agent argumentation approaches, specifically the idea of agents contributing attacks or supports (rather than arguments) to debates (Kontarinis and Toni 2015) and the restriction to trees rooted at explananda under gradual semantics from (de Tarlé, Bonzon, and Maudet 2022). We leave other interesting aspects they cover to future work, notably handling maliciousness (Kontarinis and Toni 2015), regulatory compliance (Raymond, Gunes, and Prorok 2020), and defining suitable utterances (Panisson, McBurney, and Bordini 2021).

Several approaches to obtain argumentative explanations for AI models exist (see (Cyras et al. 2021; Vassiliades, Bassiliades, and Patkos 2021) for overviews), often relying upon argumentative abstractions of the models. Our approach is orthogonal, as we assume that suitable QBAF abstractions of models and humans exist, focusing instead on formalising and validating interactive explanations.

Our AXs and agent behaviours are designed to *resolve conflicts* and are thus related to works on conflict resolution, e.g. (Black and Atkinson 2011; Fan and Toni 2012a), or centered around conflicts, e.g. (Pisano et al. 2022), but these works have different purposes to interactive XAI and use forms of argumentation other than (Q)BAFs under gradual semantics. Our agent behaviours can also be seen as attempts at *persuasion* in that they aim at selecting most efficacious arguments for changing the mind of the other agents, as e.g. in (Fan and Toni 2012b; Hunter 2018; Calegari, Riveret, and Sartor 2021; Donadello et al. 2022). Further, our AXs can be seen as supporting forms of *information-seeking and inquiry*, as they allow agents to share information, and are thus related to work in this spectrum (e.g. (Black and Hunter 2007; Fan and Toni 2015a)). Our framework however differs from general-purpose forms of argumentation-based persuasion/information-seeking/inquiry in its focus on interactive XAI supported by (Q)BAFs under gradual semantics.

The importance of machine handling of *information from humans* when explaining outputs, rather than the humans exclusively receiving information, has been highlighted e.g. for recommender systems (Balog, Radlinski, and Arakelyan 2019; Rago et al. 2020) and debugging (Lertvittayakumjorn, Specia, and Toni 2020) or other human-in-the-loop methods

(see (Wu et al. 2022) for a survey). Differently from these works, we capture *two-way* interactions.

Some works advocate *interactivity* in XAI (Paulino-Passos and Toni 2022), but do not make concrete suggestions on how to support it. Other works advocate dialogues for XAI (Lakkaraju et al. 2022), but it is unclear how these can be generated. We contribute to grounding the problem of generating interactive explanations by a computational framework implemented in a simulated environment.

3 Preliminaries

A BAF (Cayrol and Lagasquie-Schiex 2005) is a triple $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ such that $\mathcal{X}$ is a finite set (whose elements are *arguments*), $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ (called the *attack* relation) and $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{X}$ (called the *support* relation), where $\mathcal{A}$ and $\mathcal{S}$ are disjoint. A QBAF (Baroni et al. 2015) is a quadruple $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ such that $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ is a BAF and $\tau : \mathcal{X} \to \mathbb{I}$ ascribes *base scores* to arguments; these are values in some given $\mathbb{I}$ representing the arguments' intrinsic strengths. Given BAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ or QBAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$, for any $a \in \mathcal{X}$, we call $\{b \in \mathcal{X} | (b, a) \in \mathcal{A}\}$ the *attackers* of a and $\{b \in \mathcal{X} | (b, a) \in \mathcal{S}\}$ the *supporters* of a.

We make use of the following notation: given BAFs $\mathcal{B} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$, $\mathcal{B}' = \langle \mathcal{X}', \mathcal{A}', \mathcal{S}' \rangle$, we say that $\mathcal{B} \sqsubseteq \mathcal{B}'$ iff $\mathcal{X} \subseteq \mathcal{X}'$, $\mathcal{A} \subseteq \mathcal{A}'$ and $\mathcal{S} \subseteq \mathcal{S}'$; also, we use $\mathcal{B}' \setminus \mathcal{B}$ to denote $\langle \mathcal{X}' \setminus \mathcal{X}, \mathcal{A}' \setminus \mathcal{A}, \mathcal{S}' \setminus \mathcal{S} \rangle$. Similarly, given QBAFs $\mathcal{Q} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$, $\mathcal{Q}' = \langle \mathcal{X}', \mathcal{A}', \mathcal{S}', \tau' \rangle$, we say that $\mathcal{Q} \sqsubseteq \mathcal{Q}'$ iff $\mathcal{X} \subseteq \mathcal{X}'$, $\mathcal{A} \subseteq \mathcal{A}'$, $\mathcal{S} \subseteq \mathcal{S}'$ and $\forall a \in \mathcal{X} \cap \mathcal{X}'$ (which, by the other conditions, is exactly $\mathcal{X}$), it holds that $\tau'(a) = \tau(a)$. Also, we use $\mathcal{Q}' \setminus \mathcal{Q}$ to denote $\langle \mathcal{X}' \setminus \mathcal{X}, \mathcal{A}' \setminus \mathcal{A}, \mathcal{S}' \setminus \mathcal{S}, \tau'' \rangle$, where τ'' is τ' restricted to the arguments in $\mathcal{X}' \setminus \mathcal{X}$.[1] Given a BAF $\mathcal{B}$ and a QBAF $\mathcal{Q} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$, with an abuse of notation we use $\mathcal{B} \sqsubseteq \mathcal{Q}$ to stand for $\mathcal{B} \sqsubseteq \langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ and $\mathcal{Q} \sqsubseteq \mathcal{B}$ to stand for $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle \sqsubseteq \mathcal{B}$. For any BAFs or QBAFs $\mathcal{F}, \mathcal{F}'$, we say that $\mathcal{F} = \mathcal{F}'$ iff $\mathcal{F} \sqsubseteq \mathcal{F}'$ and $\mathcal{F}' \sqsubseteq \mathcal{F}$, and $\mathcal{F} \sqsubset \mathcal{F}'$ iff $\mathcal{F} \sqsubseteq \mathcal{F}'$ but $\mathcal{F} \neq \mathcal{F}'$.

Both BAFs and QBAFs may be equipped with a *gradual semantics* σ, e.g. as in (Baroni et al. 2017) for BAFs and as in (Potyka 2018) for QBAFs (see (Baroni, Rago, and Toni 2019) for an overview), ascribing to arguments a *dialectical strength* from within some given $\mathbb{I}$ (which, in the case of QBAFs, is typically the same as for base scores): thus, for a given BAF or QBAF $\mathcal{F}$ and argument a, $\sigma(\mathcal{F}, a) \in \mathbb{I}$.

Inspired by (de Tarlé, Bonzon, and Maudet 2022)'s use of (abstract) argumentation frameworks (Dung 1995) of a restricted kind (amounting to trees rooted with a single argument of focus), we use restricted BAFs and QBAFs:

Definition 1. *Let $\mathcal{F}$ be a BAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ or QBAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$. For any arguments $a, b \in \mathcal{X}$, let a* path *from a to b be defined as $(c_0, c_1), \ldots, (c_{n-1}, c_n)$ for some $n > 0$ (referred to as the* length *of the path) where $c_0 = a$, $c_n = b$ and, for any $1 \leq i \leq n$, $(c_{i-1}, c_i) \in \mathcal{A} \cup \mathcal{S}$.[2] Then, for $e \in \mathcal{X}$, $\mathcal{F}$ is a* BAF/QBAF *(resp.)* for e *iff i) $\exists (e, a) \in \mathcal{A} \cup \mathcal{S}$; ii) $\forall a \in \mathcal{X} \setminus \{e\}$, there is a path from a to e; and iii) $\nexists a \in \mathcal{X}$ with a path from a to a.*

Here e plays the role of an *explanandum*.[3] When interpreting the BAF/QBAF as a graph (with arguments as nodes and attacks/supports as edges), i) amounts to sanctioning that e admits no outgoing edges, ii) that e is reachable from any other node, and iii) that there are no cycles in the graph (and thus, when combining the three requirements, the graph is a multi-tree rooted at e). The restrictions in Definition 1 impose that every argument in a BAF/QBAF for e are "related" to e, in the spirit of (Fan and Toni 2015b).

In all illustrations (and in some of the experiments in §7) we use the *DF-QuAD* gradual semantics (Rago et al. 2016) for QBAFs for explananda. This uses $\mathbb{I} = [0, 1]$ and:

- a *strength aggregation function* Σ such that $\Sigma(()) = 0$ and, for $v_1, \ldots, v_n \in [0, 1]$ $(n \geq 1)$, if $n = 1$ then $\Sigma((v_1)) = v_1$, if $n = 2$ then $\Sigma((v_1, v_2)) = v_1 + v_2 - v_1 \cdot v_2$, and if $n > 2$ then $\Sigma((v_1, \ldots, v_n)) = \Sigma(\Sigma((v_1, \ldots, v_{n-1})), v_n)$;
- a *combination function* c such that, for $v^0, v^-, v^+ \in [0, 1]$: if $v^- \geq v^+$ then $c(v^0, v^-, v^+) = v^0 - v^0 \cdot \mid v^+ - v^- \mid$ and if $v^- < v^+$, then $c(v^0, v^-, v^+) = v^0 + (1 - v^0) \cdot \mid v^+ - v^- \mid$.

Then, for $\mathcal{F} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ and any $a \in \mathcal{X}$, given $\mathcal{A}(a) = \{b \in \mathcal{X} | (b, a) \in \mathcal{A}\}$ and $\mathcal{S}(a) = \{b \in \mathcal{X} | (b, a) \in \mathcal{S}\}$, $\sigma(\mathcal{F}, a) = c(\tau(a), \Sigma(\sigma(\mathcal{F}, \mathcal{A}(a))), \Sigma(\sigma(\mathcal{F}, \mathcal{S}(a))))$ where, for any $S \subseteq \mathcal{X}$, $\sigma(\mathcal{F}, S) = (\sigma(\mathcal{F}, a_1), \ldots, \sigma(\mathcal{F}, a_k))$ for $(a_1, \ldots, a_k)$, an arbitrary permutation of S.

4 Argumentative Exchanges (AXs)

We define AXs as a general framework in which *agents* argue with the goal of conflict resolution. The conflicts may arise when agents hold different *stances* on explananda. To model these settings, we rely upon QBAFs for explananda as abstractions of agents' internals. Specifically, we assume that each agent α is equipped with a QBAF and a gradual semantics (σ): the former provides an abstraction of the agent's knowledge/reasoning, with the base score (τ) representing *biases* over arguments; the latter can be seen as an *evaluation method* for arguments. To reflect the use of QBAFs in our multi-agent explanatory setting, we adopt this terminology (of biases and evaluation methods) in the remainder. Intuitively, biases and evaluations represent agents' views on the quality of arguments before and after, resp., other arguments are considered. For illustration, in the setting of Figure 1, biases may result from aggregations of votes from reviews for the machine and from personal views for the human, and evaluation methods allow the computation of the machine/human stance on the recommendation during the interaction (as in (Cocarascu, Rago, and Toni 2019)). Agents may choose their own *evaluation range* for measuring biases/evaluating arguments.

Definition 2. *An* evaluation range $\mathbb{I}$ *is a set equipped with a pre-order $\leq$ (where, as usual $x < y$ denotes $x \leq y$ and $y \not\leq x$) such that $\mathbb{I} = \mathbb{I}^+ \cup \mathbb{I}^0 \cup \mathbb{I}^-$ where $\mathbb{I}^+$, $\mathbb{I}^0$ and $\mathbb{I}^-$ are disjoint and for any $i \in \mathbb{I}^+$, $j \in \mathbb{I}^0$ and $k \in \mathbb{I}^-$, $k < j < i$. We refer to $\mathbb{I}^+$, $\mathbb{I}^0$ and $\mathbb{I}^-$, resp., as* positive, neutral *and* negative evaluations.

[1] Note that $\mathcal{B}' \setminus \mathcal{B}$, $\mathcal{Q}' \setminus \mathcal{Q}$ may not be BAFs, QBAFs, resp., as they may include no arguments but non-empty attack/support relations.

[2] Later, we will use $\mathtt{paths}(a, b)$ to indicate the set of all paths between arguments a and b, leaving the (Q)BAF implicit, and use $|p|$ for the length of path p. Also, we may see paths as sets of pairs.

[3] Other terms to denote the "focal point" of BAFs/QBAFs could be used. We use *explanandum* given our focus on the XAI setting.

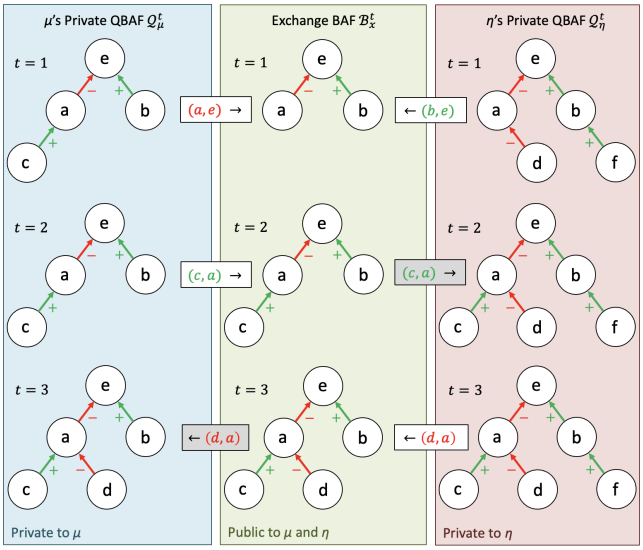

Figure 2: AX for explanandum e amongst agents $AG = \{\mu, \eta\}$, with the exchange BAF representing an interactive explanation. White (grey) boxes represent contributions (learnt relations, resp.).

Thus, an evaluation range discretises the space of possible evaluations into three categories.[4]

Definition 3. *A* private triple *for an agent α and an explanandum e is $(\mathbb{I}_\alpha, \mathcal{Q}_\alpha, \sigma_\alpha)$ where:*

- $\mathbb{I}_\alpha = \mathbb{I}^+_\alpha \cup \mathbb{I}^-_\alpha \cup \mathbb{I}^0_\alpha$ *is an evaluation range, referred to as α's* private evaluation range*;*
- $\mathcal{Q}_\alpha = \langle \mathcal{X}_\alpha, \mathcal{A}_\alpha, \mathcal{S}_\alpha, \tau_\alpha \rangle$ *is a QBAF for e, referred to as α's* private QBAF*, such that $\forall a \in \mathcal{X}_\alpha$, $\tau_\alpha(a) \in \mathbb{I}_\alpha$;*
- σ_α *is an evaluation method, referred to as α's* private evaluation method*, such that, for any QBAF $\mathcal{Q} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ ($\tau : \mathcal{X} \to \mathbb{I}_\alpha$) and, for any $a \in \mathcal{X}$, $\sigma_\alpha(\mathcal{Q}, a) \in \mathbb{I}_\alpha$.*

Agents' stances on explananda are determined by their private biases and evaluation methods.

Definition 4. *Let $(\mathbb{I}_\alpha, \mathcal{Q}_\alpha, \sigma_\alpha)$ be a private triple for agent α (for some e), with $\mathcal{Q}_\alpha = \langle \mathcal{X}_\alpha, \mathcal{A}_\alpha, \mathcal{S}_\alpha, \tau_\alpha \rangle$. Then, for $a \in \mathcal{X}_\alpha$, α's* stance on a *is defined, for $* \in \{-, 0, +\}$, as $\Sigma_\alpha(\mathcal{Q}_\alpha, a) = *$ iff $\sigma_\alpha(\mathcal{Q}_\alpha, a) \in \mathbb{I}^*_\alpha$.*

Note that a may be the explanandum or any other argument (namely, an agent may hold a stance on any arguments in its private QBAF). Also, abusing notation, we will lift the pre-order over elements of $\mathbb{I}$ to stances, whereby $- < 0 < +$.

In general, agents may hold different evaluation ranges, biases, QBAFs and evaluation methods, but the discretisation of the agents' evaluation ranges to obtain their stances allows for direct comparison across agents.

Example 1. *Consider a machine agent μ and a human agent η equipped resp. with private triples $(\mathbb{I}_\mu, \mathcal{Q}_\mu, \sigma_\mu)$ and $(\mathbb{I}_\eta, \mathcal{Q}_\eta, \sigma_\eta)$, with $\mathcal{Q}_\mu = \langle \mathcal{X}_\mu, \mathcal{A}_\mu, \mathcal{S}_\mu, \tau_\mu \rangle$, $\mathcal{Q}_\eta = \langle \mathcal{X}_\eta, \mathcal{A}_\eta, \mathcal{S}_\eta, \tau_\eta \rangle$ QBAFs for the same e and:*

- $\mathbb{I}^-_\mu = \mathbb{I}^-_\eta = [0, 0.5)$, $\mathbb{I}^0_\mu = \mathbb{I}^0_\eta = \{0.5\}$ *and* $\mathbb{I}^+_\mu = \mathbb{I}^+_\eta = (0.5, 1]$*;*
- $\mathcal{X}_\mu = \{e, a, b, c\}$, $\mathcal{A}_\mu = \{(a, e)\}$, $\mathcal{S}_\mu = \{(b, e), (c, a)\}$ *(represented graphically on the top left of Figure 2) and* $\tau_\mu(e) = 0.7$, $\tau_\mu(a) = 0.8$, $\tau_\mu(b) = 0.4$, *and* $\tau_\mu(c) = 0.6$*;*
- $\mathcal{X}_\eta = \{e, a, b, d, f\}$, $\mathcal{A}_\eta = \{(a, e), (d, a)\}$, $\mathcal{S}_\eta = \{(b, e), (f, b)\}$ *(represented on the top right of Figure 2) and* $\tau_\eta(e) = 0.6$, $\tau_\eta(a) = 0.8$, $\tau_\eta(b) = 0.2$, $\tau_\eta(d) = 0.6$ *and* $\tau_\eta(f) = 0.5$.
- σ_μ *is the DF-QuAD semantics, giving* $\sigma_\mu(\mathcal{Q}_\mu, e) = 0.336$, $\sigma_\mu(\mathcal{Q}_\mu, a) = 0.92$, $\sigma_\mu(\mathcal{Q}_\mu, b) = 0.4$, *and* $\sigma_\mu(\mathcal{Q}_\mu, c) = 0.6$*;*
- σ_η *is also DF-QuAD, giving* $\sigma_\eta(\mathcal{Q}_\eta, e) = 0.712$, $\sigma_\eta(\mathcal{Q}_\eta, a) = 0.32$, $\sigma_\eta(\mathcal{Q}_\eta, b) = 0.6$, $\sigma_\eta(\mathcal{Q}_\eta, d) = 0.6$, $\sigma_\eta(\mathcal{Q}_\eta, f) = 0.5$.

Thus, the machine and human agents hold entirely different views on the arguments (based on their private QBAFs and their evaluations) and $\Sigma_\mu(\mathcal{Q}_\mu, e) = -$ while $\Sigma_\eta(\mathcal{Q}_\eta, e) = +$. Thus, there is a conflict between the agents' stances on e.

We define AXs so that they can provide the ground to identify and resolve conflicts in stance amongst agents.

Definition 5. *An* Argumentative eXchange (AX) *for an explanandum e amongst agents AG (where $|AG| \geq 2$) is a tuple $\langle \mathcal{B}^0_x, \ldots, \mathcal{B}^n_x, AG^0, \ldots, AG^n, \mathcal{C} \rangle$ where $n > 0$ and:*

- *for every timestep $0 \leq t \leq n$:*
  - $\mathcal{B}^t_x = \langle \mathcal{X}^t_x, \mathcal{A}^t_x, \mathcal{S}^t_x \rangle$ *is a BAF for e, called the* exchange BAF *at t, such that $\mathcal{X}^0_x = \{e\}$, $\mathcal{A}^0_x = \mathcal{S}^0_x = \emptyset$ and for $t > 0$, $\mathcal{B}^{t-1}_x \sqsubseteq \mathcal{B}^t_x$;*
  - AG^t *is a set of private triples $(\mathbb{I}^t_\alpha, \mathcal{Q}^t_\alpha, \sigma^t_\alpha)$ for e, one for each agent $\alpha \in AG$, where, for $t > 0$, $\mathbb{I}^{t-1}_\alpha = \mathbb{I}^t_\alpha$, $\sigma^{t-1}_\alpha = \sigma^t_\alpha$, $\mathcal{Q}^{t-1}_\alpha \sqsubseteq \mathcal{Q}^t_\alpha$ and $\mathcal{Q}^t_\alpha \setminus \mathcal{Q}^{t-1}_\alpha \sqsubseteq \mathcal{B}^t_x \setminus \mathcal{B}^{t-1}_x$;*
- $\mathcal{C}$*, referred to as the* contributor mapping*, is a mapping such that, for every $(a, b) \in \mathcal{A}^n_x \cup \mathcal{S}^n_x$: $\mathcal{C}((a, b)) = (\alpha, t)$ with $0 < t \leq n$ and $\alpha \in AG$.*

Agents' private triples thus change over time during AXs, with several restrictions, in particular that agents do not change their evaluation ranges and methods, and that their biases on known arguments propagate across timesteps (but note that Definition 5 does not impose any restriction on the agents' private triples at timestep 0, other than they are all for e). The restriction that all BAFs/QBAFs in exchanges are *for the explanandum*, means that all contributed attacks and supports (and underlying arguments) are "relevant" to the explanandum. Implicitly, while we do not assume that agents share arguments, we assume that they agree on an underpinning 'lingua franca', so that, in particular, if two agents are both aware of two arguments, they must agree on any attack or support between them, e.g. it cannot be that an argument attacks another argument for one agent but not for another (in line with other works, e.g. (de Tarlé, Bonzon, and Maudet 2022; Raymond, Gunes, and Prorok 2020)). We leave to future work the study of the impact of this assumption in practice when AXs take place between machines and humans.

During AXs, agents contribute elements of the attack/support relations, thus "arguing" with one another. These elements cannot be withdrawn once contributed, in

[4] We choose three discrete values only for simplicity. This may mean that very close values, e.g. 0.49 and 0.51, belong to different categories. We leave to future work the analysis of further value categorisations, e.g. a distinction between strongly and mildly positive values or *comfort zones* (de Tarlé, Bonzon, and Maudet 2022).

line with human practices, and, by definition of $\mathcal{C}$, each element is said once by exactly one agent, thus avoiding repetitions that may occur in human exchanges. Note that we do not require that all agents contribute something to an AX, namely it may be that $\{\alpha | \mathcal{C}((a,b)) = (\alpha, t), (a,b) \in \mathcal{A}^n_x \cup \mathcal{S}^n_x\} \subset AG$. Also, we do not force agents to contribute something at every timestep (i.e. it may be the case that $\mathcal{B}^{t-1}_x = \mathcal{B}^t_x$ at some timestep t). Further, while the definition of AX does not impose that agents are truthful, from now on we will focus on truthful agents only and thus assume that if $(a,b) \in \mathcal{A}^n_x$ or $\mathcal{S}^n_x$ and $\mathcal{C}((a,b)) = (\alpha, t)$ (with $0 < t \leq n$), then, resp., $(a,b) \in \mathcal{A}^{t-1}_\alpha$ or $\mathcal{S}^{t-1}_\alpha$.

In the remainder, we may denote the private triple $(\mathbb{I}^t_\alpha, \mathcal{Q}^t_\alpha, \sigma^t_\alpha)$ as α^t and the stance $\Sigma_\alpha(\mathcal{Q}^t_\alpha, a)$ as $\Sigma^t_\alpha(a)$.

Example 2. *An AX amongst $\{\mu, \eta\}$ from Example 1 may be $\langle \mathcal{B}^0_x, \mathcal{B}^1_x, AG^0, AG^1, \mathcal{C}\rangle$ such that (see top row of Figure 2):*

- $\mathcal{B}^0_x = \langle \{e\}, \varnothing, \varnothing\rangle$, $\mathcal{B}^1_x = \langle \{e, a, b\}, \{(a,e)\}, \{(b,e)\}\rangle$;
- $\mu^0 = \mu^1$ *and* $\eta^0 = \eta^1$ *are as in Example 1;*
- $\mathcal{C}((a,e)) = (\mu, 1)$ *and* $\mathcal{C}((b,e)) = (\eta, 1)$, *i.e.* μ *and* η *contribute, resp., attack* (a,e) *and support* (b,e) *at 1.*

Here, each agent contributes a single attack or support justifying their stances (negative for μ and positive for η), but, in general, multiple agents may contribute multiple relations at single timesteps, or no relations at all.

When contributed attacks/supports are new to agents, they may *(rote) learn* them, with the arguments they introduce.

Definition 6. *Let $\langle \mathcal{B}^0_x, \ldots, \mathcal{B}^n_x, AG^0, \ldots, AG^n, \mathcal{C}\rangle$ be an AX amongst agents AG. Then, for any $\alpha \in AG$, with private tuples $(\mathbb{I}^0_\alpha, \mathcal{Q}^0_\alpha, \sigma^0_\alpha)$, …, $(\mathbb{I}^n_\alpha, \mathcal{Q}^n_\alpha, \sigma^n_\alpha)$:*

- *for any $0 < t \leq n$, for $\langle \mathcal{X}_\alpha, \mathcal{A}_\alpha, \mathcal{S}_\alpha, \tau_\alpha\rangle = \mathcal{Q}^t_\alpha \setminus \mathcal{Q}^{t-1}_\alpha$, $\mathcal{X}_\alpha, \mathcal{A}_\alpha$, and $\mathcal{S}_\alpha$ are, resp., the* learnt arguments, attacks, and supports *by α* at timestep *t;*
- *for $\langle \mathcal{X}_\alpha, \mathcal{A}_\alpha, \mathcal{S}_\alpha, \tau_\alpha\rangle = \mathcal{Q}^n_\alpha \setminus \mathcal{Q}^0_\alpha$, $\mathcal{X}_\alpha, \mathcal{A}_\alpha$, and $\mathcal{S}_\alpha$ are, resp., the* learnt arguments, attacks, and supports *by α.*

Note that, by definition of AXs, all learnt arguments, attacks and supports are from the (corresponding) exchange BAFs. Note also that in Example 2 neither agent learns anything, as indeed each contributed an attack/support already present in the other agent's private QBAF.

Example 3. *Let us extend the AX from Example 2 to obtain $\langle \mathcal{B}^0_x, \mathcal{B}^1_x, \mathcal{B}^2_x, AG^0, AG^1, AG^2, \mathcal{C}\rangle$ such that (see the top two rows of Figure 2):*

- $\mathcal{B}^2_x = \langle \{e, a, b, c\}, \{(a,e)\}, \{(b,e), (c,a)\}\rangle$
- $\mu^2 = \mu^1 = \mu^0$; η^2 *is such that* $\mathcal{Q}^2_\eta \supset \mathcal{Q}^1_\eta$ *where* $\mathcal{X}^2_\eta = \mathcal{X}^1_\eta \cup \{c\}$, $\mathcal{A}^2_\eta = \mathcal{A}^1_\eta$, $\mathcal{S}^2_\eta = \mathcal{S}^1_\eta \cup \{(c,a)\}$ *and* $\tau^2_\eta(c) = 0.2$;
- $\mathcal{C}((c,a)) = (\mu, 2)$, *namely μ contributes the support (c,a) in $\mathcal{B}^2_x$ at timestep 2.*

We will impose that any attack/support which is added to the exchange BAF by an agent is *learnt* by the other agents, alongside any new arguments introduced by those attacks/supports. Thus, for any $\alpha \in AG$ and $t > 0$, $\mathcal{B}^t_x \setminus \mathcal{B}^{t-1}_x \sqsubseteq \mathcal{Q}^t_\alpha$. However, agents have a choice on their biases on the learnt arguments. These biases could reflect, e.g., their trust on the contributing agents or the intrinsic quality of the arguments. Depending on these biases, learnt attacks and supports may influence the agents' stances on the explanandum differently. For illustration, in Example 3, η opted for a low bias (0.2) on the learnt argument c, resulting in $\sigma^2_\eta(\mathcal{Q}^2_\eta, e) = 0.648$, $\sigma^2_\eta(\mathcal{Q}^2_\eta, a) = 0.32$ and $\sigma^2_\eta(\mathcal{Q}^2_\eta, c) = 0.2$, and thus $\Sigma^2_\eta(e) = +$ still, as in Examples 1, 2. If, instead, η had chosen a high bias on the new argument, e.g. $\tau^2_\eta(c) = 1$, this would have given $\sigma^2_\eta(\mathcal{Q}^2_\eta, e) = 0.432$, $\sigma^2_\eta(\mathcal{Q}^2_\eta, a) = 0.88$ and $\sigma^2_\eta(\mathcal{Q}^2_\eta, c) = 1$, leading to $\Sigma^2_\eta(e) = -$, thus resolving the conflict. This illustration shows that learnt attacks, supports and arguments may fill gaps, change agents' stances on explananda and pave the way to the resolution of conflicts.

Definition 7. *Let $E = \langle \mathcal{B}^0_x, \ldots, \mathcal{B}^n_x, AG^0, \ldots, AG^n, \mathcal{C}\rangle$ be an AX for explanandum e amongst agents AG such that $\Sigma^0_\alpha(e) \neq \Sigma^0_\beta(e)$ for some $\alpha, \beta \in AG$. Then:*

- *E is* resolved at timestep *t, for some $0 < t \leq n$, iff $\forall \alpha, \beta \in AG$, $\Sigma^t_\alpha(e) = \Sigma^t_\beta(e)$, and is* unresolved at *t otherwise;*
- *E is* resolved *iff it is resolved at timestep n and it is unresolved at every timestep $0 \leq t < n$;*
- *E is* unresolved *iff it is unresolved at every $0 < t \leq n$.*

Thus, a resolved AX starts with a conflict between at least two agents and ends when no conflicts amongst any of the agents exist or when the agents give up on trying to find a resolution. Practically, AXs may be governed by a *turn-making function* $\pi : \mathbb{Z}^+ \to 2^{AG}$ determining which agents should contribute at any timestep. Then, an AX may be deemed to be unresolved if, for example, all agents decide, when their turn comes, against contributing.

Note that, while agents' biases and evaluations are kept private during AXs, we assume that agents share their stances on the explanandum, so that they are aware of whether the underpinning conflicts are resolved. Agents' stances, when ascertaining whether an AX is resolved, are evaluated internally by the agents, without any shared evaluation of the exchange BAF, unlike, e.g. in (de Tarlé, Bonzon, and Maudet 2022) and other works we reviewed in §2.

Finally, note that our definition of AX is neutral as to the role of agents therein, allowing in particular that agents have symmetrical roles (which is natural, e.g., for inquiry) as well as asymmetrical roles (which is natural, e.g., when machines explain to humans: this will be our focus from §5).

5 Explanatory Properties of AXs

Here we focus on singling out desirable properties that AXs may need satisfy to support interactive XAI. Let us assume as given an AX $E = \langle \mathcal{B}^0_x, \ldots, \mathcal{B}^n_x, AG^0, \ldots, AG^n, \mathcal{C}\rangle$ for e as in Definition 5. The first three properties impose basic requirements on AXs so that they result in fitting explanations.

Property 1. E satisfies connectedness *iff for any $0 \leq t \leq n$, if $|\mathcal{X}^t_x| > 1$ then $\forall a \in \mathcal{X}^t_x$, $\exists b \in \mathcal{X}^t_x$ such that $(a,b) \in \mathcal{A}^t_x \cup \mathcal{S}^t_x$ or $(b,a) \in \mathcal{A}^t_x \cup \mathcal{S}^t_x$.*

Basically, connectedness imposes that there should be no floating arguments and no "detours" in the exchange BAFs, at any stage during the AX. It is linked to directional connectedness in (Cyras, Kampik, and Weng 2022). A violation

of this property would lead to counter-intuitive (interactive) explanations, with agents seemingly "off-topic".

Property 2. E satisfies acyclicity *iff for any* $0 \leq t \leq n$, $\nexists a \in \mathcal{X}_x^t$ *such that* $\mathsf{paths}(a, a) \neq \emptyset$.

Acyclicity ensures that all reasoning is directed towards the explanandum in AXs. A violation of this property may lead to seemingly non-sensical (interactive) explanations.

Property 3. E satisfies contributor irrelevance *iff for any AX for* e $\langle \mathcal{B}_x^{0\prime}, \ldots, \mathcal{B}_x^{n\prime}, AG^{0\prime}, \ldots, AG^{n\prime}, \mathcal{C}' \rangle$, *if* $\mathcal{B}_x^{0\prime} = \mathcal{B}_x^0$, $\mathcal{B}_x^{n\prime} = \mathcal{B}_x^n$, $AG^{0\prime} = AG^0$, *then* $\forall \alpha \in AG$: $\Sigma_\alpha^n(\mathcal{Q}_\alpha^n, e) = \Sigma_\alpha^n(\mathcal{Q}_\alpha^{n\prime}, e)$.

Contributor irrelevance ensures that the same final exchange BAF results in the same stances for all agents, regardless of the contributors of its attacks and supports or the order in which they were contributed.

These three properties are basically about the exchange BAFs in AXs, and take the viewpoint of an external "judge" for the explanatory nature of AXs. These basic properties are all satisfied, by design, by AXs:[5]

Proposition 1. *Every AX satisfies Properties 1 to 3.*

We now introduce properties which AXs may not always satisfy, but which, nonetheless, may be desirable if AXs are to generate meaningful (interactive) explanations. First, we define notions of *pro* and *con* arguments in AXs, amounting to positive and negative reasoning towards the explanandum.

Definition 8. *Let* $\mathcal{B} = \langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ *be any BAF for* e. *Then, the* pro arguments *and* con arguments *for* $\mathcal{B}$ *are, resp.:*

- $\mathsf{pro}(\mathcal{B}) = \{a \in \mathcal{X} \mid \exists p \in \mathsf{paths}(a, e),$ *where* $|p \cap \mathcal{A}|$ *is even*$\}$;
- $\mathsf{con}(\mathcal{B}) = \{a \in \mathcal{X} \mid \exists p \in \mathsf{paths}(a, e),$ *where* $|p \cap \mathcal{A}|$ *is odd*$\}$.

Note that the intersection of pro and con arguments may be non-empty as multiple paths to explananda may exist, so an argument may bring both positive and negative reasoning.

Pro/con arguments with an even/odd, resp., number of attacks in their path to e are related to chains of supports (*supported/indirect defeats*, resp.) in (Cayrol and Lagasquie-Schiex 2005) (we leave the study of formal links to future work). Pro/con arguments are responsible for increases/decreases, resp., in e's strength using DF-QuAD:

Proposition 2. *For any* $\alpha \in AG$, *let* σ_α *indicate the evaluation method by DF-QuAD. Then, for any* $0 < t \leq n$:

- *if* $\sigma_\alpha(\mathcal{Q}_\alpha^t, e) > \sigma_\alpha(\mathcal{Q}_\alpha^{t-1}, e)$, *then* $\mathsf{pro}(\mathcal{B}_x^t) \supset \mathsf{pro}(\mathcal{B}_x^{t-1})$;
- *if* $\sigma_\alpha(\mathcal{Q}_\alpha^t, e) < \sigma_\alpha(\mathcal{Q}_\alpha^{t-1}, e)$, *then* $\mathsf{con}(\mathcal{B}_x^t) \supset \mathsf{con}(\mathcal{B}_x^{t-1})$.

We conjecture (but leave to future work) that this result (and more later) holds for other gradual semantics satisfying *monotonicity* (Baroni, Rago, and Toni 2019) or *bi-variate monotony/reinforcement* (Amgoud and Ben-Naim 2018).

Property 4. E *satisfies* resolution representation *iff* E *is resolved and* $\forall \alpha \in AG$: *if* $\Sigma_\alpha^n(e) > \Sigma_\alpha^0(e)$, *then* $\mathsf{pro}(\mathcal{B}_x^n) \neq \emptyset$; *and if* $\Sigma_\alpha^n(e) < \Sigma_\alpha^0(e)$, *then* $\mathsf{con}(\mathcal{B}_x^n) \neq \emptyset$.

This property also takes the viewpoint of an external "judge", by imposing that the final exchange BAF convincingly represents a resolution of the conflicts between agents' stances, thus showing why stances were changed. Specifically, it imposes that a changed stance must be the result of pro or con arguments (depending on how stances have changed). For example, in Figure 2, b, d, f are pro arguments which could justify an increase in stance for e, while a, c are con arguments which could justify its decrease. Note that this property does not hold in general, e.g., given an agent which (admittedly counter-intuitively) increases its evaluation of arguments when they are attacked. However, it holds for some evaluation models, notably DF-QuAD again:

Proposition 3. *If* E *is resolved and* $\forall \alpha \in AG$, σ_α *is DF-QuAD, then* E *satisfies resolution representation.*

The final property we consider concerns unresolved AXs, in the same spirit as resolution representation.

Property 5. E *satisfies* conflict representation *iff* E *is unresolved,* $\mathsf{pro}(\mathcal{B}_x^n) \neq \emptyset$ *and* $\mathsf{con}(\mathcal{B}_x^n) \neq \emptyset$.

This property thus requires that the conflict in an unresolved AX is apparent in the exchange BAF, namely it includes both pro and con arguments (representing the conflicting stances). For example, if the AX in Figure 2 concluded unresolved at $t = 2$, this property requires that $\mathcal{B}_x^2$ contains both pro arguments for e (e.g. a or c) and con arguments against it (e.g. b). This property does not hold in general, e.g. for an agent who rejects all arguments by imposing on them minimum biases and contributes no attack or support. Proving that this property holds requires consideration of the agents' behaviour, which we examine next.

6 Agent Behaviour in AXs for XAI

All our examples so far have illustrated how AXs may support explanatory interactions amongst a *machine* μ and a *human* η. This specific XAI setting is our focus in the remainder, where we assume $AG = \{\mu, \eta\}$. Also, for simplicity, we impose (as in all illustrations) that $\mathbb{I}_\mu = \mathbb{I}_\eta = [0, 1]$, $\mathbb{I}_\mu^- = \mathbb{I}_\eta^- = [0, 0.5)$, $\mathbb{I}_\mu^0 = \mathbb{I}_\eta^0 = \{0.5\}$ and $\mathbb{I}_\mu^+ = \mathbb{I}_\eta^+ = (0.5, 1]$. We also restrict attention to AXs governed by a turn-making function π imposing a strict interleaving such that $\pi(i) = \{\mu\}$ if i is odd, and $\pi(i) = \{\eta\}$ otherwise (thus, in particular, the machine starts the interactive explanation process).

In line with standard argumentative XAI, the machine may draw the QBAF in its *initial private triple* (at $t = 0$) from the model it is explaining. This QBAF may be obtained by virtue of some abstraction methodology or may be the basis of the model itself (see (Cyras et al. 2021)). The humans, instead, may draw the QBAF in their initial private triple, for example, from their own knowledge, biases, and/or regulations on the expected machine's behaviour. The decision on the evaluation method, for machines and humans, may be dictated by specific settings and desirable agent properties therein. Here we focus on how to formalise and evaluate interactive explanations between a machine and a human using AXs, and ignore how their initial private triples are obtained.

Below we define various behaviours dictating how machines and humans can engage in AXs for XAI, focusing on ways to i) determine their biases and ii) decide their contributions (attacks/supports) to (unresolved) AXs.

Biases. As seen in §4, the degree to which learnt attacks/supports impact the stances of agents on explananda is determined by the agents' biases on the learnt arguments.

[5] Proofs for all propositions are in arxiv.org/abs/2303.15022.

In XAI different considerations regarding this learning apply to machines and humans. Firstly, not all machines may be capable of learning: simple AI systems which provide explanations but do not have the functionality for understanding any input from humans are common in AI. Secondly, machines capable of learning may assign different biases to the learnt arguments: a low bias indicates scepticism while a high bias indicates credulity. Machines may be designed to give low biases to arguments from sources which cannot be trusted, e.g. when the expertise of a human is deemed insufficient, or high biases to arguments when the human is deemed competent, e.g. in debugging. Here, we refrain from accommodating such challenges and focus instead on the restrictive (but sensible, as a starting point) case where machines assign constant biases to arguments from humans.

Definition 9. *Let $c \in [0,1]$ be a chosen constant. For any learnt argument $a \in \mathcal{X}_\mu^t \setminus \mathcal{X}_\mu^{t-1}$ at timestep t, $\tau_\mu^t(a) = c$.*

If $c = 0$ then the machine is unable to learn, whereas $0 < c < 1$ gives partially sceptical machines and $c = 1$ gives credulous machines. The choice of c thus depends on the specific setting of interest, and may have an impact on the conflict resolution desideratum for AXs. For example, let μ use DF-QuAD as its evaluation method: if $c = 1$ we can derive guarantees of rejection/weakening or acceptance/strengthening of arguments which are attacked or supported, resp., by learnt arguments,[6] demonstrating the potential (and dangers) of credulity in machines (see §7).

Humans, meanwhile, typically assign varying biases to arguments based on their own internal beliefs. These assignments may reflect cognitive biases such as the *confirmation bias* (Nickerson 1998) – the tendency towards looking favourably at evidence which supports one's prior views. In §7 we model humans so that they assign random biases to learnt arguments, but explore confirmation bias by applying a constant offset to reduce the bias assigned by the human. This differs, e.g., from the modelling of confirmation bias in (de Tarlé, Bonzon, and Maudet 2022), acting on the probability of an argument being learned. We leave the exploration of alternatives for assigning biases to future work.

Attack/Support Contributions. We consider *shallow*, *greedy* and *counterfactual* behaviours: intuitively, the first corresponds to the one-shot explanations in most XAI, the second contributes the (current) strongest argument in favour of the agent position, and the third considers how each attack/support may (currently) affect the exchange BAF before it is contributed. All behaviours identify argument pairs to be added to the exchange BAF as attacks or supports reflecting their role in the private QBAFs from which they are drawn. We use the following notion:

Definition 10. *For E resolved at timestep t, if $\Sigma_\mu^t(e) > \Sigma_\eta^t(e)$ then the* states of *μ and η* at *t are, resp.,* arguing for *and* arguing against *e (else, the states are reversed).*

The agents' states point to a "window for persuasion", whereby an agent arguing for (against) e may wish to attempt to increase (decrease, resp.) the stance of the other agent, without accessing their private QBAFs, thus differing from other works, e.g. (de Tarlé, Bonzon, and Maudet 2022), which rely on shared evaluations: in our case, reasoning is shared but it is not evaluated in a shared manner.

The *shallow* behaviour selects a (bounded by max) maximum number of supports for/attacks against the explanandum if the agent is arguing for/against, resp., it, as follows:

Definition 11. *Let $max \in \mathbb{N}$. Agent $\alpha \in AG$ exhibits* shallow behaviour *(wrt max) iff, at any $0 \leq t < n$ where $\pi(t) = \{\alpha\}$, $C = \{(a,b) | \mathcal{C}((a,b)) = (\alpha, t)\}$ is a maximal (wrt cardinality) set $\{(a_1, e), \ldots, (a_p, e)\}$ with $p \leq max$ such that:*

- *if α is arguing for e then $C \subseteq \mathcal{S}_\alpha^{t-1} \setminus \mathcal{S}_x^{t-1}$ where $\forall i \in \{1, \ldots, p\}, \nexists (b,e) \in \mathcal{S}_\alpha^{t-1} \setminus (\mathcal{S}_x^{t-1} \cup C)$ with $\sigma_\alpha^{t-1}(b) > \sigma_\alpha^{t-1}(a_i)$;*
- *if α is arguing against e then $C \subseteq \mathcal{A}_\alpha^{t-1} \setminus \mathcal{A}_x^{t-1}$ where $\forall i \in \{1, \ldots, p\}, \nexists (b,e) \in \mathcal{A}_\alpha^{t-1} \setminus (\mathcal{A}_x^{t-1} \cup C)$ with $\sigma_\alpha^{t-1}(b) > \sigma_\alpha^{t-1}(a_i)$.*

This behaviour thus focuses on reasoning for or against the explanandum e exclusively. It selects supports or attacks in line with the agent's stance on e and with the highest evaluation in the contributing agent's private QBAF. This behaviour is inspired by *static* explanation methods in XAI, which deliver all information in a single contribution. Clearly, if we let μ exhibit this shallow behaviour and η be *unresponsive*, i.e. never contribute any attack/support, then the AX cannot satisfy conflict representation.

The *greedy* behaviour allows an agent arguing for e to support the pro or attack the con arguments, while that arguing against can support the con or attack the pro arguments.

Definition 12. *Agent $\alpha \in AG$ exhibits* greedy behaviour *iff, at any $0 \leq t < n$ where $\pi(t) = \{\alpha\}$, $C = \{(a,b) | \mathcal{C}((a,b)) = (\alpha, t)\}$ is empty or amounts to a single attack or support $(a,b) \in (\mathcal{A}_\alpha^{t-1} \cup \mathcal{S}_\alpha^{t-1}) \setminus (\mathcal{A}_x^{t-1} \cup \mathcal{S}_x^{t-1})$ such that:*

1. *if α is arguing for e then: $(a,b) \in \mathcal{S}_\alpha^{t-1}$ and $b \in \mathsf{pro}(\mathcal{B}_x^{t-1}) \cup \{e\}$; or $(a,b) \in \mathcal{A}_\alpha^{t-1}$ and $b \in \mathsf{con}(\mathcal{B}_x^{t-1})$; if α is arguing against e then: $(a,b) \in \mathcal{S}_\alpha^{t-1}$ and $b \in \mathsf{con}(\mathcal{B}_x^{t-1})$; or $(a,b) \in \mathcal{A}_\alpha^{t-1}$ and $b \in \mathsf{pro}(\mathcal{B}_x^{t-1}) \cup \{e\}$;*
2. *$\nexists (a', b') \in (\mathcal{A}_\alpha^{t-1} \cup \mathcal{S}_\alpha^{t-1}) \setminus (\mathcal{A}_x^{t-1} \cup \mathcal{S}_x^{t-1})$ satisfying 1. such that $\sigma_\alpha^{t-1}(a') > \sigma_\alpha^{t-1}(a)$;*
3. *$\nexists (a'', b'') \in (\mathcal{A}_\alpha^{t-1} \cup \mathcal{S}_\alpha^{t-1}) \setminus (\mathcal{A}_x^{t-1} \cup \mathcal{S}_x^{t-1})$ satisfying 1. such that $\sigma_\alpha^{t-1}(a'') = \sigma_\alpha^{t-1}(a)$ and $|argmin_{P'' \in \mathsf{paths}(a'', e)} |P''| | < |argmin_{P \in \mathsf{paths}(a,e)} |P| |$.*

Intuitively, 1. requires that the attack or support, if any, is in line with the agent's views; 2. ensures that the attacking or supporting argument has maximum strength; and 3. ensures that it is "close" to the explanandum. We posit that enforcing agents to contribute at most one argument per turn will aid *minimality* without affecting conflict resolution negatively wrt the shallow behaviour (see §7). Minimality is a common property of explanations in XAI, deemed beneficial both from a machine perspective, e.g. wrt computational aspects (see *computational complexity* in (Sokol and Flach 2020)), and from a human perspective, e.g. wrt cognitive load and privacy maintenance (see *parsimony* in (Sokol and Flach 2020)). Naturally, however, conflict resolution in AXs should always take precedence over minimality, as prioritising the latter would force AXs to remain empty.

[6] Propositions on such effects are in arxiv.org/abs/2303.15022.

Proposition 4. *If E is unresolved and $\forall \alpha \in AG$: α exhibits greedy behaviour and $\{(a,b) \in \mathcal{A}_x^n \cup \mathcal{S}_x^n | \mathcal{C}((a,b)) = (\alpha, t), t \in \{1,\ldots,n\}\} \neq \emptyset$, then E satisfies conflict representation.*

Proposition 5. *If $\forall \alpha \in AG$, for all $0 \leq t < n$ and $\forall a \in \mathcal{X}_\alpha^t$, $\mathsf{paths}((a,e)) = \{(a,e)\}$, then the shallow (with $max = 1$) and greedy behaviours are aligned.*

The greedy behaviour may not always lead to resolutions:

Example 4. *Let us extend the AX from Example 3 to $\langle \mathcal{B}_x^0, \ldots, \mathcal{B}_x^3, AG^0, \ldots, AG^3, \mathcal{C} \rangle$ such that (see Figure 2):*

- $\mathcal{B}_x^3 = \langle \{e,a,b,c,d\}, \{(a,e),(d,a)\}, \{(b,e),(c,a)\} \rangle$;
- $\eta^3 = \eta^2$; μ^3 *is such that* $\mathcal{Q}_\mu^3 \supset \mathcal{Q}_\mu^2$ *where* $\mathcal{X}_\mu^3 = \mathcal{X}_\mu^2 \cup \{d\}$, $\mathcal{A}_\mu^3 = \mathcal{A}_\mu^2 \cup \{(d,a)\}$, $\mathcal{S}_\mu^3 = \mathcal{S}_\mu^2$, $\tau_\mu^3(d) = 0.6$; *then, the argument evaluations are* $\sigma_\mu^3(\mathcal{Q}_\mu^3, e) = 0.42$, $\sigma_\mu^3(\mathcal{Q}_\mu^3, a) = 0.8$, $\sigma_\mu^3(\mathcal{Q}_\mu^3, b) = 0.4$, $\sigma_\mu^3(\mathcal{Q}_\mu^3, c) = 0.6$, $\sigma_\mu^3(\mathcal{Q}_\mu^3, d) = 0.6$;
- $\mathcal{C}((d,a)) = (\eta, 3)$, *i.e.* η *contributes attack* (d,a) *at* $t = 3$.

Here, in line with the greedy behaviour, μ learns the attack (d,a) contributed by η at timestep 3. Then, even if μ assigns the same bias to these learnt arguments as η (which is by no means guaranteed), this is insufficient to change the stance, i.e. $\Sigma_\mu^3(e) = -$, and so the AX remains unresolved.

The final *counterfactual* behaviour takes greater consideration of the argumentative structure of the reasoning available to the agents in order to maximise the chance of conflict resolution with a limited number of arguments contributed. This behaviour is defined in terms of the following notion.

Definition 13. *Given an agent $\alpha \in AG$,* a private view of the exchange BAF by α at timestep t *is any* $\mathcal{Q}_{\alpha v}^t = \langle \mathcal{X}_{\alpha v}^t, \mathcal{A}_{\alpha v}^t, \mathcal{S}_{\alpha v}^t, \tau_{\alpha v}^t \rangle$ *such that* $\mathcal{B}_x^t \sqsubseteq \mathcal{Q}_{\alpha v}^t \sqsubseteq \mathcal{Q}_\alpha^t$.

An agent's private view of the exchange BAF thus projects their private biases onto the BAF, while also potentially accommodating *counterfactual reasoning* with additional arguments. Based on arguments' evaluations in an agent's private view, the agent can then judge which attack or support *it perceives* will be the most effective.

Definition 14. *Given an agent $\alpha \in AG$, α's* perceived effect on e *at $0 < t \leq n$ of any $(a,b) \in (\mathcal{A}_\alpha^{t-1} \cup \mathcal{S}_\alpha^{t-1}) \setminus (\mathcal{A}_x^{t-1} \cup \mathcal{S}_x^{t-1})$, where $a \in \mathcal{X}_\alpha^{t-1} \setminus \mathcal{X}_x^{t-1}$ and $b \in \mathcal{X}_x^{t-1}$, is $\epsilon((a,b), \mathcal{Q}_\alpha^t) = \sigma_\alpha(\mathcal{Q}_{\alpha v}^t, e) - \sigma_\alpha(\mathcal{Q}_{\alpha v}^{t-1}, e)$ for $\mathcal{Q}_{\alpha v}^t \supset \mathcal{Q}_{\alpha v}^{t-1}$ a private view of the exchange BAF at t by α such that $\mathcal{X}_{\alpha v}^t = \mathcal{X}_{\alpha v}^{t-1} \cup \{a\}$, $\mathcal{A}_{\alpha v}^t = (\mathcal{X}_{\alpha v}^t \times \mathcal{X}_{\alpha v}^t) \cap \mathcal{A}_\alpha^{t-1}$ and $\mathcal{S}_{\alpha v}^t = (\mathcal{X}_{\alpha v}^t \times \mathcal{X}_{\alpha v}^t) \cap \mathcal{S}_\alpha^{t-1}$.*

The counterfactual view underlying this notion of perceived effect relates to (Kampik and Cyras 2022), although we consider the effect of adding an attack or support, whereas they consider an argument's *contribution* by removing it. It also relates to the *hypothetical value* of (de Tarlé, Bonzon, and Maudet 2022), which however amounts to the explanandum's evaluation in the shared graph.

Definition 15. *Agent $\alpha \in AG$ exhibits* counterfactual behaviour *iff, at any $0 \leq t < n$ where $\pi(t) = \{\alpha\}, C = \{(a,b) | \mathcal{C}((a,b)) = (\alpha, t)\}$ is empty or is $\{(a,b)\}$ such that:*

- *if α is arguing for e then $\epsilon((a,b), \mathcal{Q}_\alpha^t) > 0$ and (a,b) is $argmax_{(a',b') \in (\mathcal{A}_\alpha^{t-1} \cup \mathcal{S}_\alpha^{t-1}) \setminus (\mathcal{A}_x^{t-1} \cup \mathcal{S}_x^{t-1})} \epsilon((a',b'), \mathcal{Q}_\alpha^t)$;*
- *if α is arguing against e then $\epsilon((a,b), \mathcal{Q}_\alpha^t) < 0$ and (a,b) is $argmin_{(a',b') \in (\mathcal{A}_\alpha^{t-1} \cup \mathcal{S}_\alpha^{t-1}) \setminus (\mathcal{A}_x^{t-1} \cup \mathcal{S}_x^{t-1})} \epsilon((a',b'), \mathcal{Q}_\alpha^t)$.*

Identifying attacks and supports based on their effect on the explanandum is related to *proponent* and *opponent arguments* in (Cyras, Kampik, and Weng 2022), defined however in terms of *quantitative dispute trees* for BAFs.

The counterfactual behaviour may better consider argumentative structure, towards resolved AXs, as shown next.

Example 5. *Consider the AX from Example 4 but where:*

- $\mathcal{B}_x^3 = \langle \{e,a,b,c,f\}, \{(a,e)\}, \{(b,e),(c,a),(f,b)\} \rangle$;
- μ^3 *is such that* $\mathcal{Q}_\mu^3 \supset \mathcal{Q}_\mu^2$ *where* $\mathcal{X}_\mu^3 = \mathcal{X}_\mu^2 \cup \{f\}$, $\mathcal{A}_\mu^3 = \mathcal{A}_\mu^2$, $\mathcal{S}_\mu^3 = \mathcal{S}_\mu^2 \cup \{(f,b)\}$, $\tau_\mu^3(f) = 0.5$; *then, the argument evaluations are* $\sigma_\mu^3(\mathcal{Q}_\mu^3, e) = 0.546$, $\sigma_\mu^3(\mathcal{Q}_\mu^3, a) = 0.92$, $\sigma_\mu^3(\mathcal{Q}_\mu^3, b) = 0.7$, $\sigma_\mu^3(\mathcal{Q}_\mu^3, c) = 0.6$ *and* $\sigma_\mu^3(\mathcal{Q}_\mu^3, f) = 0.5$;
- $\mathcal{C}((f,b)) = (\eta, 3)$, *i.e.* η *contributes support* (f,b) *to* $\mathcal{B}_x^3$.

Here, η contributes (f,b) in line with the counterfactual behaviour as $\epsilon((f,b), \mathcal{Q}_\eta^3) = 0.24 > \epsilon((d,a), \mathcal{Q}_\eta^3) = 0.216$. This sufficiently modifies μ's private QBAF such that $\Sigma_\mu^3 = +$, and the AX is now resolved: the counterfactual behaviour succeeds where the greedy behaviour did not (Example 4).

We end showing some conditions under which conflict representation is satisfied by the counterfactual behaviour.

Proposition 6. *If E is unresolved and is such that $\forall \alpha \in AG$: α exhibits counterfactual behaviour; σ_α is DF-QuAD; $\{(a,b) \in \mathcal{A}_x^n \cup \mathcal{S}_x^n | \mathcal{C}((a,b)) = (\alpha, t), t \in \{1,\ldots,n\}\} \neq \emptyset$; then E satisfies conflict representation.*

7 Evaluation

We now evaluate sets of AXs obtained from the behaviours from §6 via simulations, using the following metrics:[7]

Resolution Rate (RR): the proportion of resolved AXs.

Contribution Rate (CR): the average number of arguments contributed to the exchange BAFs in the resolved AXs, in effect measuring the total information exchanged.

Persuasion Rate (PR): for an agent, the proportion of resolved AXs in which the agent's initial stance is the other agent's final stance, measuring the agent's persuasiveness.

Contribution Accuracy (CA): for an agent, the proportion of the contributions which, if the agent was arguing for (against) e, would have maximally increased (decreased, resp.) e's strength in the other agent's private QBAF.

We tested PR and CA for machines only. Let *unresponsive behaviour* amount to contributing nothing (as in §6). Then, our hypotheses were:

H1: For a *shallow machine* and an *unresponsive human*, as the max *constant* increases, RR, CR and CA increase.

H2: For a *shallow machine* and an *unresponsive human*, as the human's *confirmation bias* increases, RR decreases.

H3: For a *greedy machine* and a *counterfactual human*, RR increases *relative to a shallow machine* and an *unresponsive human*.

H4: For a *greedy machine* and a *counterfactual human*, as the machine's *bias on learnt arguments* increases, RR increases while CR and PR decrease.

H5: For a *counterfactual machine* and a *counterfactual human*, RR and CA increase *relative to a greedy machine*.

[7] See arxiv.org/abs/2303.15022 for exact formulations.

Experimental Setup. For each AX for e (restricted as in §6), we created a "universal BAF", i.e. a BAF for e of which all argumentation frameworks are subgraphs. We populated the universal BAFs with 30 arguments by first generating a 6-ary tree with e as the root. Then, any argument other than e had a 50% chance of having a directed edge towards a random previous argument in the tree, to ensure that multiple paths to the explanandum are present. 50% of the edges in the universal BAF were randomly selected to be attacks, and the rest to be supports. We built agents' private QBAFs from the universal BAF by performing a random traversal through the universal BAF and stopped when the QBAFs reached 15 arguments, selecting a random argument from each set of children, as in (de Tarlé, Bonzon, and Maudet 2022). We then assigned random biases to arguments in the agents' QBAFs, and (possibly different) random evaluation methods to agents amongst QuAD (Baroni et al. 2015), DF-QuAD, REB (Amgoud and Ben-Naim 2017) and QEM (Potyka 2018) (all with evaluation range $[0, 1]$). We used different evaluation methods to simulate different ways to evaluate arguments in real-world humans/machines. We repeated this process till agents held different stances on e.

For each hypothesis, we ran 1000 experiments per configuration, making sure the experiments for different strategies are run with the same QBAFs. We ran the simulations on the NetLogo platform using BehaviorSpace.[8] We tested the significance between testing conditions in a pairwise manner using the chi-squared test for the discrete measures RR and PR, and Student's t-test for the continuous measures CR and CA. We rejected the null hypotheses when $p < 0.01$.

Experimental Results. Table 1 reports the results of our simulations: all hypotheses were (at least partially) verified.

H1: As expected, increasing max for shallow machines results in significantly higher RR, CR and CA up to max = 3 ($p < 0.005$ for max values of 1 vs 2 and 2 vs 3 for all metrics). Above this limit (max values of 3 vs 4 and 4 vs 5), this trend was no longer apparent, suggesting that there was a limit to the effectiveness of contributing arguments at this distance from e. Note that the machine's PR is always 100% here, since the (unresponsive) human does not contribute.

H2: We fixed $max = 4$ (the value with the maximum RR for H1) and found that increasing the confirmation bias in the human significantly decreased the machine's RR initially ($p < 0.01$ for 0 vs -0.1 and -0.1 vs -0.2), before the effect tailed off as RR became very low ($p = 0.09$ for -0.2 vs -0.3 and $p = 0.03$ for -0.3 vs -0.4), demonstrating the need for behaviours which consider deeper reasoning than the shallow behaviour to achieve higher resolution rates.

H3: From here onwards we tested with a counterfactual human[9] and fixed the level of confirmation bias therein to -0.2. We compared shallow against greedy machines, also limiting the number of arguments they contributed to maxima of three and four to compare fairly with the shallow machine with the fixed *max* constant. RR increased significantly with the greedy behaviour ($p < 0.001$), over the shallow machine which remained statistically significant when we restricted the greedy machine's contributed arguments to 4 ($p < 0.005$), but not to 3 ($p = 0.202$).

	Behaviour		**Learning**		**RR**	**CR**	**PR$_\mu$**	**CA$_\mu$**
	μ	η	μ	η				
H1	S (1)	-	-	0	5.4	1	100	45.4
	S (2)	-	-	0	9.6	1.96	100	51.9
	S (3)	-	-	0	13.0	2.76	100	56.7
	S (4)	**-**	**-**	**0**	**13.9**	**3.22**	**100**	**58.1**
	S (5)	-	-	0	13.7	3.38	100	58.3
H2	S (4)	-	-	-0.1	11.2	3.26	100	57.6
	S (4)	**-**	**-**	**-0.2**	**8.6**	**3.27**	**100**	**58.0**
	S (4)	-	-	-0.3	6.7	3.30	100	58.3
	S (4)	-	-	-0.4	5.3	3.38	100	58.5
H3	G (≤3)	C	0	-0.2	9.8	3.15	83.7	38.8
	G (≤4)	C	0	-0.2	11.9	3.88	79.0	37.1
	G	**C**	**0**	**-0.2**	**18.8**	**7.16**	**79.3**	**35.7**
H4	**G**	**C**	**0.5**	**-0.2**	**42.2**	**6.73**	**31.5**	**37.5**
	G	C	1.0	-0.2	55.5	5.24	20.4	38.2
H5	C	C	0.5	-0.2	48.4	7.37	41.5	50.5

Table 1: Results in the simulations for the five hypotheses for three behaviours: Shallow (*max* constant given in parentheses); Greedy (where any limit on the number of contributed arguments by the agent is in brackets); and Counterfactual. Learning amounts to c in Definition 9 for μ and to the confirmation bias offset for η (where appropriate). We report RR, PR$_\mu$ and CA$_\mu$ as percentages. We indicate in bold the chosen baseline for the next hypothesis.

H4: RR increased significantly with the bias on learnt arguments ($p < 0.001$ for both comparisons of learning configurations: 0 vs 0.5 and 0.5 vs 1). However, the machine's CR and PR fell significantly ($p < 0.001$ for similar pairwise comparisons, except for 0 vs 0.5 for CR, where $p = 0.27$). highlighting the naive nature of machines learning credulously (i.e. assigning all learnt arguments the top bias).

H5: The counterfactual behaviour outperformed the greedy behaviour significantly in terms of both RR ($p < 0.01$) and CA ($p < 0.001$), showing, even in this limited setting, the advantages in taking a counterfactual view, given that the strongest argument (as selected by the greedy behaviour) may not always be the most effective in persuading.

8 Conclusions

We defined the novel concept of AXs, and deployed AXs in the XAI setting where a machine and a human engage in interactive explanations, powered by non-shallow reasoning, contributions from both agents and modelling of agents' learning and explanatory behaviour. This work opens several avenues for future work, besides those already mentioned. It would be interesting to experiment with any number of agents, besides the two that are standard in XAI, and to identify restricted cases where hypotheses H1-H5 are guaranteed to hold. It would also be interesting to accommodate mechanisms for machines to model humans, e.g. as in opponent modelling (Hadjinikolis et al. 2013). Also fruitful could be an investigation of how closely AXs can represent machine and human behaviour. Further, while we used AXs in XAI, they may be usable in various multi-agent settings.

[8] See github.com/CLArg-group/argumentative_exchanges.

[9] Experiments with greedy humans gave similar findings.

Acknowledgements

This research was partially funded by the ERC under the EU's Horizon 2020 research and innovation programme (No. 101020934, ADIX) and by J.P. Morgan and by the Royal Academy of Engineering, UK.

References

Albini, E.; Lertvittayakumjorn, P.; Rago, A.; and Toni, F. 2020. DAX: deep argumentative explanation for neural networks. *CoRR* abs/2012.05766.

Amgoud, L., and Ben-Naim, J. 2017. Evaluation of arguments in weighted bipolar graphs. In *ECSQARU 2017*, 25–35.

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

Amgoud, L., and Ben-Naim, J. 2022. Axiomatic foundations of explainability. In *IJCAI 2022*, 636–642.

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

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards artificial argumentation. *AI Magazine* 38(3):25–36.

Balog, K.; Radlinski, F.; and Arakelyan, S. 2019. Transparent, scrutable and explainable user models for personalized recommendation. In *SIGIR 2019*, 265–274.

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

Baroni, P.; Comini, G.; Rago, A.; and Toni, F. 2017. Abstract games of argumentation strategy and game-theoretical argument strength. In *PRIMA 2017*, 403–419.

Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds. 2018. *Handbook of Formal Argumentation*. College Publications.

Baroni, P.; Rago, A.; and Toni, F. 2018. How many properties do we need for gradual argumentation? In *AAAI 2018*, 1736–1743.

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

Bertrand, A.; Belloum, R.; Eagan, J. R.; and Maxwell, W. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *AIES '22*, 78–91.

Black, E., and Atkinson, K. 2011. Choosing persuasive arguments for action. In *AAMAS 2011*, 905–912.

Black, E., and Hunter, A. 2007. A generative inquiry dialogue system. In *AAMAS 2007*, 241.

Calegari, R.; Omicini, A.; Pisano, G.; and Sartor, G. 2022. Arg2P: an argumentation framework for explainable intelligent systems. *J. Log. Comput.* 32(2):369–401.

Calegari, R.; Riveret, R.; and Sartor, G. 2021. The burden of persuasion in structured argumentation. In *ICAIL 2021*, 180–184.

Cawsey, A. 1991. Generating interactive explanations. In *AAAI 1991*, 86–91.

Cayrol, C., and Lagasquie-Schiex, M. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *ECSQARU 2005*, 378–389.

Cocarascu, O.; Rago, A.; and Toni, F. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *AAMAS 2019*, 1261–1269.

Cyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A survey. In *IJCAI 2021*, 4392–4399.

Cyras, K.; Kampik, T.; and Weng, Q. 2022. Dispute trees as explanations in quantitative (bipolar) argumentation. In *ArgXAI 2022 co-located with COMMA 2022*.

de Tarlé, L. D.; Bonzon, E.; and Maudet, N. 2022. Multiagent dynamics of gradual argumentation semantics. In *AAMAS 2022*, 363–371.

Donadello, I.; Hunter, A.; Teso, S.; and Dragoni, M. 2022. Machine learning for utility prediction in argument-based computational persuasion. In *AAAI 2022*, 5592–5599.

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

Fan, X., and Toni, F. 2012a. Argumentation dialogues for two-agent conflict resolution. In *COMMA 2012*, 249–260.

Fan, X., and Toni, F. 2012b. Mechanism design for argumentation-based persuasion. In *COMMA 2012*, 322–333.

Fan, X., and Toni, F. 2015a. Mechanism design for argumentation-based information-seeking and inquiry. In *PRIMA 2015*, 519–527.

Fan, X., and Toni, F. 2015b. On computing explanations in argumentation. In *AAAI 2015*, 1496–1502.

Hadjinikolis, C.; Siantos, Y.; Modgil, S.; Black, E.; and McBurney, P. 2013. Opponent modelling in persuasion dialogues. In *IJCAI 2013*, 164–170.

Hirsch, T.; Soma, C. S.; Merced, K.; Kuo, P.; Dembe, A.; Caperton, D. D.; Atkins, D. C.; and Imel, Z. E. 2018. "It's hard to argue with a computer": Investigating psychotherapists' attitudes towards automated evaluation. In *DIS 2018*, 559–571.

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

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. Abduction-based explanations for machine learning models. In *AAAI 2019*, 1511–1519.

Kampik, T., and Cyras, K. 2022. Explaining change in quantitative bipolar argumentation. In *COMMA 2022*, 188–199.

Kontarinis, D., and Toni, F. 2015. Identifying malicious behavior in multi-party bipolar argumentation debates. In *EUMAS/AT 2015*, 267–278.

Lakkaraju, H.; Slack, D.; Chen, Y.; Tan, C.; and Singh, S. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *CoRR* abs/2202.01875.

Lertvittayakumjorn, P.; Specia, L.; and Toni, F. 2020. FIND: human-in-the-loop debugging deep text classifiers. In *EMNLP 2020*, 332–348.

Lundberg, S. M., and Lee, S. 2017. A unified approach to interpreting model predictions. In *NIPS 2017*, 4765–4774.

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

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2:175 – 220.

Panisson, A. R.; McBurney, P.; and Bordini, R. H. 2021. A computational model of argumentation schemes for multi-agent systems. *Argument Comput.* 12(3):357–395.

Paulino-Passos, G., and Toni, F. 2022. On interactive explanations as non-monotonic reasoning. In *XAI 2022 co-located with IJCAI 2022*.

Pisano, G.; Calegari, R.; Prakken, H.; and Sartor, G. 2022. Arguing about the existence of conflicts. In *COMMA 2022*, 284–295.

Potyka, N. 2018. Continuous dynamical systems for weighted bipolar argumentation. In *KR 2018*, 148–157.

Potyka, N. 2021. Interpreting neural networks as quantitative argumentation frameworks. In *AAAI 2021*, 6463–6470.

Rago, A.; Baroni, P.; and Toni, F. 2022. Explaining causal models with argumentation: the case of bi-variate reinforcement. In *KR 2022*, 505–509.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *KR 2016*, 63–73.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; and Toni, F. 2020. Argumentation as a framework for interactive explanations for recommendations. In *KR 2020*, 805–815.

Rago, A.; Cocarascu, O.; and Toni, F. 2018. Argumentation-based recommendations: Fantastic explanations and how to find them. In *IJCAI 2018*, 1949–1955.

Raymond, A.; Gunes, H.; and Prorok, A. 2020. Culture-based explainable human-agent deconfliction. In *AAMAS 2020*, 1107–1115.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A symbolic approach to explaining bayesian network classifiers. In *IJCAI 2018*, 5103–5111.

Sokol, K., and Flach, P. A. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *FAT* 2020*, 56–67.

Teso, S.; Alkan, Ö.; Stammer, W.; and Daly, E. 2023. Leveraging explanations in interactive machine learning: An overview. *Frontiers Artif. Intell.* 6.

Vassiliades, A.; Bassiliades, N.; and Patkos, T. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36:e5.

Wachter, S.; Mittelstadt, B. D.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR* abs/1711.00399.

Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; and He, L. 2022. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* 135:364–381.