Research Article

Explanation–Question–Response dialogue: An argumentative tool for explainable AI

Argument & Computation
2025, Vol. 16(2) 133–150

DOI: 10.3233/AAC-230015
journals.sagepub.com/home/arg

Federico Castagna[1], Peter McBurney[2] and Simon Parsons[3]

Abstract

Advancements and deployments of AI-based systems, especially Deep Learning-driven generative language models, have accomplished impressive results over the past few years. Nevertheless, these remarkable achievements are intertwined with a related fear that such technologies might lead to a general relinquishing of our lives's control to AIs. This concern, which also motivates the increasing interest in the eXplainable Artificial Intelligence (XAI) research field, is mostly caused by the opacity of the output of deep learning systems and the way that it is generated, which is largely obscure to laypeople. A dialectical interaction with such systems may enhance the users' understanding and build a more robust trust towards AI. Commonly employed as specific formalisms for modelling intra-agent communications, dialogue games prove to be useful tools to rely upon when dealing with user's explanation needs. The literature already offers some dialectical protocols that expressly handle explanations and their delivery. This paper fully formalises the novel Explanation–Question–Response (EQR) dialogue and its properties, whose main purpose is to provide satisfactory information (i.e., justified according to argumentative semantics) whilst ensuring a simplified protocol, in comparison with other existing approaches, for humans and artificial agents.

Keywords

Explainable AI, dialogue games, computational argumentation, dialogue protocols, large language models

1. Introduction

The surging interest towards Large Language Models (LLMs) rests upon the ambition of programming machines capable of communicating like human beings. Generally, language models (LM) tackle this challenge by studying the generative likelihood of word sequences so as to predict the probabilities of subsequent tokens in such sequences. LLMs can be identified as the last stage of a progressive development in language models research [96]: starting from statistical LMs (e.g., n-grams models and Markov chains) and moving through neural LMs (e.g., recurrent neural networks), a significative milestone has been achieved with the introduction of the pre-trained LM paradigm (mostly based on the concurrent rise of Transformer models [84]). Scaling up those pre-trained models resulted in LLMs, i.e., large(-sized pre-trained) language models displaying surprising ability in solving complex tasks [96]. The famous family of ChatGPT systems[1] is a noteworthy application of LLMs's ability to produce human-level conversations.

Nevertheless, according to the study conducted by Hinton and Wagemans [42], the output of generative large language models as GPT-3 [14] (one of the foundational models upon which ChatGPT is based) does not provide satisfactory

[1] https://chat.openai.com

[1]Department of Computer Science, Brunel University, London, United Kingdom
[2]Department of Informatics, King's College London, London, United Kingdom
[3]School of Computer Science, University of Lincoln, Lincoln, United Kingdom

Corresponding author:
Federico Castagna, Department of Computer Science, Brunel University, London, United Kingdom.
Email: federico.castagna@brunel.ac.uk

argumentative replies. The authors elaborate on such a conclusion after a thorough application of the Comprehensive Assessment Procedure for Natural Argumentation (CAPNA) protocol [41]. GPT-3 is able to produce different argument types (thus identifying common human dialectical patterns), but it fails when it comes to providing their acceptability, mostly generating fallacious arguments. The entailed consequence is that the capability of arguing, intended as an exchange of reasoning between intelligent entities, should be learnt by AIs if their purpose aims for more than just acquiring and repeating information. Overall, there is a current urge to provide clear explanations about what drives AIs decisions. This is also advocated by the technical report that OpenAI released about GPT-4 performances: *"Despite its capabilities, GPT-4 has similar limitations to earlier GPT models: it is not fully reliable (e.g. can suffer from "hallucinations"), has a limited context window, and does not learn from experience. Care should be taken when using the outputs of GPT-4, particularly in contexts where reliability is important"* [68]. In particular, the same authors encourage more investigation into AIs' explainability given their current nature as "black-box" models.[2] OpenAI researchers have already started to face the challenge of explaining single neurons within a deep learning network, however, there are several limitations that should be addressed in the future [9].

XAI is the field that studies how to improve artificial intelligence models' interpretability by analysing different tools and strategies that provide adequate explanations (i.e., that satisfy specific desirable properties). Interestingly, different works propose an account of explanations that is primarily argumentative [1,24]. Similarly, the survey of [83] concludes that using argumentation to justify why an event started, or what led to a decision, can enhance explainability. These intuitions are also backed by [60], where it is suggested that AI systems should adopt an argumentation-based approach to explanations consisting of dialogue protocols characterising the interactions between an *explainer* and an *explainee*. Such a dialectical interplay, sketched in [60] as the Explanation–Question–Response (EQR) dialogue, will be developed here into a fully-fledged formal model. Combined with LLMs or any other AI-driven model, the EQR protocol would provide an informative method to deliver deliberated explanations to end-users, while also ensuring detailed replies to follow-on queries. Indeed, the underlying computational argumentation engine supplies strong rationales that justify the given explanations.

The paper is structured as follows. Section 2 covers the background topics leveraged throughout the article, whilst Section 3 provides a detailed description of the EQR protocol which greatly extends that of [16,60], including syntax, semantics and an overview of the involved utterances, and lays the groundwork for the full formalisation (connected with computational argumentation extension-based approach) that is given in Section 4. Section 5 discusses and analyses the results achieved and compares them with related works before examining potential future lines of research. Lastly, Section 6 concludes the paper with a summary of the contents and the main findings.

1.1. Contributions

The main contribution of this study is the development of the complete formal model of EQR protocol, based on the idea from [16,60], and its theoretical analysis. Though the idea of an EQR dialogue is not new, all [60] provides is the basic concept of a dialogue that conveys an explanation through question and response, and the characterisation of that dialogue as having aspects of both persuasion, information-seeking and information-giving. The detailed characterisation of such dialogues, in Section 3, and the formalization and analysis of Section 4 are all novel. The engineering of the EQR protocol entails: (1) the creation of a new kind of dialogue that embeds multiple other types without requiring an external formal layer to switch from one protocol to another. The simplicity of its architecture is an advantage that largely favours its implementation. (2) EQR dialogues enable explanatory interactions that satisfy desirable properties of explanations. (3) Terminated EQR dialogue instantiations are proved to be sound and complete with respect to admissible extensions of Dung's Abstract Argumentation Frameworks occurring when the explainer successfully answers its interlocutor's questions. This ensures that each conveyed information is justified by a set of dialectically compelling reasons.

2. Background

In this section, we are going to outline the main theoretical subjects employed for the generation of EQR formalism and its main features. As we will see, computational argumentation and dialogue games can be combined to serve as a powerful XAI tool.

[2] Consider, indeed, the puzzling appearance of 'emergent abilities'. Such an unpredictable phenomenon consists of specific competencies that occur only in large-scale models but not in smaller ones. Thus, it is not possible to anticipate the 'emergence' of these abilities by simply analysing smaller-scale models [90]. From the perspective of developing transparent AI systems with reliable and predictable behaviour, this is certainly problematic.

2.1. Computational argumentation

A promising paradigm for modelling reasoning in the presence of conflict and uncertainty, computational argumentation has come to be increasingly central as a core study within Artificial Intelligence [6]. According to such theory, in order to determine if a piece of information is acceptable (which, roughly speaking, means that it can be treated as if it is true given the current state of an agent's beliefs), it will suffice to prove that the argument, in which the considered information is embedded, is justified (which, roughly speaking, means that it is not seriously contested) under specific semantics. Dung's abstract argumentation framework (AF) [29] is the most widely adopted formalism for determining the acceptability of arguments. In a nutshell, an argument is justified (i.e., acceptable) only if it is defended against any attacks from counterarguments.

Definition 1 (AFs, and Dung's Semantics)**.** An argumentation framework is a pair $\mathrm{AF} = \langle AR, C\rangle$ where AR is a set of arguments, and C is a binary relation on AR, i.e. $C \subseteq AR \times AR$. In addition, let $S \subseteq AR$, then:

- S is *conflict free* iff $\forall X, Y \in S$: $(X,\ Y) \notin C$;
- $X \in AR$ is acceptable w.r.t. S iff $\forall Y \in AR$ such that $(Y, X) \in C$: $\exists Z \in S$ such that $(Z, Y) \in C$.
- A conflict free set S is an *admissible* extension iff $X \in S$ implies X is acceptable w.r.t. S;
- An admissible extension S is a *complete* extension iff $\forall X \in AR$: X is acceptable w.r.t. S implies $X \in S$;
- The least complete extension (with respect to set inclusion) is called the *grounded extension*;
- A maximal complete extension (with respect to set inclusion) is called a *preferred extension*.

In the previous definition, we could describe Z as the argument that *defends* X, thus granting its acceptability. Given sufficiently large AFs, *indirect defences* might also occur. That is to say, given a sequence of acceptable arguments ending with X and starting with A, it can be the case that X (indirectly) defends A.

Definition 2 (Indirect defence and attack)**.** Let $\langle AR, C\rangle$ be an AF, and $X, A \in AR$. According to the following recursive definition, an argument X *indirectly defends* an argument A if:

a) X *defends* A;
b) X *defends* Z, and Z *indirectly defends* A.

Notice that an unattacked argument is, trivially, defended and indirectly defended by itself. A similar recursive definition can be described such that X *indirectly attacks* an argument A if:

a) X *attacks* A;
b) X *attacks* Z, and Z *indirectly defends* A.

Dung's AF can be extended such that 'preferences' can be taken into account as well. It can be useful indeed to have a way of deciding, among two or more conflicting arguments, which ones are preferred, hence, which attacks will succeed as *defeats*. This leads to the formal definition of defeats:

Definition 3 (Defeats)**.** Let $\langle AR, C\rangle$ be an AF. Then $D \subseteq C$ is the defeat relation defined by the strict partial ordering $\prec$ over AR, such that:

$$(X,\ Y) \in D \quad \text{iff} \quad (X,\ Y) \in C \quad \text{and} \quad X \not\prec Y$$

That is to say, an argument X defeats an argument Y if and only if X attacks Y and Y is not strictly preferred over X.

Overall, abstract AFs represent general frameworks capable of providing argumentative characterisations of non-monotonic logic.[3] That is to say, given a set of formulae Δ of some logical language L, AFs can be instantiated by such formulae. The conclusions of justified arguments defined by the instantiating Δ are equivalent to those obtained from Δ by the inference relation of the logic L. These instantiations paved the way for a plethora of different studies concerning the so-called 'structured' argumentation (as opposed to the abstract approach) [7,66,81].

[3] In [29], Dung employs Reiter's Default logic [71] and Pollock's Inductive Defeasible logic [69] as an example of non-monotonic reasoning rendered via abstract argumentation.

2.2. Dialogues games

The view of computation as distributed cognition and interaction [53] led to the rise of multi-agent systems, where agents are software entities with control over their own execution. This new paradigm required the design of means of communication between such intelligent agents [59]. The choice fell upon formal dialogues, due to their potential expressivity despite being still subject to specific restrictions. Dialogue games are rule-governed interactions among players (i.e., agents) that take turns in making utterances (i.e., moves) following the protocol of the game.[4]

2.2.1. Types of dialogue. Dialogue games are commonly categorized according to elements such as: what the participants know, what the participants seek to get from the dialogue, and what the dialogue rules are intended to bring about [11]. The following is an extended list (with no ambition of being exhaustive) of the standard dialogue types presented in [89]:

- *Information-Seeking*: one participant seeks the answer to some question(s) from another participant, who is believed by the first to know the answer(s). On the other hand, its complementary *Information-Giving* dialogue is characterised by a protocol that focuses on the participant who gives the response to the requesting interlocutor (e.g., [44]).
- *Inquiry*: the participants collaborate to answer some question(s) whose answers are not known to any one participant (e.g., [10]).
- *Persuasion*: one participant seeks to persuade another to accept a proposition she does not currently endorse. This can mean that the persuadee holds the opposite or is agnostic about the position put forward by the persuader. (e.g., [70]).
- *Negotiation*: the participants bargain over the division of some scarce resources. If a negotiation dialogue terminates with an agreement, then the resource has been divided in a manner acceptable to all participants. (e.g., [62]).
- *Deliberation*: the participants collaborate (hence, share the responsibility) to decide what action or course of action should be adopted in some situations. Appeals to value assumptions, such as goals and preferences, may influence the agents' deliberation (e.g., [55]).
- *Eristic*: the participants quarrel verbally as a substitute for physical fighting, aiming to vent perceived grievances (e.g., [13]).
- *Verification*: one participant seeks the answer to some question from another agent. The first participant wants to verify if the second believes that p (i.e., the proposition with which the dialogue is concerned) is true (e.g., [22]).
- *Query*: one participant always challenges the answer about p from another agent. The first participant's interest lies more on the second's arguments rather than if the second agent believes p or not (e.g., [22]).
- *Command*: One participant tells another what to do. If challenged, instructions may be justified, possibly by referencing further actions which the commanded action is intended to enable (e.g., [34]).
- *Education*: One participant wants to teach another something. Unlike information-seeking dialogues, in education dialogues the tutor, or asking agent, does know the answer to the question she is posing (i.e., she is *quizzing* the learner). (e.g., [79]).
- *Discovery*: A new idea arises out of exchanges between participants. Unlike inquiry dialogues, here the focus is on the discovery of something not previously known. The question of whose truth is to be ascertained may (or not) emerge in the course of the dialogue (e.g., [56]).

2.2.2. Dialogues combinations and control layers. In general, a dialogue game can be composed of multiple mixtures of dialogues, each of which might be of a different type. Drawing from the classification detailed in [57], we can identify the combination patterns listed in Table 1.[5] The selection and transitions between different dialogue types can be rendered via a *Control Layer* [57], defined in terms of *atomic dialogue-types* and *control dialogues*. The first element is based upon a finite set of dialogue-types. Control dialogues, instead, are dialogues that have as their discussion subjects not topics but other dialogues. They include the so-called Commencement and Termination Dialogues in charge of opening (respectively, closing) the subject dialogue, thus contributing to the management of dialogue combinations and their transitions.

[4] Notice that the literature also presents 'argument games', another dialectical formalisation which may seem similar to dialogue games in many aspects [18,65]. Nonetheless, the former can be intuitively regarded as a dialogue that an agent performs 'within itself', whereas the latter is more suited to model a public conversation that can simultaneously engage multiple agents.

[5] Similarly, Walton and Krabbe [89] studied the interaction between multiple dialogues. Their analysis resulted in an informal classification of possible dialogue shifts: (a) 'from one type to another', a sequence composed of multiple kinds of dialogues; (b) 'internal shifts', which occur within the same dialogue type without normative changes; (c) 'from one flavour to another', where the transitions concern only flavours (i.e., "[…] *secondary type of dialogue* […] *[presented] in a more subdued or less explicit form.*" [89]).

Table 1. Dialogue combinations. These are familiar from computer programming, specifically they are the operations used to model computer programs in dynamic logic.

Combination	Description
Iteration	Let D be a dialogue. The *iteration* of D to its n-fold repetition is also a dialogue, where each occurrence is undertaken until closure, and then is followed immediately by the next occurrence.
Sequencing	If D_1 and D_2 are both dialogues, then their *sequence* is also a dialogue, which consists of undertaking D_1 until its closure and then immediately undertaking D_2.
Parallelization	If D_1 and D_2 are both dialogues, then conducting them in *parallel* can be also considered as a dialogue, which consists of undertaking D_1 and D_2 simultaneously until each is closed.
Embedding	If D_1 and D_2 are both dialogues, then their *embedding* is also a dialogue, which consists of undertaking D_1 and then switching to dialogue D_2 which is undertaken until its closure, whereupon dialogue D_1 resumes immediately after the point where it was interrupted and continues until closure.

2.2.3. Dialogue components. Following the study outlined in [59], we can now summarize the three main features of formal dialogues: *syntax*, *semantics* and *pragmatics*.

Syntax. The syntax of a language prescribes instructions on how to form words, phrases and their combinations. Similarly, determining the syntax of a dialogue game involves the specification of the utterances available to the agents and the rules that govern the interactions among such utterances. In addition, it is standard to consider utterances as composed of (1) an inner layer comprising the topics of discussion and (2) an outer (or wrapper) layer comprising the locutions.

Semantics. Research concerning dialogue games is at a crossroads between multiple fields of study. Indeed, the interplay among participants in the dialogue is a form of communication that draws from human linguistics knowledge. However, the language must also be necessarily formal and interpretable by computers (an issue tackled by research such as [91]). It might then be helpful to consider different types of semantics according to the specific focus, and final deployment, of the dialogue.

(1) *Axiomatic*: defines each locution in terms of its pre and (possibly) post-conditions. Pre-conditions identify what must exist before the locution can be uttered, and post-conditions determine the consequences of such utterance. Public axiomatic approaches enable access to all conditions from each agent in the dialogue, whereas private axiomatic approaches restrict such access to a smaller subset.
(2) *Operational*: considers each locution as a computational instruction that operates successively on the states of some abstract machine. That is to say, it interprets these locutions as commands in some computer programme language.
(3) *Denotational*: assigns, for each element of the language syntax, a relationship to an abstract mathematical entity, its denotation.[6]

While the dialogue unfolds, agents usually incur *commitments*. That is to say, a speaker asserting the truth of a statement, may be committed to justifying such statement (even if it does not correspond to their real beliefs) against opponents' challenges or retract its assertion. The commitments of all the agents are then tracked and stored in a public database, called a *commitment store*. This position adopts Hamblin's understanding of commitments as purely dialectical obligations [37]. Walton and Krabbe consider instead commitments as obligations connected to a course of action that subsumes under this paradigm also dialectical commitments: *"*[…] *whose partial strategies assign dialogical actions that center on one proposition"* [89]. On the other hand, Singh [78] and Colombetti [23] regard commitments as *social*, i.e., expressions of wider interpersonal, social, business or legal relationships between the participants, and utterances in a dialogue are a means by which these relationships may be manipulated or modified.

Pragmatics. Pragmatics deals with those aspects of the language that do not involve considerations about truth and falsity. Such aspects usually include the *illocutionary force* of the utterances along with *speech acts*, i.e., non-propositional utterances intended to or perceived to change the state of the world.[7] More precisely, drawing from the analysis of [35] based on relevant literature on the topic, such as Austins and Searle's works [3,75,76], we can define speech acts as 'verbal actions' that accomplish something. Locution would correspond to the simple performance of an utterance, whereas

[6] The possible worlds (i.e., Kripkean) semantics is an example of denotational semantics for a logical language [48].
[7] An example of analysis of the different pragmatical meanings existing between, say, 'commands' and 'promises' can be found in [58]. Furthermore, the work presented in [61] introduces a specific syntax that accounts for the pragmatical uptake and revocation of utterances over actions.

illocution would be the actual intention of the speaker behind the locution meaning. For example, the sentence "You're standing on my foot" uttered in a crowded place is a statement (locution) with the illocutionary force of a command (that is to say, the real meaning is the imperative "move away").

2.2.4. Burden of proof. One last important aspect considered by the dialogue literature regards the so-called *'burden of proof'*. Multiple authors have investigated the matter and proposed different definitions. For instance, according to Walton [88], the burden of proof is *"an allocation made in reasoned dialogue which sets a strength (weight) of argument required by one side to reasonably persuade the other side."*, whereas van Eemeren and Grootendorst [30] described it as occurring when *"a party that advances the* [dialogue] *standpoint is obliged to defend it if the other party asks him to do so"*. In general, we could say that participants in a dialogue incur a burden of proof when declaring a proposition as their thesis, thereby compelling them to offer evidence or backing when such a thesis is challenged. In an evenly matched dispute, where the plausibility of the participants' thesis is balanced, any new argument moved may tilt the burden of proof. Nevertheless, in some specific circumstances, the burden of proof can be much heavier on one particular side. As an example, consider any criminal trial: the prosecutor must prove guilt "beyond reasonable doubt" to win her case, which means that she bears a greater encumbrance than her counterpart who does not have to show that their client did not commit the crime, but merely that there is reasonable doubt that their client did so. The notion of burden of proof may be considered as a dialectical obligation which is 'stronger' than the previously examined ('weaker') *commitments*. Indeed, while the latter always occurs in a dialogue, this is not the case for the former, as Walton concluded *"If there is no thesis to be proved or cast into doubt* […]*, there is no burden of proof in that dialogue"* [86].

While computational arguments can be embedded in any dialogue game's inner syntactic layer to handle the current topic, argument semantics can justify the rationale behind each utterance. On the other hand, leveraging dialogue game protocols and their combinations allow for the generation of efficient strategies on a variety of subjects, which proves to be a useful feature to exploit for the eXplainable AI (XAI) research field.

2.3. Explainable AI

The design and implementation of tools capable of enhancing artificial intelligence models' interpretability, thus addressing the well-known opacity of their black-box algorithms, constitutes the core focus of XAI. The underlying idea revolves around the possibility of generating exhaustive explanations disclosing salient information about systems operations [5]. Noticeably, Article 22 of the General Data Protection Regulation (GDPR)[8] introduces the right to obtain an explanation of the inferences produced by automated decision-making models. The necessity to abide by this new regulation contributes to making explainability a current hot topic in the AI research landscape. Nevertheless, there is still no consensus on a unique definition of explanation [85], especially because most scholars seem to be influenced by their subjective intuitions of what an XAI approach should entail. To help clarify essential aspects of explanations by drawing from social science studies, Miller [63] identified contextuality (which is the product of merging different explanatory features, e.g., selectivity and causality of information) as the most relevant factor. Similarly, Bex and Walton [8] view explanations as speech acts used to help understand something. Gunning et al. [36] focus instead on pinpointing the current main issues regarding XAI and present them in a list that includes challenges such as: accuracy vs interpretability, the use of abstractions to simplify explanations or preference of competencies over decisions as core elements of information delivery. Another problem is related to the requirement of tailoring explanations to the end-user who is interacting with the system. From this perspective, consider that explanations can also be provided in a dialogical form: given an initial reason, additional information and answers to follow-on queries may be delivered while the dialectical interaction unfolds. This enables a collaborative process where the explainer is capable of determining what information it is that the user wants (i.e., tailored to the user's needs). Furthermore, a study from Lakkaraju et al. argues that decision-makers largely prefer interactive (dialectical) explanations such that: *"natural language dialogues for explainability could enhance the [AI] model's understanding with greater ease than current one-off explanations"* [50]. The EQR protocol herein presented will follow precisely this explanatory strategy.

2.3.1. Desirable properties of explanations. A plethora of research from different scholars has studied the general properties that effective explanations should fulfil. For simplicity, we will single out the most intuitive (and, we claim, more desirable) of such features:

[8] Regulation (EU) 2016/679

- [**Exhaustivity**] A convincing explanation should cover every aspect of the explained procedure: what the AI system has done, what it is doing now and what it is going to do next [5]. Additional formulation of the exhaustivity principle can be found in the recent literature [12,33] and in [49] where it is denoted as 'completeness'.
- [**Selectivity**] A compelling explanation should present only the causes that are relevant (e.g., necessary and sufficient) to the explanatory purpose [39,40,51,64,82].
- [**Transfer of Understanding**] A successful explanation should be able to transfer understanding from one party to another, where understanding is intended as 'common knowledge' shared between those parties [8,74]. This may also involve the *learning process* [52,92].
- [**Contextuality**] An effective explanation should account for the circumstances in which it occurs and the end-users to whom it is addressed[9] [28,36,38,63,67,80].

As we are going to see, the dialogical model fleshed out in the following sections enables explanatory interactions that satisfy all of the above properties. Notice that we also selected such features due to their extensive scope, which makes them suited for any kind of explanation. Nevertheless, different researchers may prefer to focus on more specific aspects of the XAI procedure (e.g., post-hoc explanations for AI-assisted human decisions [45], principles of interactive explanations via natural language interactions [50], etc.), thus identifying diverse (and narrower) properties from the one we defined.

3. Explanation–Question–Response (EQR) dialogue

We have already mentioned how evaluation and assessment of explanations are particularly suited to be modelled as dialogical interactions between an *explainer*, i.e., an agent capable and willing to answer questions concerning the explanation and an *explainee*, i.e., an agent seeking to determine the validity of such answers. The research conducted in [16,60] suggest that a *new type* of dialogue denoted as Explanation–Question–Response (EQR) might be helpful for explanation, and sketched it as being halfway between a persuasion, an information-giving/seeking and a query dialogue, without the need for any complicated shifting formalism (as the Control Layer) that would account for different simultaneous discussions taking place. As such, the EQR dialogue is engineered to provide a simple and efficient way to capture multiple kinds of dialectical interactions that might occur when the topic revolves around the explanation of an issue. In the following sections, we are going to comprehensively describe one possible Explanation–Question–Response dialogue that fits the high level description from [60], detail a formal account of this dialogue, along with a protocol, and prove that it provides explanations in a specific, technical, sense.

3.1. EQR dialogue syntax

Each utterance of an EQR dialogue presents two syntactic layers: (i) an innermost layer in which the contents of the utterances are expressed in a formal way through propositional logic; (ii) an outermost layer which expresses the locutionary force of the single utterances. We, therefore, denote as 'arguments' the components of the former (that, indeed, can be rendered as computational arguments as per Section 2.1), whereas the outermost wrapping layer can be represented by listing all the possible locutions of the dialogue as detailed in the first column of Table 2 and Table 3. In particular, Table 2 depicts a series of structural locutions (such as **change player** and **enter/leave dialogue**), which roughly emulate the termination function of the Control Layer, thus intuitively justifying the reason why the Control Layer is not required.

Resolution of conflicts. While the dialogue unfolds, different kinds of conflicts may occur, each of which entails different possible resolutions. Such conflicts depend upon a specific class of dialectical attacks listed as [2]:

- **Attacks concerning factual disagreement** These kinds of attacks involve the nature of the current state of the world (including causal relations).
- **Attacks concerning representation** These kinds of attacks involve issues related to the language and logic being used.
- **Attacks concerning different preferences** These kinds of attacks involve the different ranking of the players' preferences.
- **Attacks concerning clarification of a position** These kinds of attacks involve questioning a specific position of the contender.

[9] The importance of providing a context is also emphasised in studies concerning robot failure explanations: it improves the resolution and identification of shortcomings and increases users' trust [26,72].

Table 2. Locutions to control the dialogue.

Locution	Pre-conditions	Post-conditions
Enter dialogue	– Speaker has not already uttered enter dialogue	– Speaker has entered dialogue
Leave dialogue	– Speaker has uttered enter dialogue	– Speaker has left dialogue
Turn start	– Speaker has not already made their move	– Speaker has started their turn
Turn finished	– Speaker has started their turn – Speaker has finished making their move	– Speaker has finished their turn
Concede	– Hearer has made an attack **or** – Hearer has asked a question on an element of speaker's position **or** – Hearer has answered a question asked by the speaker	– Speaker committed to the negation of the element that was denied by the hearer **or** – Speaker does not know the answer **or** – Speaker committed to the statement given as a response by the hearer
Reject	– Hearer has made an attack **or** – Hearer has answered a question asked by the speaker	– Disagreement reached
Change player	– Speaker has uttered turn finished	– Speaker and hearer switch roles so new speaker can now make a move

Table 3. Locutions to unfold the dialogue.

Locution	Pre-conditions	Post-conditions
State 'something'	– Speaker has uttered enter dialogue – Speaker not committed to the 'something'	– Speaker committed to the stated 'something'
Ask 'something'	– Speaker has uttered enter dialogue – Hearer has uttered enter dialogue – Speaker not committed to the asked 'something'	– Hearer must reply with state 'something' **or** – Hearer doesn't know the asked 'something'
Deny 'something'	– Speaker has uttered enter dialogue – Hearer has uttered enter dialogue – Hearer committed to the denied 'something'	– Speaker committed to deny 'something'

Assuming that every participant of the EQR dialogue pre-emptively agrees on the involved ontology (i.e., state of the world, language and logic employed), it is then possible to identify two forms of resolution: (a) value-preferred defeat or (b) rational disagreement. Both types of resolution require a means for evaluating defeats according to the ranked-value order of the players. For this purpose, it is thus possible to employ any computational argumentation theory capable of handling defeats. The rational disagreement is then formalised via the generation of two different (and conflicting) admissible extensions, each of which is related to the preference of one (team of) player.

3.2. EQR dialogue semantics

An axiomatic semantics for the EQR dialogue presents the pre-conditions necessary for the legal utterance of each locution under the protocol, and any post-conditions arising from their legal utterance (second and third columns of Table 2 and Table 3). Such pre and post-conditions influence the *commitment store* of each agent participating in the dialogue. These commitment stores are intended according to Hamblin's definition [37], i.e., public statements that the agents have to defend in the dialogue (unless withdrawn), but they might not correspond to the agent's real beliefs or intentions.

3.3. Turn structure and winning conditions

Having in mind the semantic pre and post-conditions of each locution of the EQR dialogue protocol, we can informally identify the ordered sequence of locutions that distinguishes every turn by a player. We can determine two parties of agents (which can also be composed of one element each) playing the dialogue: the proponent team (PRO), i.e., the *explainer* agent/s that advance(s) the explanation (say, X) which is willing to back with additional information; the opponent team (OPP), i.e., the *explainee* agent/s trying instead to challenge the proponent statement. The goal of OPP is to successfully attack X, the initial argument moved by PRO which, in turn, has to counter every such attack via elucidating replies. Notice that the purpose of the explainee is to retrieve data and understand the rationale behind the received explanation X (recall

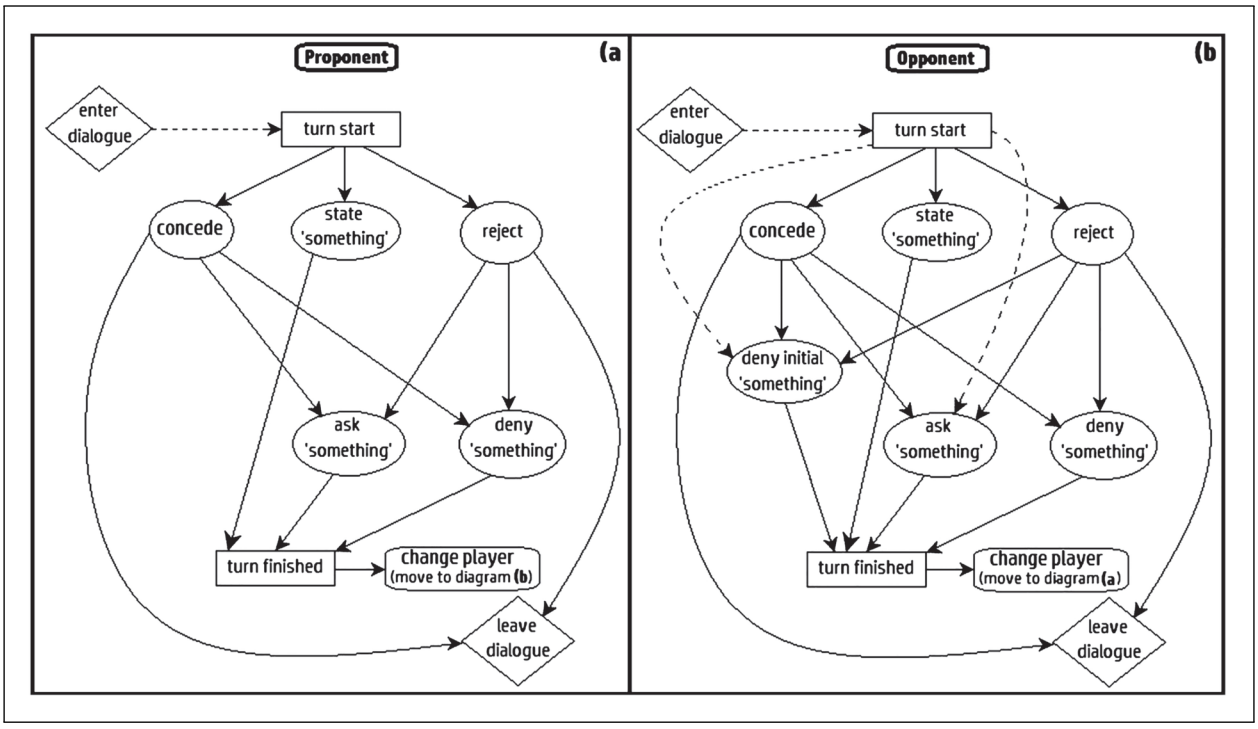

Figure 1. Ordered sequences of locutions describing the turns of each player. The dashed arrows denote moves that must be performed during the first turn of each participant only (e.g., in all the subsequent turns, the player will start from the locution **turn start** rather than **enter dialogue**). Notice also that the opponent will always prefer to utter a **deny initial 'something'** rather than a **deny 'something'** or **ask 'something'**. This preference is emphasized in the graph by the different positions of the locutions.

that an EQR dialogue is a mixture of persuasion, information-giving/seeking[10] and query dialogues) rather than suggesting its own view on the subject. The ordered sequence of locutions can then be depicted as in Figure 1 and summarized as follows:

- PRO is the first to play. Its first turn will consist only in the utterance of the locution **state 'something'**, where 'something' corresponds to the initially posited explanation X (which can also be trivial).
- OPP is the second to play. Its first turn is characterised only by the utterance of either **deny initial 'something'** or **ask 'something'**.[11] Obviously, the *explainee* cannot immediately **concede** X: if that was the case, there would have been no need for a dialogue to occur in the first place.
- A turn can finish only after:
  - (*PRO's turn*) an attack has been moved, a question has been asked or a statement has been uttered. That is to say, if a locution **deny 'something'**, **ask 'something'** (where 'something' refers to the target of the attack or the query) or **state 'something'** (where 'something' comprises all the information needed to answer a question) is uttered by the team of players. Recall that PRO's task is to counter every challenge against the initial argument X, as such, *it must* utter locutions that serve this purpose.
  - (*OPP's turn*) an attack has been moved, a question has been asked or a statement has been uttered. That is to say, if a locution **deny 'something'**, **deny initial 'something'**, **ask 'something'** (where 'something' refers to the target of the attack or the query) or **state 'something'** (where 'something' comprises all the information needed to answer a question) is uttered by the team of players. Recall that OPP's task is to challenge the initial argument X moved by PRO, as such, *it must* utter locutions that serve this purpose.

[10] Such that the explainer gives information and the explainee seeks information.
[11] Observe that **deny initial 'something'** identifies an attack that directly targets the initial argument X moved by PRO.

- The team to whom the attack, question or response of the previous turn was addressed must begin its current turn with the locution **concede**, **reject** (if the party is, respectively, accepting the result of an attack, accepting the response to its previous query, claiming it does not know an answer, or disagreeing with the result of an attack or a response) or the locution **state 'something'** (if the party is answering the 'something' asked by the other team's previous question). Notice that OPP is the only one *not obliged* to counter the previous move from PRO: even if it could do otherwise, it can simply concede it (uttering **concede**) and move another attack or ask another question.
- No player can perform more than one locution per turn except for the ones designed to control the dialogue (Table 2).
- **enter dialogue** (as well as **leave dialogue**) is a locution performed one-time only throughout the whole dialogue.

Winning conditions. In an EQR dialogue, where we need to assess the reliability of an explanation, the 'burden of proof' lies on PRO, the *explainer*. Indeed, it is the proponent who needs to show the validity of its initial argument and persuade its contender via compelling reasons, while for OPP it suffices to successfully attack or question it such that PRO cannot respond with other than a **concede** locution. We can then informally define the winning condition of the two teams of players as:

(PRO) *The proponent wins if the opponent leaves the dialogue*.[12] PRO has countered every possible attack/answered every possible question moved by the contender party which is now persuaded about the validity of the initial argument X. This means that OPP has uttered the locution **concede** before moving the locution **leave dialogue**.

(OPP) *The opponent wins if the proponent leaves the dialogue*. OPP successfully attacks/inquires the initial argument X of the contender party raising at least one doubt about its validity. This means that PRO has uttered the locution **concede** before moving the locution **leave dialogue**.

(Draw) The utterance of **reject** before **leave dialogue** (by either contender) implies that the last arguments of each player have the same level of preference. As a result, this might create two different and conflicting admissible extensions (according to Dung's semantics [29]). Such extensions would represent the *rational disagreement* reached by the two parties of agents and their respective positions.[13]

4. EQR formal protocol

In this section, we provide a formal description of an EQR protocol, and prove some of its properties.

4.1. Protocol definition

Before defining the protocol, it may be useful to briefly outline the role of each locution and determine their functions within the EQR dialogue frame. In particular, most of them can be rendered as challenges towards other uttered arguments.

ask This locution aims at asking the ground on which it is the case that 'something' (say a) and not otherwise. As such, it can be seen as an argument attacking another argument on a.

state Except with PRO's first turn (in which **state** corresponds just with the presentation of the initial explanation), the purpose of the locution is to answer the query moved by the **ask** locution. This will establish the reason why the questioned 'something' is the case. **state** can then be seen as identifying such rationale via an argument attacking the argument that posed the previous question.

deny Intuitively, this locution denotes a refutation against the 'something' (say a) it is addressed. As such, it can be straightforwardly seen as an argument attacking another argument on a. The same reasoning can be applied to **deny initial** (which is merely a locution *deny* that directly targets the initial argument X moved by PRO).

The role of locutions such as **enter/leave dialogue**, **turn start/finished**, **concede**, **reject** and **change player** is, instead, to administer the dialogue in a more 'structural' way that is not directly connected to challenges among arguments. Their purpose is to identify specific phases of each turn and different stages of the overall dialogue.

[12] A similar idea of winning conditions is proposed by Krabbe: "[…] whosoever abandons a chain of arguments has lost that chain of arguments, and that who loses the last chain of arguments, loses the discussion as a whole" [47].
[13] Consider that a draw can be solved by an additional inquiry dialogue to adjust the preference ordering between players.

Remark 1 (Notation)**.** In the next sections, we are going to employ the following notation. The dialogue locutions and their conveyed arguments will be formally denoted by Pl [locution] : argument, where Pl represents the team of players (either P or O, abbreviations for PRO and OPP) that utters them. (Pl_n)turn identifies, instead, the set of locutions and conveyed arguments uttered by either team as the nth-move of the unfolding dialogue.[14] In addition, we are going to use $\mathcal{A}$, $\mathcal{A}'$, $\mathcal{A}''$, etc. as variables denoting arbitrary arguments.

Finally, $\mathcal{A}^\star = \mathcal{A}_1 - \mathcal{A}_2 - \cdots - \mathcal{A}_i$ designates a sequence of alternating PRO and OPP arguments (each of which is part of a locution, i.e., Pl [locution] : argument) ending with argument $\mathcal{A}_i$ (where $i = 2k + 1$ for $k \geqslant 0$). The sequence is such that $\mathcal{A}_1 - \mathcal{A}_2$ identifies the conflict of information occurring from argument $\mathcal{A}_1$, moved by one team, to $\mathcal{A}_2$, moved by the other team. Writing $(\mathcal{A}, \mathcal{A}^\star) \in \mathcal{C}$ means that $\mathcal{A}$ (attacks or) indirectly attacks $\mathcal{A}_i$, i.e., the last argument of the sequence $\mathcal{A}^\star$.

To clarify, let us consider a brief conversation as an example, and label each utterance with the corresponding notation of Remark 1. For simplicity, we are omitting the control locutions of Table 2.

Example 1. The following natural language interaction can be thought of as a trivial instance of an EQR dialogue where the explainer (PRO) tries to justify their initial argument about the weather forecast.

P[**state**]:"It is windy and dark clouds are coming. This means it is going to rain soon"
O[**ask**]:"Why do you say it is going to rain, given that it is actually sunny and not that windy?"
P[**state**]:"I heard it on the forecast of Channel 717"
O[**deny**]:"Channel 717 is a very unreliable source of information"
P[**ask**]:"Even for the weather forecast?"
O[**state**]:"Especially for the weather forecast…"

Suppose the proponent decides to believe the opponent's statement about Channel 717 (hence uttering the **concede** locution). Given that PRO has now resigned from her initial argument, OPP can be deemed to have successfully defeated its counterpart's argument, thus winning this particular dispute.

We now have all the elements to formally introduce the dialogue protocol. Similar to a list of instructions, this protocol determines the legal moves that can be performed by the participants. The conversation unfolds as a result of the legal arguments uttered and terminates when there are no more valid moves available. When this happens, the status of the initial explanation X is evaluated.

Definition 4 (Protocol)**.** Let Arg be the initial explanation (i.e., the argument we previously referred to as X). Let us also make use of the same notation as described in Remark 1. Then, an EQR dialogue unfolds in the following way:

(4.0) PRO moves the first set of locutions, i.e., (P_1)turn:

(a) P[**enter dialogue**], P[**turn start**]
(b) P[**state**] : Arg
(c) P[**turn finished**], P[**change player**]

(4.1) OPP moves the second set of locutions, i.e., (O_2)turn:

(a) O[**enter dialogue**], O[**turn start**]
(b) O[**deny initial**] : $\mathcal{A}'$ or O[**ask**] : $\mathcal{A}'$, where $(\mathcal{A}', Arg) \in \mathcal{C}$
(c) O[**turn finished**], O[**change player**]

(4.2) If (Pl_n)turn and $n = 2k + 1$ (for $k > 0$), then it is PRO's turn to move, i.e., (P_n)turn. Otherwise, if (Pl_n)turn and $n = 2k$ (for $k > 1$), then it is OPP's turn to move, i.e., (O_n)turn. That is to say, PRO moves on odd turns, while OPP moves on even turns.

(4.3) A generic PRO's turn, i.e., (P_n)turn, is such that (in order):

(a) P[**turn start**]

[14] That is to say, (P_n)turn and (O_n)turn identify the set of locutions and conveyed arguments moved by PRO, respectively OPP, as the nth-move of the unfolding dialogue.

(*b*) PRO chooses one among:

* P[**concede**] : $\mathcal{A}'$, if $\mathcal{A}' \in (O_{n-1})\texttt{turn}$ and corresponds either to O[**deny**] : $\mathcal{A}'$, O[**ask**] : $\mathcal{A}'$, or O[**state**] : $\mathcal{A}'$
* P[**reject**] : $\mathcal{A}'$, if $\mathcal{A}' \in (O_{n-1})\texttt{turn}$ and corresponds either to O[**deny**] : $\mathcal{A}'$ or O[**state**] : $\mathcal{A}'$. Alternatively, O[**deny initial**] : $\mathcal{A}'$ if $(\mathcal{A}', Arg) \in \mathcal{C}$
* P[**state**] : $\mathcal{A}''$, if $(\mathcal{A}'', \mathcal{A}') \in \mathcal{C}$, $\mathcal{A}' \in (O_{n-1})\texttt{turn}$ and corresponds to O[**ask**] : $\mathcal{A}'$

(*c*) According to the choice of point (4.3(*b*)), PRO selects one among:

⋄ P[**deny**] : $\mathcal{A}''$, or P[**ask**] : $\mathcal{A}''$, where $(\mathcal{A}'', \mathcal{A}''') \in \mathcal{C}$, and $\mathcal{A}''' \in (O_i)\texttt{turn}$ (for $i < n$) if uttered after a P[**concede**] : $\mathcal{A}'$, such that $\mathcal{A}''' \neq \mathcal{A}'$

⋄ P[**deny**] : $\mathcal{A}''$, or P[**ask**] : $\mathcal{A}''$, where $(\mathcal{A}'', \mathcal{A}') \in \mathcal{C}$ if uttered after a P[**reject**] : $\mathcal{A}'$

(*d*) No unattacked argument $\mathcal{A}$ moved in the dialogue is such that $(\mathcal{A}, \mathcal{A}^\star) \in \mathcal{C}$ and Arg is the last argument in $\mathcal{A}^\star$.

(*e*) P[**turn finished**], P[**change player**]

(4.4) A generic OPP's turn, i.e., $(O_n)\texttt{turn}$, is such that (in order):

(*a*) O[**turn start**]

(*b*) OPP chooses one among:

* O[**concede**] : $\mathcal{A}'$, if $\mathcal{A}' \in (P_{n-1})\texttt{turn}$, $\mathcal{A}' \neq Arg$ and corresponds either to P[**deny**] : $\mathcal{A}'$, P[**ask**] : $\mathcal{A}'$, or P[**state**] : $\mathcal{A}'$
* O[**reject**] : $\mathcal{A}'$, if $\mathcal{A}' \in (P_{n-1})\texttt{turn}$ and corresponds either to P[**deny**] : $\mathcal{A}'$ or P[**state**] : $\mathcal{A}'$
* O[**state**] : $\mathcal{A}''$, if $(\mathcal{A}'', \mathcal{A}') \in \mathcal{C}$, $\mathcal{A}' \in (P_{n-1})\texttt{turn}$ and corresponds to P[**ask**] : $\mathcal{A}'$

(*c*) According to the choice of point (4.4(*b*)), OPP selects one among:

⋄ O[**deny**] : $\mathcal{A}''$, or O[**ask**] : $\mathcal{A}''$, where $(\mathcal{A}'', \mathcal{A}''') \in \mathcal{C}$ and $\mathcal{A}''' \in (P_i)\texttt{turn}$ (for $i < n$), if uttered after a O[**concede**] : $\mathcal{A}'$, such that $\mathcal{A}''' \neq \mathcal{A}'$

⋄ O[**deny**] : $\mathcal{A}''$, or O[**ask**] : $\mathcal{A}''$, where $(\mathcal{A}'', \mathcal{A}') \in \mathcal{C}$ if uttered after a O[**reject**] : $\mathcal{A}'$

⋄ O[**deny initial**] : $\mathcal{A}''$ such that $(\mathcal{A}'', Arg) \in \mathcal{C}$ if uttered after a O[**reject**] : Arg or O[**concede**] : $\mathcal{A}'$

(*d*) There exists an unattacked argument $\mathcal{A}$ moved in the dialogue such that $(\mathcal{A}, \mathcal{A}^\star) \in \mathcal{C}$ and Arg is the last argument in $\mathcal{A}^\star$.

(*e*) Every argument $\mathcal{A}''$ of points (4.4(*b*)) and (4.4(*c*)) is such that it has not already been moved and attacked (and not defended) in $(O_i)\texttt{turn}$ (for $i < n$)

(*f*) O[**turn finished**], O[**change player**]

(4.5) The turn of the first team having no more locutions [**deny**], [**deny initial**], [**ask**] or [**state**] available, i.e., $(Pl_n)\texttt{turn}$, is such that it overwrites any other previous move requirements and (in order):

(*a*) Pl[**turn start**]

(*b*) Pl[**concede**] : $\mathcal{A}'$ or Pl[**reject**] : $\mathcal{A}'$, where $\mathcal{A}' \in (Pl_{n-1})\texttt{turn}$ and corresponds either to Pl[**deny**] : $\mathcal{A}'$, Pl[**state**] : $\mathcal{A}'$ or Pl[**ask**] : $\mathcal{A}'$ (alternatively, it corresponds to O[**deny initial**] : $\mathcal{A}'$, if $(Pl_{n-1})\texttt{turn} = (O_{n-1})\texttt{turn}$)

(*c*) Pl[**leave dialogue**]

Definition 4 formalises the moves available to each team of participants during an *EQR dialogue*. PRO starts the game and utters a specific set of locutions [(4.0)], after which it will be OPP's turn to move [(4.1)]. Then, the two teams will alternate, uttering ordered lists of locutions in accordance with the previous moves and phases of the dialogue [(4.3), (4.4)]. Notice that both PRO and OPP will have to abide by the respective *relevance conditions*. These are rules that force the teams to change the (temporary) outcome of the game to their advantage, thus avoiding any detour. That is to say, at the end of its turn, PRO must have (directly or indirectly) defended the initial argument Arg [(4.3(*d*))]. On the other hand, at the end of its turn, OPP must have (directly or indirectly) challenged the validity of Arg [(4.4(*c*)), (4.4(*d*))]. Finally, consider also the importance of the *non-repetition rule* [(4.4(*e*))]: without this constraint, OPP will be allowed to repeat the same attacks that have already been countered by PRO, generating, in this way, an infinite dialogue.

The conclusion of the dialogue occurs whenever one of the two teams terminates its available move [(4.5)]. At this point, PRO will be proclaimed the winner if OPP leaves the dialogue first by uttering the locutions **concede** and **leave**

dialogue. The opposite circumstance will instead grant the victory to OPP. Otherwise, the result will be a draw. Recall that a victory means being able to assess the validity of *Arg* (for PRO) or to dialectically show its invalidity (for OPP).

4.2. Soundness and completeness with respect to Dung's AFs

The termination of the dialogue leaves us with a number of arguments that, along with the existing attacks, can be semantically evaluated. Since those arguments can be considered as being members of an AF (as defined in Section 2.1), we can thus show the following proof:

Theorem 1 (Soundness and Completeness)**.** *Let* Arg *be the argument of an AF. A terminated EQR dialogue starting with* Arg *is won or tied by PRO iff* Arg *is a member of an admissible extension* Adm *of the AF.*

Proof. (Sketched). Let *Arg* be the member of a set $\mathcal{S}$, i.e., a subset of arguments moved by PRO in the terminated dialogue. The theorem can be proven if we show that: ($\Rightarrow$) $\mathcal{S}$ corresponds to an admissible extension and that ($\Leftarrow$) it is possible for PRO to win or tie an EQR dialogue starting with *Arg* by employing only arguments from Adm.

($\Rightarrow$) By Definition 4 (in particular (4.3(*b*–*d*))), *Arg* is defended by PRO's arguments, which, in turn, are indirectly defended by undefeated PRO's arguments. Assume that the set $\mathcal{S}$ is composed exactly of those moves and *Arg*, whilst also being conflict free (otherwise, OPP would have posited the conflicting argument against $\mathcal{S}$, preventing the inferred PRO's victory or draw). However, these are the same features that identify an admissible extension, as stated in Definition 1.

($\Leftarrow$) Suppose that $Arg \in$ Adm. We can play an EQR dialogue starting with *Arg* and following the protocol of Definition 4. At each challenge moved by OPP, we can respond with a locution that conveys an argument member of Adm. Since Adm is admissible, the dialogue can terminate only with the victory of PRO or a tie between the contenders. □

Theorem 1 proves the semantic connection that exists between EQR dialogues and computational argumentation which provides an additional formalism to establish the rationale behind PRO's initial explanation and the subsequent arguments defending it. Furthermore, such an equivalency allows us to evaluate the EQR dialogue moves using any proof theory, algorithmic procedures, or methodologies semantically associated with computational argumentation. In other words, because the protocol is defined in terms of high-level argumentation concepts like Dung's notions of attacks and admissibility, Theorem 1 will hold for any argumentation representation that respects these notions. Thus, for example, were we to implement agents equipped with knowledge in the form of ASPIC$^+$ statements [66], construct arguments and identify attacks using ASPIC$^+$, then an EQR dialogue between the agents would be sound and complete in the sense of Theorem 1.

4.3. EQR dialogue and desirable properties of explanation

Having fleshed out the EQR formal protocol, we can now illustrate how an interaction ensuing from the unfolding of such a dialogue results in an explanatory interplay that satisfies the desirable properties outlined in Section 2.3.1, as demonstrated in the following:

Theorem 2. *Any instantiation of the EQR dialogue protocol characterised by the victory of PRO enjoys the exhaustivity, selectivity, transfer of understanding and contextuality properties of explanation.*

Proof. (Sketched).

- **[Exhaustivity]** Since the explainer (PRO) replies to each question, doubt or rebuttal that the explainee (OPP) may have, covering every aspect of the requested explanations, any explanatory process resulting from the unfolding of an EQR dialogue that terminates with PRO's victory, enjoy the exhaustivity property.
- **[Selectivity]** Theorem 1 shows how a victory (or a draw) from PRO in an EQR dialogue generates an admissible set of arguments defending the initial explanation *X*. Intuitively, these acceptable arguments provide the necessary and sufficient rationale for the justification of the explanatory process. Indeed, they constitute the minimum amount of information required to validate *X*.
- **[Transfer of Understanding]** As indicated in Section 3.1, the participant of an EQR dialogue pre-emptively agrees on the ontology that will characterise the interaction, thus sharing a joint comprehension (i.e., a 'common knowledge') of the topic and the capability of transferring such understanding.
- **[Contextuality]** The EQR protocol provides the explainee with legal moves that allow them to freely investigate the initial explanation. This flexibility, when the EQR dialogue terminates with PRO's victory, guarantees the delivery of information specifically tailored to the explainee's needs and background. □

This result emphasises the suitability of an EQR dialogue as a formal tool capable of conveying relevant (selectivity), user-friendly (contextuality) explanations based on a shared comprehension (transfer of understanding) whilst covering each significant aspect of the procedure (exhaustivity).

5. Related and future work

Related work. In the literature, there are a number of other explanation protocols which have some similarity with the EQR dialogue protocol. One of the older examples may be found in the works of Walton [87] and the subsequent joint research from Bex and Walton [8] where the authors design a dialogue and detail a complete list of its locutions. However, to evaluate the explanation that this protocol provides, the explainee needs to resort to a different dialogue protocol (denoted an "examination"). A similar, multi-protocol, approach is adopted by Madumal et al. [54]. They devise a study for modelling explanation dialogues by following a data-driven approach in which the resulting formalisation embeds (possibly several) argumentation dialogues nested in the outer layer of the explanation protocol. The dialogue structure proposed by Sassoon et al. [73] in the context of explanations for wellness consultation also exploits multiple dialogue types (e.g., *persuasion*, *deliberation* and *information seeking*) and their respective protocols whilst mostly focusing on the course of action to undertake. All of these approaches differ from the EQR dialogue, both in the sense that the EQR protocol is partway between *persuasion, information-giving/seeking and query*, and also in the sense that we believe the EQR protocol more comprehensively incorporates locutions for handling each of these tasks without the need for adopting a Control Layer (Section 2.2.2) or switching between protocols. This allows for a simpler formalisation and ensures a closer approximation to real-world dialogues. Less directly related to EQR is the formalism proposed by Dennis and Oren [27], where another theoretical analysis of explanations is conducted. Here the interactions focus on BDI agents and the paper outlines properties strictly related to the introduced protocol rather than those that are scalable to general explanations like the features enjoyed by EQR.

Other relevant work on argumentation-based explanations is provided by the studies of Fan and Toni [32], and Shams et al. [77]. These mainly focus on explanations whose justification is rendered through argumentation semantics via specific dialogue formalisations, and there is thus an equivalence between the dialogues and the argumentation semantics. The EQR protocol enjoys a similar equivalence with argumentation semantics (Theorem 1) whilst also generating explanations that satisfy the aforementioned properties (Section 2.3.1). Another interesting approach consists of the formal protocol presented by Buisman [15], who outlines a dialogue hinging on the delivery of tailored explanations to target audiences. To ensure personalised interactions, a variety of purposedly designed locutions is added to the allowed moves list. Adopting a different method, with the EQR formalisation, we tried instead to create a simple protocol: we argue that few locutions suffice to unfold the dialogue and provide appropriate explanations for diverse end-users thanks to the selection of the conveyed argument. Finally, we note that there are similarities between the work of [79], on education dialogues, and EQR. These similarities are largely in terms of the intuitions behind both formalisms, since the two approaches differ in substantial respects. [79] presents three variants characterising the whole spectrum of possible interactions between a tutor and a learner: the tutor either (a) quizzes or (b) refines her perception about the learner, whereas the learner asks a clarifying question to the tutor (c1) or another learner (c2). Given these categories, we could consider (c1) as an instance of an *information-seeking* (respectively *information-giving*) protocol, while (c2) would better depict an *inquiry* kind of dialogue. On the other hand, (b) mirrors a *query* protocol, whilst (a) portrays a particular version of an *information-seeking* interplay where the tutor asks questions to which she already knows the answers. In short, although *information-giving/seeking* and *query* dialogue elements are shared by both education and EQR interactions, *inquiry* and *'quiz'* aspects pertain only to the former. The EQR protocol focuses more on the *persuasion* facet and combines (and handles) all of its features together without requiring variants.

Future work. We can envisage several different extensions of the research we have presented. For example, the fully-fledged EQR dialogue that we introduced here could be implemented via the EQRbot chatbot proposed by Castagna et al. [17,19,20]. In the cited works, we investigated instantiations of argument schemes as a way to deliver explanations to patients according to the treatment recommended by a clinical decision support system (the specific decision support system considered was the CONSULT system [4,31,46]). Such information delivery can be enhanced by a dialectical protocol purposely designed to strategically convey explanations such as the EQR dialogue. Following the EQRbot implementation, another potential direction to pursue would involve representing the arguments associated with each locution of the EQR protocol as structured (rather than abstract) entities. ASPIC$^+$ [66] would be a suitable formalism to use for this, especially in its dialectical version D-ASPIC$^+$ which accounts for resource-bounded agents [25]. Consider also that, in real-world dialogues, human agents do not always make fully formed arguments. Instead, they often make incomplete arguments, called enthymemes [43,93,94]. Incorporating enthymemes in the EQR dialogue protocol would thus generate

a better approximation of the everyday exchange of arguments performed by real-world agents, and that would be a further possible direction for future work. In addition, all of the previous approaches could, and arguably should, be tested via specific user studies to evaluate the quality of the proposed explanations. Finally, it may also be interesting to provide a comparison of the EQR protocol and other state-of-the-art explanation mechanisms in the realm of LLMs, such as the Tree of Thoughts (ToT) method [95]. Devised as an advanced reasoning strategy for LLMs' ability in problem-solving tasks that harnesses the exploration and evaluation of multiple *thoughts* (i.e., coherent language sequences), ToT may also improve the interpretability of the decisions of a model. This approach employs search algorithms and backtrack processes to probe all of an LLM's thoughts. In comparison to ToT, EQR dialogues leverage AFs and computational argumentation semantics (given the equivalence proved in Theorem 1) in order to provide the best 'argumentative path' leading to explanations, thus resulting in a more intuitive and human-friendly approach. Notice also that such paths account for divergent information, therefore mimicking and (potentially) outperforming the recent CCoT (Contrastive Chain of Thought) prompting technique, which mostly handles only one contrastive explanation at a time [21].

6. Conclusion

Stemming as a novel approach to argumentation-based explanations for addressing XAI concerns, the Explanation–Question–Response dialogue introduced herein presents a fully-fledged protocol and a set of fundamental characteristics. These features include: (1.) a simple protocol that avoids meta-level locutions to manage the dialectical interplay whilst conveniently embedding multiple dialogue types. Compared to other dialogues that require a Control Layer, the simplicity of the EQR design favours its implementation. (2.) EQR exchanges of arguments result in interactions satisfying desirable properties of explanations (i.e., exhaustivity, selectivity, transfer of understanding and contextuality). Lastly, (3.) the information conveyed by a terminated EQR dialogue proves to be justified by a series of compelling reasons. Indeed, such an explanation turns out to be sound and complete with respect to Dung's AFs admissible semantics: this equivalency thus allows us to evaluate the EQR dialogue moves using any proof theory, algorithmic procedures, or methodologies semantically associated with computational argumentation. Future investigation will focus on ways for extending the introduced protocol and testing our hypothesis that EQR dialogue provides a valid instrument in progressing XAI towards the recent challenges posed by the rise of LLMs.

Acknowledgements

This research was partially funded by the UK Engineering & Physical Sciences Research Council (EPSRC) under grant #EP/P010105/1.

References

[1] C. Antaki and I. Leudar, Explaining in conversation: Towards an argument model, in: *European Journal of Social Psychology*, Vol. 22, Wiley Online Library, 1992, pp. 181–194.

[2] K. Atkinson, T. Bench-Capon and P. McBurney, Computational representation of practical argument, in: *Synthese*, Vol. 152, Springer, 2006, pp. 157–206.

[3] J.L. Austin, *How to do Things with Words*, Oxford University Press, 1962.

[4] P. Balatsoukas, T. Porat, I. Sassoon, K. Essers, N. Kokciyan, M. Chapman, A. Drake, S. Modgil, M. Ashworth, E.I. Sklar et al., User involvement in the design of a data-driven self-management decision support tool for stroke survivors, in: *IEEE EUROCON 2019 – 18th International Conference on Smart Technologies*, IEEE, 2019, pp. 1–6.

[5] V. Bellotti and K. Edwards, Intelligibility and accountability: Human considerations in context-aware systems, in: *Human–Computer Interaction*, Vol. 16, Taylor & Francis, 2001, pp. 193–212.

[6] T.J. Bench-Capon and P.E. Dunne, Argumentation in artificial intelligence, in: *Artificial Intelligence*, Vol. 171, Elsevier, 2007, pp. 619–641.

[7] P. Besnard and A. Hunter, *Elements of Argumentation*, MIT press, Cambridge, 2008.

[8] F. Bex and D.N. Walton, Combining explanation and argumentation in dialogue, in: *Argument & Computation*, Vol. 7, IOS Press, 2016, pp. 55–68.

[9] S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu and W. Saunders, Language models can explain neurons in language models, 2023.

[10] E. Black and A. Hunter, A generative inquiry dialogue system, in: *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, Association for Computing Machinery, 2007, pp. 1–8.

[11] E. Black, N. Maudet and S. Parsons, Argumentation-based dialogue, in: *Handbook of Formal Argumentation*, D. Gabbay, M. Giacomin, G. Simari and M. Thimm, eds, Vol. 2, College Publications, 2021, p. 511. ISBN 978-1-84890-336-4.

[12] T. Blanchard, Explanatory abstraction and the Goldilocks problem: Interventionism gets things just right, *The British Journal for the Philosophy of Science* (2020).

[13] T. Blount, Modelling eristic and rhetorical argumentation on the social web, PhD thesis, University of Southampton, 2018.
[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language models are few-shot learners, *Advances in neural information processing systems* **33** (2020), 1877–1901.
[15] V. Buisman, Designing a formal dialogue system with personalized explanations for XAI, Master's thesis, 2023.
[16] F. Castagna, Towards a fully-fledged formal protocol for the Explanation–Question–Response dialogue, in: *Online Handbook of Argumentation for AI*, 2022, pp. 17–21.
[17] F. Castagna, Dialectical argumentative characterisations for real-world resource-bounded agents, PhD thesis, King's College London, 2022.
[18] F. Castagna, Dialectical argument game proof theories for classical logic, *Journal of Applied Logics* **2631**(3) (2023), 279.
[19] F. Castagna, A. Garton, P. McBurney, S. Parsons, I. Sassoon and E.I. Sklar, EQRbot: A chatbot delivering EQR argument-based explanations, *Frontiers in Artificial Intelligence* **6** (2023), 1045614. doi:10.3389/frai.2023.1045614.
[20] F. Castagna, S. Parsons, I. Sassoon and E.I. Sklar, Providing explanations via the EQR argument scheme, in: *Computational Models of Argument: Proceedings of COMMA 2022*, IOS Press, 2022, pp. 351–352.
[21] Y.K. Chia, G. Chen, L.A. Tuan, S. Poria and L. Bing, Contrastive chain-of-thought prompting, 2023, arXiv preprint arXiv:2311.09277.
[22] E. Cogan, S. Parsons and P. McBurney, New types of inter-agent dialogues, in: *International Workshop on Argumentation in Multi-Agent Systems*, Springer, 2005, pp. 154–168.
[23] M. Colombetti and M. Verdicchio, An analysis of agent speech acts as institutional actions, in: *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 3*, Association for Computing Machinery, 2002, pp. 1157–1164.
[24] K. Čyras, A. Rago, E. Albini, P. Baroni and F. Toni, Argumentative XAI: A survey, in: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) Survey Track*, International Joint Conferences on Artificial Intelligence, 2021.
[25] M. D'Agostino and S. Modgil, A fully rational account of structured argumentation under resource bounds, in: *International Joint Conference on Artificial Intelligence (IJCAI-20), IJCAI*, 2020, pp. 1841–1847.
[26] D. Das, S. Banerjee and S. Chernova, Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery, in: *Proceedings of the 2021 ACM/IEEE International Conference on Human–Robot Interaction, HRI '21*, Association for Computing Machinery, New York, NY, USA, 2021, pp. 351–360. ISBN 9781450382892. doi:10.1145/3434073.3444657.
[27] L.A. Dennis and N. Oren, Explaining BDI agent behaviour through dialogue, *Autonomous Agents and Multi-Agent Systems* **36**(2) (2022), 29. doi:10.1007/s10458-022-09556-8.
[28] J. Díez, K. Khalifa and B. Leuridan, General theories of explanation: Buyer beware, *Synthese* **190**(3) (2013), 379–396. doi:10.1007/s11229-011-0020-8.
[29] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and N-person games, in: *Artificial Intelligence*, Vol. 77, Elsevier, 1995, pp. 321–357.
[30] F.H.V. Eemeren and R. Grootendorst, *Argumentation, Communication, and Fallacies: A Pragma-Dialectical Perspective*, Lawrence Erlbaum Associates, Inc., 1992.
[31] K. Essers, M. Chapman, N. Kokciyan, I. Sassoon, T. Porat, P. Balatsoukas, P. Young, M. Ashworth, V. Curcin, S. Modgil et al., The CONSULT system, in: *Proceedings of the 6th International Conference on Human–Agent Interaction*, 2018, pp. 385–386. doi:10.1145/3284432.3287170.
[32] X. Fan and F. Toni, On computing explanations in argumentation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29, 2015.
[33] L.R. Franklin-Hall, High-level explanation and the interventionist's 'variables problem', *The British Journal for the Philosophy of Science* (2016).
[34] R.A. Girle, Commands in dialogue logic, in: *Practical Reasoning*, D.M. Gabbay and H.J. Ohlbach, eds, Springer, Berlin Heidelberg, 1996, pp. 246–260. ISBN 978-3-540-68454-1. doi:10.1007/3-540-61313-7_77.
[35] M. Green, Speech acts, in: *The Stanford Encyclopedia of Philosophy*, Fall 2021 edn, Research Lab, Stanford University, 2021.
[36] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf and G.-Z. Yang, XAI – explainable artificial intelligence, in: *Science Robotics*, Vol. 4, American Association for the Advancement of Science, 2019, p. eaay7120.
[37] C.L. Hamblin, Fallacies, Methuen, London, 1970.
[38] N.R. Hanson, *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*, CUP Archive, 1965.
[39] G. Hesslow, The problem of causal selection, *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988), 11–32.
[40] D. Hilton, Social attribution and explanation, in: *Oxford Handbook of Causal Reasoning*, 2017.

[41] M. Hinton, CAPNA – the comprehensive assessment procedure for natural argumentation, in: *Evaluating the Language of Argument*, Springer International Publishing, Cham, 2021, pp. 167–194. ISBN 978-3-030-61694-6. doi:10.1007/978-3-030-61694-6_11.

[42] M. Hinton and J.H. Wagemans, How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator, *Argument & Computation* (2022), 1–16.

[43] S.A. Hosseini, S. Modgil and O. Rodrigues, Dialogues incorporating enthymemes and modelling of other agents' beliefs, PhD thesis, King's College London, 2017.

[44] J. Hulstijn, Dialogue models for inquiry and transaction, PhD thesis, Universiteit Twente, Enschede, The Netherlands, 2000.

[45] L. Ibrahim, M.M. Ghassemi and T. Alhanai, Do explanations improve the quality of AI-assisted human decisions? An algorithm-in-the-loop analysis of factual & counterfactual explanations, in: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '23*, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2023, pp. 326–334. ISBN 9781450394321.

[46] N. Kokciyan, M. Chapman, P. Balatsoukas, I. Sassoon, K. Essers, M. Ashworth, V. Curcin, S. Modgil, S. Parsons and E.I. Sklar, A collaborative decision support tool for managing chronic conditions, in: *The 17th World Congress of Medical and Health Informatics*, 2019.

[47] E.C. Krabbe, Dialogue logic revisited, in: *Aristotelian Society Supplementary Volume*, Vol. 75, Oxford University Press, Oxford, UK, 2001, pp. 33–49.

[48] S.A. Kripke, A completeness theorem in modal logic1, *The journal of symbolic logic* **24** (1959), 1–14, Cambridge University Press.

[49] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan and W.-K. Wong, Too much, too little, or just right? Ways explanations impact end users' mental models, in: *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, IEEE, 2013, pp. 3–10. doi:10.1109/VLHCC.2013.6645235.

[50] H. Lakkaraju, D. Slack, Y. Chen, C. Tan and S. Singh, Rethinking explainability as a dialogue: A practitioner's perspective, 2022, arXiv preprint arXiv:2202.01875.

[51] P. Lipton, Contrastive explanation, *Royal Institute of Philosophy Supplements* **27** (1990), 247–266. doi:10.1017/S1358246100005130.

[52] T. Lombrozo, The structure and function of explanations, *Trends in cognitive sciences* **10**(10) (2006), 464–470. doi:10.1016/j.tics.2006.08.004.

[53] M. Luck, P. McBurney, O. Shehory and S. Willmott, *Agent Technology, Computing as Interaction: A Roadmap for Agent Based Computing*, University of Southampton on behalf of AgentLink III, 2005.

[54] P. Madumal, T. Miller, L. Sonenberg and F. Vetere, A grounded interaction protocol for explainable artificial intelligence, 2019, arXiv preprint arXiv:1903.02409.

[55] P. McBurney, D. Hitchcock and S. Parsons, The eightfold way of deliberation dialogue, in: *International Journal of Intelligent Systems*, Vol. 22, Wiley Online Library, 2007, pp. 95–132.

[56] P. McBurney and S. Parsons, Chance discovery using dialectical argumentation, in: *Annual Conference of the Japanese Society for Artificial Intelligence*, Springer, 2001, pp. 414–424.

[57] P. McBurney and S. Parsons, Games that agents play: A formal framework for dialogues between autonomous agents, *Journal of logic, language and information* **11** (2002), 315–334, Springer.

[58] P. McBurney and S. Parsons, Retraction and revocation in agent deliberation dialogs, in: *Argumentation*, Vol. 21, Springer, 2007, pp. 269–289.

[59] P. McBurney and S. Parsons, Dialogue games for agent argumentation, in: *Argumentation in Artificial Intelligence*, Springer, 2009, pp. 261–280. doi:10.1007/978-0-387-98197-0_13.

[60] P. McBurney and S. Parsons, Argument schemes and dialogue protocols: Doug Walton's legacy in artificial intelligence, *Journal of Applied Logics* **8** (2021), 263–286, College Publications.

[61] P. McBurney, S. Parsons, K. Atkinson, H. Prakken and A. Wyner, Talking about doing, in: *From Knowledge Representation to Argumentation in AI, Law and Policy Making*, College Publications, 2013, pp. 151–166.

[62] P. McBurney, R.M. Van Eijk, S. Parsons and L. Amgoud, A dialogue game protocol for agent purchase negotiations, in: *Autonomous Agents and Multi-Agent Systems*, Vol. 7, Springer, 2003, pp. 235–273.

[63] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, in: *Artificial Intelligence*, Vol. 267, Elsevier, 2019, pp. 1–38.

[64] T. Miller, P. Howe and L. Sonenberg, Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences, 2017, arXiv preprint arXiv:1712.00547.

[65] S. Modgil and M. Caminada, Proof theories and algorithms for abstract argumentation frameworks, in: *Argumentation in Artificial Intelligence*, Springer, 2009.

[66] S. Modgil and H. Prakken, A general account of argumentation with preferences, in: *Artificial Intelligence*, Vol. 195, Elsevier, 2013, pp. 361–397.

[67] J.D. Moore and C.L. Paris, Planning text for advisory dialogues: Capturing intentional and rhetorical information, Technical report, University of Southern California Marina Del Rey Information Sciences Institute, 1993.

[68] OpenAI, GPT-4 Technical report, arXiv, 2023.

[69] J.L. Pollock, Defeasible reasoning, in: *Cognitive Science*, Vol. 11, Elsevier, 1987, pp. 481–518.

[70] H. Prakken, Formal systems for persuasion dialogue, in: *The Knowledge Engineering Review*, Vol. 21, Cambridge University Press, 2006, pp. 163–188.

[71] R. Reiter, A logic for default reasoning, in: *Artificial Intelligence*, Vol. 13, Elsevier, 1980, pp. 81–132.

[72] D.A. Robb, X. Liu and H. Hastie, Explanation styles for trustworthy autonomous systems, in: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '23*, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2023, pp. 2298–2300. ISBN 9781450394321.

[73] I. Sassoon, N. Kökciyan, E.I. Sklar and S. Parsons, Explainable argumentation for wellness consultation, in: *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer, 2019, pp. 186–202. doi:10.1007/978-3-030-30391-4_11.

[74] R. Schank, Understanding mechanically and creatively, 1986.

[75] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, 1969.

[76] J. Searle, A taxonomy of illocutionary acts, in: *Language, Mind, and Knowledge*, University of Minnesota Press, 1975, pp. 1–29.

[77] Z. Shams, D.V. Marina, O. Nir and P. Julian, *Normative Practical Reasoning via Argumentation and Dialogue*, 2020.

[78] M.P. Singh, An ontology for commitments in multiagent systems, in: *Artificial Intelligence and Law*, Vol. 7, Springer, 1999, pp. 97–113.

[79] E.I. Sklar and S. Parsons, Towards the application of argumentation-based dialogues for education, in: *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, Vol. 4, IEEE Computer Society, 2004, pp. 1420–1421.

[80] K. Sokol and P. Flach, One explanation does not fit all: The promise of interactive explanations for machine learning transparency, *KI-Künstliche Intelligenz* **34**(2) (2020), 235–250. doi:10.1007/s13218-020-00637-y.

[81] F. Toni, A tutorial on assumption-based argumentation, *Argument & Computation* **5**(1) (2014), 89–117. doi:10.1080/19462166.2013.869878.

[82] T. Trabasso and J. Bartolone, Story understanding and counterfactual reasoning, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **29**(5) (2003), 904.

[83] A. Vassiliades, N. Bassiliades and T. Patkos, Argumentation and explainable artificial intelligence: A survey, in: *The Knowledge Engineering Review*, Vol. 36, Cambridge University Press, 2021.

[84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017).

[85] G. Vilone and L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion* **76** (2021), 89–106. doi:10.1016/j.inffus.2021.05.009.

[86] D. Walton, Types of dialogue and burdens of proof, in: *COMMA*, IOS Press, 2010, pp. 13–24.

[87] D. Walton, A dialogue system specification for explanation, *Synthese* **182** (2011), 349–374. doi:10.1007/s11229-010-9745-z.

[88] D.N. Walton, Burden of proof, in: *Argumentation*, Vol. 2, Kluwer Academic Publishers, 1988, pp. 233–254.

[89] D.N. Walton and E.C. Krabbe, *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*, SUNY Press, 1995.

[90] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler et al., Emergent abilities of large language models, 2022, arXiv preprint arXiv:2206.07682.

[91] S. Wells and C.A. Reed, A domain specific language for describing diverse systems of dialogue, *Journal of Applied Logic* **10**(4) (2012), 309–329. doi:10.1016/j.jal.2012.09.001.

[92] D.A. Wilkenfeld and T. Lombrozo, Inference to the best explanation (IBE) versus explaining for the best inference (EBI), *Science & Education* **24**(9) (2015), 1059–1077. doi:10.1007/s11191-015-9784-4.

[93] A. Xydis, C. Hampson, S. Modgil and E. Black, Enthymemes in dialogues, in: *COMMA*, IOS Press, 2020, pp. 395–402.

[94] A. Xydis, C. Hampson, S. Modgil and E. Black, A sound and complete dialogue system for handling misunderstandings, in: *4th International Workshop on Systems and Algorithms for Formal Argumentation, SAFA 2022*, CEUR-WS, 2022, pp. 19–32.

[95] S. Yao, D. Yu, J. Zhao, I. Shafran, T.L. Griffiths, Y. Cao and K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, 2023, arXiv preprint arXiv:2305.10601.

[96] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., A survey of large language models, 2023, arXiv preprint arXiv:2303.18223.