# Argumentation Theoretical Frameworks for Explainable Artificial Intelligence

**Martijn H. Demollin***
Laboratory of The New Ethos,
Faculty of Administration and Social Sciences,
Warsaw University of Technology, Poland
Martijn.Demollin@pw.edu.pl

**Qurat-ul-ain Shaheen***
IIIA-CSIC, Spain
qurat@iiia.csic.es

**Katarzyna Budzynska**
Laboratory of The New Ethos,
Faculty of Administration and Social Sciences,
Warsaw University of Technology, Poland
Katarzyna.Budzynska@pw.edu.pl

**Carles Sierra**
IIIA-CSIC, Spain
sierra@iiia.csic.es

## Abstract

This paper discusses four major argumentation theoretical frameworks with respect to their use in support of explainable artificial intelligence (XAI). We consider these frameworks as useful tools for both system-centred and user-centred XAI. The former is concerned with the generation of explanations for decisions taken by AI systems, while the latter is concerned with the way explanations are given to users and received by them.

## 1 Introduction

The enforcement of GDPR (https://gdpr-info.eu/) by the EU has made eXplainable Artificial Intelligence (XAI) into a rapidly growing area of research over the last two years. While there is no standard definition of explainable AI systems yet, the need itself is undisputed as evidenced by the GDPR requirements. Also, there is agreement that explainability for AI systems is as diverse as the systems themselves. Neerincx et al. have defined three phases in the explanation of an AI system: (1) explanation generation, (2) explanation communication, and (3) explanation reception (Neerincx et al., 2018). Based on this, recent XAI literature can be divided into two types: *system-centred* and *user-centred* XAI.

System-centred XAI is focused on phase 1. Broadly, systems fall into two main categories: black-box subsymbolic systems such as those based on deep learning and white-box symbolic systems like decision trees or rule-based. A consequence of the GDPR implementation has been a recent explosion in grey-box systems, which aim to add some symbolic layer to black-box systems to add transparency (Guidotti et al., 2018; Chakraborty et al., 2017; Tjoa and Guan, 2015).

User-centred XAI, which is concerned with aspects related to user-interaction and experience (Ribera Turró and Lapedriza, 2019), is mainly focused on phases 2 and 3 and aims to integrate a user into the loop of an AI system's decision making as much as possible (Anjomshoae et al., 2019). Phase 2 deals with what is exactly to be provided to the end-user and how to present it, while phase 3 is concerned with the level of understanding that is achieved in an end-user with an explanation.

For these varying tasks identified within system-centred and user-centred XAI, it is useful to consider which argumentation theoretical framework can best provide the output that is most effective in a particular setting. In this paper, we briefly discuss the roles that some of the main argumentation theories can play for both system-centred and user-centred XAI approaches. Section 2 presents the role of Dung's theories and Walton's dialogue for achieving system-centred XAI, while Section 3 explores how Pragma-dialectics and Inference Anchoring Theory contribute towards user-centred XAI. Finally, Section 4 makes some final observations about the suitability of each theory to XAI.

## 2 System-centred XAI

Most of the literature on system-centred XAI does not differentiate between *interpretability* and *explainability* of learning models. Guidotti et al.

* Both M.D. and Q.S. contributed equally in the writing of this paper. Specifically, M.D. focused on Sections 3 and 4 and Q.S. focused on Sections 1 and 2.



*Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020)*, pages 44–49,
Dublin, Ireland, 18 December 2020. 

(Guidotti et al., 2018) consider *explainability* as an interface between interpretable models and human users. They formalise four types of explanations for black boxes: (1) simulating them with an equivalent symbolic model that explains its working, (2) explaining only the black-box outcome rather than its working, (3) providing visual representation of the black-box mechanism for inspection, and (4) a transparent model that is fully explainable on its own without needing any supplementary explanations model. Rudin makes a distinction between *interpretable* ML and *explainable* ML (Rudin, 2019) where the latter involves the first three types of explanations as identified by Guidotti et al. while the former includes the last type. Based on this discussion, recent approaches to system-centred XAI can be classified into two main types: *interpretable* and *non-interpretable*. Interpretable black-box models can either be purely symbolic models or grey-box models, that is, those that generate intermediate symbols which can be leveraged for generating a trace of the reasoning process used by the model. Non-interpretable models will then refer to black-boxes for which only input and output are available. Thus, achieving explainability nails down to answering the question of how to generate or extract the symbols out of the black-boxes that make up the explanations. In the next two sections, we explore some preliminary ideas on how Abstract Argumentation Framework (AF) and dialogue theory can help us answer this question.

### 2.1 Abstract Argumentation Framework

An $AF$ as defined by Dung (Dung, 1995) is a pair, $\langle A, R \rangle$, where $A$ is a set of arguments and $R$ is a binary relation on the set $A$ which determines the attack relations between the arguments (Baroni and Giacomin, 2009). The arguments in an AF are atomic entities without any internal structure. This abstraction allows generalising the properties of the framework independently of the internal argument structure and possibly re-using the argumentation model across specific problems. AFs can be used as the formalism underpinning explanations for a black-box as we can see next.

For any black-box model, the data can be classified into three types: input, output and intermediate symbols which are generated during the learning process. Given such a black-box model, we consider different routes to XAI. A simple one would be to use a decision tree based approach as an initial step to build an AF. First, we apply a classification algorithm such as ID3 (Quinlan, 1986) over a table that contains the input (and possibly the intermediary data) as features of items and the output as their classes. The arguments of the AF could then be extracted from the decision tree. The labels (arguments) in $A$ would be any subset of the nodes of the tree, including singletons. The label (argument) of the set of nodes in a path leading to a class $C_i$ would attack the label representing the set of nodes of a path leading to a different class $C_j$. Other attack relations could be found, as well as other relationships for variants of Dungs model like Bipolar argumentation systems (Cayrol and Lagasquie-Schiex, 2005). For instance, labels representing the nodes in paths leading to the same class $C_i$ support each other. Then explanations of an output (a class) can become the arguments (nodes and paths) in some preferred semantics over AF. This approach makes sense only if the input data is symbolic.

Figure 1 shows the schema of the data set for the classification algorithm.Each row in the table corresponds to a training example. Each column represents a feature label. *Input features* represent the input (independent) variables represented by $i_{pn}$ where $p \in$ row number and $n \in$ column number. *Intermediary features* represent the intermediate symbols such as outputs of hidden layers generated from a black box model such as a neural network. These are represented by $m_{pm}$ where $p \in$ row number as before and $m \in$ column number. *Output class* indicates the corresponding classification label for each row, represented by $c_p$

### 2.2 Dialogue Theory

Dialogue theory in argumentation can play a vital role in bridging the explanation gap between machine recommendations and human trust and understanding of these. During a dialogue, one party is typically seeking to persuade the other party to agree to some disputed conclusion. In contrast, while providing an explanation, one party is trying to provide some information to the other party in order to improve understanding of the already accepted conclusion (Walton, 2009). In this context, argumentation dialogues can be used to query black-box models on their intermediate symbols in order to generate more enriched explanation models. For example, consider a hypothetical decision system on the lines of COMPAS (Larson et al., 2016) which recommends parole or not for convicts on the basis of past parole violations and age. The

| Input features | | | | intermediary features | | | | Output class |
|---|---|---|---|---|---|---|---|---|
| $i_{11}$ | $i_{12}$ | ... | $i_{1n}$ | $m_{11}$ | $m_{12}$ | ... | $m_{1m}$ | $c_1$ |
| $i_{21}$ | $i_{22}$ | ... | $i_{2n}$ | $m_{21}$ | $m_{22}$ | ... | $m_{2m}$ | $c_2$ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| $i_{p1}$ | $i_{p2}$ | ... | $i_{pn}$ | $m_{p1}$ | $m_{p2}$ | ... | $m_{pm}$ | $c_p$ |

Figure 1: Input table for a classification algorithm. Sets of features become arguments in an AF.

system can be queried for an explanation of specific outcomes such as 'Why is this parole granted?' The system could use the features used for recommendation as justification such as 'Because there are no past parole violations'. In this case, the user was able to gain some information from the system.

Another scenario could be a case where the explanation model poses a question to the AI system regarding a feature which the decision system has not considered. For example, assuming that there is a query from the user to justify the outcome for the hypothetical parole system such as 'Is it because of my ethnicity?' In this case, ethnicity is not something the system has taken into account. So the system can try to find the symbols that can help it to determine the correlation and inform the user accordingly. In this way, the system is forced to look for more information resulting in not only a more enriched explanation model for the user but also more transparency for the system as it can cause hidden biases and correlations to be identified. Both these examples fall under the *Information Seeking Dialogue* type proposed by Walton where the dialogue goal is an information exchange. The argument generation approach from Section 2.1 can be combined with dialogue generation in the manner of Walton to explain black-box models as highlighted in this section.

## 3 User-centred XAI

User-centred XAI focuses on the way explanations generated by AI systems are communicated to non-expert and non-technical users, who often do not require a full understanding of the inner workings of the system they are interacting with. Instead, this type of user will be primarily interested in natural language explanations that are maximally understandable, that build trust and confidence in the system's recommendations, and that inform a user about how to alter the outcome of a decision (Hind, 2019). For example, when an AI system rejects a user's application for a bank loan, it should explain in natural language which variable (e.g. salary, or outstanding debt) is responsible for this output and what is needed in order to be eligible for a loan.

Providing explanations to non-expert users of AI systems is widely recognised as an essential component in XAI, but adapting these explanations to the particular needs of a user and the communicative setting in which they occur remains a challenging task (Anjomshoae et al., 2019). In order to endow AI systems with trustworthy and realistic interactive capabilities it is necessary to model the dialogical setting in which users interact with AI systems and to determine which types of communication and reasoning are most effective for informing users. The following two sections discuss the pragma-dialectical theory of argumentation and Inference Anchoring Theory, which are theoretical frameworks for modelling argumentation and reasoning in natural language.

### 3.1 The Pragma-dialectical Theory of Argumentation

The pragma-dialectical theory of argumentation (Van Eemeren and Grootendorst, 2004) is designed to allow for the analysis and evaluation of argumentation as it is actually used in communication. In pragma-dialectics, argumentation is considered as a complex and interlinked array of speech acts, which are directed towards fulfilling the ultimate goal of a critical discussion: the reasonable resolution of a conflict of opinion. Ideally, a discussion consists of four dialectical stages, which are (1) the confrontation stage, (2) the opening stage, (3) the argumentation stage and finally, (4) the concluding stage. In the confrontation stage, arguers establish they have a difference of opinion, which they may decide to attempt to resolve in the opening stage. The argumentation stage is dedicated to providing arguments in support of the standpoints proposed by the arguers and in the concluding stage the parties determine whether their difference of opinion has been resolved and in who's favour (Van Eemeren and Grootendorst, 2004, pp. 59-62).

In the context of user-centred XAI, this allows us to determine how the exchange of messages between a system and a user should be specified at different stages of communication, e.g. an explanation should be differently communicated depending on whether a message is provided in the confrontation stage or the argumentation stage.

The pragma-dialectical theory also stipulates ten rules for a critical discussion (Van Eemeren and Grootendorst, 2004, pp. 190-196), which represent the conditions arguers must uphold in order to ensure a reasonable discussion. Any violation of these critical discussion rules constitutes a hindrance towards the reasonable resolution of the conflict of opinion and is considered a fallacious argumentative move. These rules for a critical discussion reflect the normative element of the pragma-dialectical theory of argumentation and allow for an evaluation of the reasonableness of argumentation in actual language use. As such, the pragma-dialectical theory makes it possible to model the dialogical setting in which a user-AI interaction takes place and to establish whether the arguments that are used are fair and suited to the intended goals of an AI system's end user.

### 3.2 Inference Anchoring Theory

Inference Anchoring Theory (IAT) (Budzynska and Reed, 2011) is a theoretical framework for connecting the inferential structures that are present in argumentation with their associated illocutionary forces and dialogical processes. Consider the following example, taken from (Budzynska and Reed, 2011):

(1) a. Bob: *p is the case*
b. Wilma: *Why p?*
c. Bob: *q*.

Example (1) contains a dialogical structure that represents the order in which the propositions were uttered, which is governed by dialogical rules that stipulate how the participants in the dialogue may make communicative moves. This locutionary level (i.e. what is actually being said) of the dialogue and the transitions between the statements made are represented on the right-hand side of the diagram shown in Figure 2. Additionally, (1) can be viewed as containing a basic inferential structure, including a premise ($p$) and a conclusion ($q$). This propositional content and its logical structure are represented on the left-hand side. Central to IAT, the propositional content of a dialogue is 'anchored' in its respective locution or transitions through an illocutionary connection which represents the illocutionary force (Searle, 1969) that is exerted with a particular statement (e.g. asserting, arguing, or promising) and is represented in the middle of the diagram.

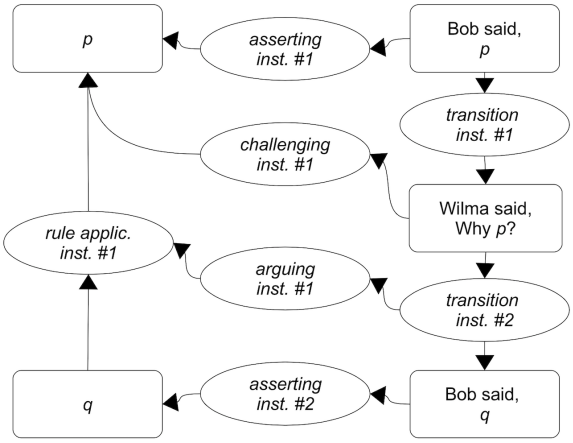


Figure 2: Interaction between argument and dialogue in IAT (Budzynska and Reed, 2011).

In summary, Inference Anchoring Theory allows to unpack four dimensions of explanations which can be then differently computed: it is possible to link (1) dialogical acts ("Bob said: *p is the case*") to (2) their propositional contents ($p$) through (3) an illocutionary connection that signifies the communicative intention of the speaker/user (*asserting instance #1*) linked to (4) ethotic conditions that allow us to express the credibility, trustworthiness, and character of a speaker (user modelling). This is particularly valuable for the task of user-centred XAI, since it enables the adaptation of argumentation and explanation to specific users.

## 4 Discussion and Future Work

In this paper, we have differentiated between system-centred and user-centred XAI, and discussed how four major argumentation theoretical frameworks can be applied to these challenges. Depending on the type of explanation required from an AI system, it is useful to consider the various theoretical tools that these approaches offer. Abstract Argumentation and dialogue theory excel in generating explanations of the inner workings of an AI system and modelling inter-system interaction. Pragma-dialectics and Inference Anchoring Theory are especially suited towards modelling the dialogical setting of human-AI interaction and identifying which type of reasoning is most effective there.

Future work on system-centred XAI could explore how Abstract Argumentation Framework and dialogue theory can be used in a multi-agent recommender system. In this case, the goal is to achieve explainability for the joint recommendation made by multiple systems after consensus. However, in order to achieve consensus, we need dialogue between the different systems. In this context, we can explore using Abstract Argumentation Framework for justifying the recommendation and dialogue theory for achieving consensus on the recommendation itself.

For user-centred XAI, we propose to investigate how pragma-dialectics and Inference Anchoring Theory can be applied for modelling users in social media. To this end, Natural Language Processing techniques such as argument mining can help create an image of a user's linguistic profile, which provides insight into their communicative behaviour and reasoning patterns (i.e. argumentation schemes). In turn, these argumentation schemes can form a blueprint for the generation of arguments and explanations that are tailored to a specific communicative situation and a particular user. In that capacity, argumentation schemes carry substantial value for tasks in explainable AI related to language generation, inter-agent communication, and personalising AI systems to end users.

To conclude, in order to further improve our understanding of, and our interaction with AI systems, we believe it is fruitful to build on existing argumentation theoretical frameworks in various ways towards more robust and accurate methods for eXplainable Artificial Intelligence.

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621.

## References

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1078–1088, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Pietro Baroni and Massimiliano Giacomin. 2009. Semantics of abstract argument systems. In *Argumentation in Artificial Intelligence*, pages 25–44. Springer US.

Katarzyna Budzynska and C. Reed. 2011. Whence inference. Technical report, University of Dundee.

C. Cayrol and M. C. Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3571 LNAI, pages 378–389. Springer Verlag.

S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram. 2017. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5).

Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):16–19.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 201 6. How We Analyzed the COMPAS Recidivism Algorithm — ProPublica.

Mark A. Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. 2018. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *Engineering Psychology and Cognitive Ergonomics*, pages 204–214, Cham. Springer International Publishing.

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

Mireia Ribera Turró and Agata Lapedriza. 2019. Can we do better explanations? A proposal of User-Centered Explainable AI.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Erico Tjoa and Cuntai Guan. 2015. A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI. Technical Report 8.

Frans H. Van Eemeren and R. Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Douglas Walton. 2009. Argumentation Theory: A Very Short Introduction. In *Argumentation in Artificial Intelligence*, pages 1–22. Springer US, Boston, MA.