Towards Argumentative Dialogue as a Humane Interface between People and Intelligent Machines

Simon Wells[1][0000−0003−4512−7868]

Edinburgh Napier University, Edinburgh, United Kingdom. s.wells@napier.ac.uk
http://www.simonwells.org

Abstract. We sketch the framework for a theoretical and applied system that we are developing which uses argumentation schemes and dialogue games to support dialectical interaction between people and machine learning systems. The goal is to support the automated justification and explanation of decisions made by AI systems, through a natural, human-oriented interface, in response to contemporary societal concerns about the impact of AI decisions upon individuals.

Keywords: Artificial Intelligence · Machine Learning · Explainable AI · Argumentation · Dialogue · Explanation · Justification

1 Introduction

Decisions are increasingly made by intelligent machines. Whilst many of these decisions are operational, for example trading decisions made by financial technology (FinTech) software, the A.I. wavefront, constituting the range of problems that can be decided by machines, is accelerating. It is likely that coming years will bring new decision making machines acting within at least the Banking, Medical, Insurance, Legal, and Transport sectors. Sectors that are heavily regulated and in which there is frequently a legal mandate that the output of predictive models, and the decisions that follow from them, can be scrutinised and interpreted. Traditionally the trade-off has been towards simpler linear predictive models, sacrificing some accuracy for simplicity and hence understandability of the model. However this has changed in recent years as findings in Machine Learning (ML) [1] and Deep Learning (DL) [2] have produced advances in capability but with the expense of increased complexity and reduced understandability.

When interpretation of results is difficult and understandability of the model is poor there is a consequent barrier to both responsible and legal use of those systems as well as wider, more general acceptance of the systems. There is also the additional risk that a form of "moral outsourcing" occurs in which operators of intelligent systems derogate from their responsibilities as a consequence of delegating to the machine, despite people legally remaining the "problem holders". Additionally, the vocal positions of a number of public commentators has fed both a popular and a legislative wariness of A.I. This has lead to, perhaps premature, moves towards legal regulation of intelligent machines, autonomous robots, and ML algorithms. Many of those calling for legislation, and many of the drafted regulations, propose explainable A.I.; systems that can generate explanations for their decisions in concert with the search for solutions. However this is not a capability of the intelligent systems that are at stake.

Conversely, techniques such as argumentative dialogue, are human friendly in a way that ML and DL based techniques are not. People communicate with each other to share knowledge using language, they argue in order to persuade others, they present arguments to justify their positions, they engage in dialogues to explore alternative positions, and they do these things frequently throughout their lives. Argumentative dialogue is thus a primary interface between people, and possesses attributes that are required for the interface between people and intelligent machines. Approaches to computational argumentation includes techniques for reasoning defeasibly about knowledge, and for managing various types of dialogue between arguers.

We propose that the explanatory role within A.I. should be taken up by the application of computational argument. Specifically that argumentative dialogue has the potential not merely to be a human computer interface but to be a humane interface between people and intelligent machines. Argumentative dialogue is a unifying technology that can enable intelligent machines not only to explain but also to justify their decisions and, when necessary, to select those arguments necessary to persuade others that the decision made was right under the circumstances. In this paper we demonstrate how the construction of such a humane interface draws together computational argument research at the logical, dialectical, procedural, and heuristic layers and applies it to the problem of understanding, explaining, and justifying decision making within intelligent machines based upon differing formalisms.

In this paper we present an overview of research to bring together increasingly capable ML techniques with increasingly sophisticated human-oriented, dialogue-based interfaces.

2 A Dual Process Model of Explanation

Dual process models have been used in psychology for many years to account for the multiple ways in which thinking happens. Whilst various authors have delineated the boundary between types of reasoning in a variety of ways, the basic split is between unconscious, automatic reasoning processes as exemplified by the majority of problems that are solved by the current generation of ML systems and conscious, controlled reasoning processes, the type of slow, rule-based, explicit, language linked logical reasoning that is exemplified by traditional symbolic AI, targeting domains such as natural language, or strategic planning. This approach accounts partly for the problem of explainable AI. Great advances have been made in solving problems associated with unconscious reasoning processes, but we conjecture that human acceptance of the results is often dependent upon conscious reasoning, an explanation of the reasoning in terms that people would usually use to explain their own behaviour to others. Furthermore, people often don't accept the base explanations for another's behaviour and require a defense of the behaviour at issue. Such verbal interactions, where the interlocutors may choose the right thing to say in order to persuade their opponent, are inherently strategic, and thus explainable AI systems, when acting in human society, will often need to go beyond mere explanation, to justify their behaviour given the context in which it occurs. To differentiate this specific subset of dialogical interactions, we refer to such Justificatory and Explanatory Techniques for Dialogues as "JET Dialogues".

Clearly a link between unconscious, ML-based, AI systems, and conscious reasoning processes to support JET dialogues is required. We propose that argumentation schemes [4], a mechanism for cataloging, relating, and criticising instances of reasoning, provide this link. Argumentation Schemes are formalisations of stereotypical patterns of reasoning, primarily as expressed in natural argumentative language. For example, the *Argumentation Scheme for the Argument from Sign* is schematized as follows:

Specific Premise: A (a finding) is true in this situation.
General Premise: B is generally indicated as true when its sign, A, is true.
Conclusion: B is true in this situation.

Schemes capture the idea that whilst the expression of a given instance of reasoning is often unique to the situation in which it occurs, the framework of expression, how the data (premises) are marshalled and linked to the result (conclusion), frequently fits into a finite range of possibilities. Furthermore, some instances of fallacious reasoning can often be equated with poor application of a scheme. Schemes have been catalogued yielding collections of hundreds of individual schemes[5] that refer to patterns of real-world reasoning.

One interpretation of Argumentation Schemes is that they can be used to form a functional link between the automatic and unconscious reasoning of modern data-driven ML systems and dialogue systems that can provide a human-oriented interface to the, otherwise opaque, reasoning processes that are happening. The argument for the link between schemes and ML-reasoning is straightforward, if modern data-driven AI systems make decisions based upon input data, then they can be described as reasoning systems. It has not been suggested that the current generation of AI can perform reasoning tasks that go beyond the kinds of reasoning that people do, but that they perhaps operate on different scales, for example, making decisions faster, more rapidly, more reliably, at greater scale, than a person could. However the actual reasoning processes are not suprahuman. Any decisions that a machine could make using such a system, and hence any conclusions that could be drawn, are not beyond the wit of humanity and will conform to existing patterns of reasoning. Patterns that have been recognised and formalised as Argumentation Schemes. It follows therefore that for each instance of ML reasoning, a corresponding argumentation scheme can be identified and instantiated for subsequent utilisation within an interactive dialogue system. In this approach, the inputs to the ML system correspond to premises within an argument, the outputs to conclusions drawn, and the ML process corresponds to a known instance of a reasoning process, identified by a named Argumentation Scheme, as illustrated in Fig. 1.

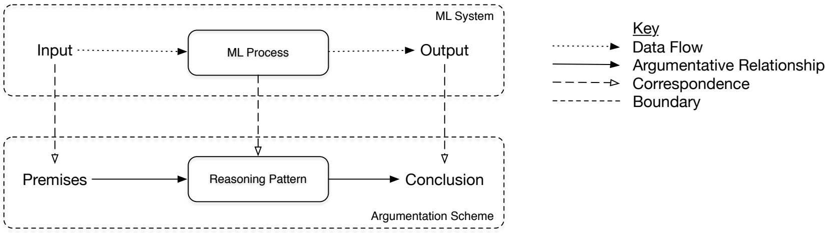

Fig. 1. Correspondences & mappings between reasoning in ML systems and reasoning archetypes described in Argumentation Schemes.

To this end we are aligning existing, identified argumentation schemes with published ML research, using the OpenML repository [3], in order to understand the kinds of reasoning that are achieved by current, state of the art processes. Argumentation Schemes that represent the reasoning process that occurs within a given ML system, together with the inputs and outputs to that system, are then used to generate utterances for use within an argumentative dialogue system as illustrated in Fig. 2. The system use the Dialogue Game Description Language (DGDL) [7] to represent various kinds of explanation and justification dialogue and dialogues are managed using A Dialogue MANagement Tool (ADAMANT), a dialogue runtime that manages dialogical interactions according to the rules expressed in DGDL descriptions and using the Argumentation scheme extensions to DGDL from [6]. The figure illustrates two agents, a ML agent and one or more other dialogue participants who interact with each other, in dialogues mediated by ADAMANT, yielding a dialogue transcript, a record of the explanation and justifications that have been produced.

3 Conclusions & Further Research

There are a number of threads of further work which roughly correspond to various sub-systems illustrated in Fig. 2. Firstly, mapping from ML representations to knowledge structures. Secondly, natural language generation. Thirdly, strategic and contextual personalisation of dialogue game interactions. This captures the notion that whilst an explanation will be fairly static, a factual statement of the reasoning process, a justification, which ideally persuades the other party, but at least is acceptable to them, may differ depending upon the circumstances

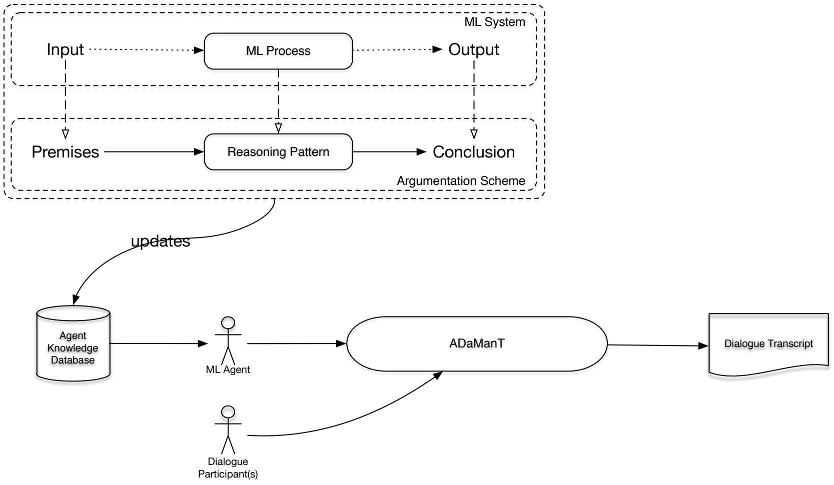

Fig. 2. Overview of data flow from ML systems into Dialogue Systems via Argumentation Schemes.

in which the dialogue takes place, and the nature of the other party. Persuasive arguments are very susceptible to the disposition, knowledge, and circumstances of the person to whom they are directed, choosing the right argument for the right person, strategy, is, as a result, extremely important in successful dialogical interaction. As these research threads are drawn together, so a humane interface to the underlying ML system is built enabling more rich, human-oriented interactions to occur.

References

1. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
3. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: Openml: Networked science in machine learning. SIGKDD Explorations **15**(2), 49–60 (2013). https://doi.org/10.1145/2641190.2641198, http://doi.acm.org/10.1145/2641190.2641198
4. Walton, D.N.: The New Dialectic. University of Toronto Press (1998)
5. Walton, D.N., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press (2008)
6. Wells, S.: Supporting argumentation schemes in argumentative dialogue games. Studies in Logic, Grammar and Rhetoric (SLGR) **36**(1), 171–191 (2014)
7. Wells, S., Reed, C.: A domain specific language for describing diverse systems of dialogue. Journal of Applied Logic **10**(4), 309–329 (2012)