Towards an Argument Scheme Classification for Ethical Reasoning

Elfia Bezou-Vrakatseli[1], Oana Cocarascu[1] and Sanjay Modgil[1]

[1]*Department of Informatics, King's College London, UK*

Abstract

This paper describes progress towards realising the long-term research goal of supporting dialogue between humans and between humans and AI systems so as to enable transparent and rational joint reasoning, in particular about matters of ethical significance. Key to realising this goal is this paper's proposed exploration and analysis of natural language moral debates, via argument schemes and critical questions. We believe this to be an important first step in identifying schemes and scheme taxonomies, specialised for ethical reasoning, and that can support both the natural language processing needed to support human-AI dialogue, as well as for scaffolding human-human dialogue.

Keywords

Dialogue, argument schemes, argument mining, ethics

1. Introduction & Motivation

Recent highly publicised successes in Artificial Intelligence (AI) applications have in large part been due to advances in machine learning (notably deep learning) [1]. However, symbolic approaches to AI will necessarily play a role in facilitating inter-agent (in particular human-AI and human-human) communication which is inherently a symbolic activity. In particular, reasoning in the presence of uncertainty, conflict, and disagreement typically benefits from multiple agents engaged in information sharing and joint reasoning; that is, in dialogical exchanges normatively governed by prescriptions encoded in logics for non-monotonic reasoning. One of the most promising paradigms for facilitating such dialogues is through argumentation-based characterisations of non-monotonic (nm) inference relations in terms of the exchange of arguments constructed from a given static belief base. These 'monological' characterisations can then be generalised to dialogical models in which agents exchange locutions (not limited to arguments), such that evaluation of a dialogical exchange in favour of a claim equates with that claim being a non-monotonic inference from the information shared during the course of the dialogue [2]. As dialogues consist of informational exchanges, they enable the understanding of the parties involved, not only as a direct result of the information each utterance conveys, but also as a means of exploring the reasoning of the interlocutors. As a result, dialogues also facilitate joint reasoning. Overall, the communication and joint reasoning of AI systems and humans leverages the strengths of AI along with those of human reasoning.

CMNA '22: Workshop on Computational Models of Natural Argument

elfia.bezou_vrakatseli@kcl.ac.uk (E. Bezou-Vrakatseli); oana.cocarascu@kcl.ac.uk (O. Cocarascu); sanjay.modgil@kcl.ac.uk (S. Modgil)

Our long-term research goal is guiding humans and AI systems alike in building and challenging arguments related to ethical issues as ethical reasoning is of crucial importance for decision-making systems that can impact human lives [3]. Moreover, enabling joint human and AI reasoning can help solve the *value alignment problem* [2] and help ensure the ethical behaviour, in accordance with human values, of agents as shown in [4]. In order to support dialogue between humans and between humans and AI systems that can enable transparent and rational joint reasoning, we need to draw from recent advances in natural language processing (NLP), and in particular argument mining [5], an area of research which focuses on extracting arguments and their relations from text (e.g. relations between premises and conclusions).

To this end, we begin by analysing natural language (NL) texts using argument schemes and critical questions. In particular, we focus on annotating moral debates using two argument scheme classifications: Walton [6] and Wagemans [7]. As the existing vast variety of taxonomies and argument schemes is one of the biggest problems in the literature, we purport to reconcile the two most used classifications and develop a theoretically well-founded, as well as practically useful, hybrid classification that can be applied to ethical debates. Said reconciliation leverages the strengths of both, while identifying the schemes particular to ethical reasoning. The novelty of our research can be found in the attempt to go beyond standard argument mining techniques (to determine the relation between premise and conclusion and identify support/attack relations between arguments) by making use of informal logic. In particular, we make use of argument schemes and critical questions that offer a *semantically* richer approach to argument classification, as premises support a conclusion *by virtue* of instantiating a scheme and support/attack relations are instigated *in response* to critical questions.

2. Background & Related Work

Argument schemes represent stereotypical patterns of reasoning and consist of a set of premises and a conclusion. Walton proposed over 60 argument schemes with corresponding sets of *critical questions* which are used to evaluate the strength of an argument [6] . For example, Walton's representation of the *argument from positive consequences* scheme is defined by a *Premise: "If A is brought about, good consequences will (plausibly) occur"* and *Conclusion: "Therefore, A should be brought about"*, with the following critical questions (CQs): *CQ1: "How strong is the likelihood that the cited consequences will (may, must) occur?"*; *CQ2: "What evidence supports the claim that the cited consequences will occur and is it sufficient to support the strength of the claim adequetly?"*; *"CQ3: Are there opposite consequences (bad as opposed to good) that should be taken into account?"*.

Wagemans proposed the Periodic Table of Arguments (PTA) [7] (see Figure 1 for the first iteration of the table), which is based on an *a priori* constrained set of possible combinations between different characterisations of argument: subject vs predicate arguments, first vs second order arguments, and argument substance. *Subject arguments* are arguments whose premise and conclusion have the same predicate but different subject (e.g. "Abortion should be illegal, because murder is illegal."), whereas *Predicate arguments* are arguments whose premise and conclusion share the same subject but have different predicates (e.g. "The death penalty should be abolished, because the death penalty carries the risk of ruining innocent people lives."). *First-order arguments* contain simple statements which cannot be split further while *Second-order*

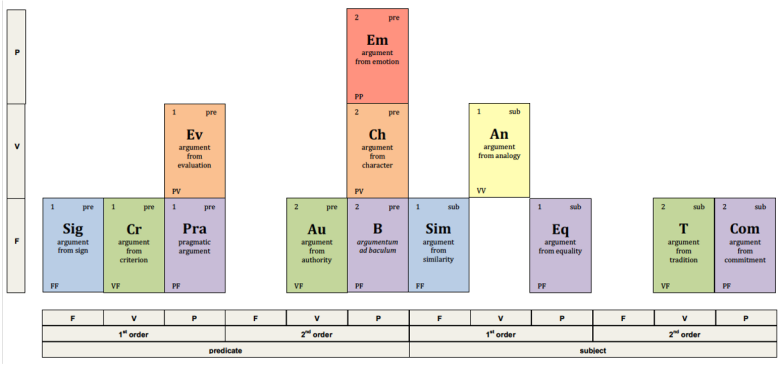

Figure 1: The first Periodic Table of Arguments with different characterisations of arguments: subject vs predicate, 1st vs 2nd order, and combinations of three types of propositions: F(act), V(alue), P(olicy) [7].

arguments are epistemological in nature and, containing at least one complex statement whose subject can be broken down, they can embed first order arguments (e.g., "We should wear masks in public because WHO suggests it."). Finally, the argument's statements are categorised as propositions of *policy* (if they express that an act should be performed), of *value* (if they contain an evaluative judgment based on a definition or assessment criteria, such as ethical judgements), or of *fact* (if their veracity can be verified empirically). We note that the two taxonomies are not incompatible and the PTA can incorporate existing schemes, for example Walton's *argument from sign* corresponds to *1-pre-FF* in the PTA, as can be seen in Figure 1.

Several works have focused on classifying arguments based on the argument scheme they instantiate using features extracted from text [8, 9]. [10] proposed guidelines for annotating argument schemes using a taxonomic hierarchy of schemes and the semantic properties of premises/claims, while [11] advocated for the creation of argument scheme templates representing clusters of schemes. Visser et al. [12] annotated election debates using Walton's schemes and Wagemans's PTA and provided an overview of the correspondence between the results obtained with the two taxonomies. In contrast, our aim is to reconcile the two approaches of classifying schemes, leveraging the strengths of both in order to develop a hybrid classification that can be applied to ethical debates so as to provide transparent and rational reasoning.

3. From Argument Classification to Reconciling Taxonomies

We make first steps towards reconciling the two taxonomies by classifying NL arguments using Walton and Wagemans, following the method of Visser et al. [12] to manually annotate arguments, but we focus on ethical debates (e.g. "Pro-life vs Pro-choice: Should abortion be legal?", "Should an Artificial General Intelligence be created?", "Should all humans be vegans?"). The most common schemes identified are: *argument from consequences*, *argument from values*, *argument from analogy*, *argument from example*, *practical reasoning*, while the majority of arguments we identify using Wagemans' classification are first-order, predicate arguments.

Classifying an argument from the debate *Pro-life vs Pro-choice.* Consider argument *A:*

Access to legal abortion improves the health and safety of pregnant people. In our analysis, we first determine that argument A is an enthymeme and assume the complete argument to be A'*: Access to legal abortion improves the health and safety of pregnant people, so pregnant people should have the right to choose abortion.*

To annotate with Walton's taxonomy, we use the Argument Scheme Key (ASK), a series of disjunctive choices based on the distinctive features of argument schemes which also groups scheme types that share particular characteristics [12]. The ASK path used to classify argument A' is as follows, with the answer in brackets: Argument relies on a source's opinion or character (No); Conclusion is about a course of action (Yes); Argument focuses on the outcome of the action (Yes); Conclusion promotes a positive outcome (Yes); Course of action assists someone else (No); Course of action promotes a goal (Yes), resulting in an *argument from positive consequences*.

To annotate with Wagemans' taxonomy, we consider the three characterisations of argument. In particular, we determine that A' is a first-order, predicate argument as the premise and the conclusion share the same subject (i.e. "access to legal abortion"). Note that A' needs to be rephrased slightly in order to match the definition of a predicate argument. Lastly, we observe that the conclusion of A' is a proposition of *policy* ("pregnant people should have the right ..."), while the premise is a *fact* ("legal access improves the health and safety of pregnant people"), thus the argument is of type *1-pre-PF*, equivalent to a *pragmatic argument*.

This example offers insights into our approach towards a hybrid taxonomy: firstly, observing the co-occurrences of argument schemes in both classification systems allows us to detect correspondences between the two in order to reconcile them. For example, Walton's *argument from positive consequences* is often classified as *1-pre-PF* using Wagemans' PTA, which was also observed in [12]. Secondly, comparing and contrasting the annotation guidelines of each taxonomy helps us reflect on them and their usefulness. For instance, deciding if an argument is first or second order is a criterion in both taxonomies, which indicates that this distinction is necessary and should be included in the criteria of classification in our hybrid taxonomy.

We also take inspiration from [13] that developed the hybrid scheme *argument from action*, which (partly) incorporates the schemes *argument from positive/negative consequences*, *argument from positive/negative values*, and *practical reasoning*. Grouping schemes in this manner can be a helpful tool in decreasing the number of argument schemes considered and a step towards developing our hybrid classification. Thus, we also consider argument A' as an *argument from action*, although the scheme is not part of the two taxonomies analysed.

Linking arguments using critical questions. Consider argument *A: A vegan society would cause the least harm to wildlife, so all humans should go vegan* and argument *B: The risk of death of wildlife increases during the transport of food, especially when the vegan food travels for thousands of miles.* Argument A is an *argument from action* which has 16 critical questions [13], one of them being *CQ: "Assuming the circumstances, does the action have the stated consequences?"*. We observe that argument B can serve as an answer to the CQ of argument A and can be seen as a counter-argument to it. In addition, argument B is as an *argument from danger*.

An important step in reconciling Walton's argument schemes and Wagemans' PTA is to identify the strengths of the two taxomies. The former is more comprehensive, while the latter is more practically useful and can be seen as an intermediate between the semantic detail of Walton and the relation between premise-conclusion used in argument mining approaches.

Critical questions represent a key aspect of argument schemes, as they can be pointers to counter-arguments or supporting arguments, since the argument answering a critical question attacks/supports the original argument by virtue of the critical question; we aim to include them in our hybrid classification and apply them to ethical debates. Part of our reconciliation process is also to decrease the number of argument schemes. Indeed, the next step of our study will focus on exploring whether other schemes can be grouped together, similarly to *argument from action* [13]. The clustering nature of the ASK algorithm along with the criteria of the PTA can be used to create a reduced, but still broad, number of argument types.

4. Conclusion

This paper describes an initial step towards realising the long-term research goal of supporting dialogue between humans and between humans and AI systems. We focus on argument annotation using Walton's argument schemes and critical questions as well as Wagemans' Periodic Table of Argument, in order to develop a new, evolved framework, specialised in ethical reasoning. We believe this is an important first step in identifying schemes and scheme taxonomies specialised for ethical reasoning and that can support dialogical scaffolding of both human and artificial agent reasoning. Our approach goes beyond standard annotation approaches for argument mining and proposes a semantically richer approach to argument classification through tools of argumentation.

References

[1] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works, Transactions of the Association for Computational Linguistics 8 (2020) 842–866.
[2] S. Modgil, Dialogical scaffolding for human and artificial agent reasoning, in: Proceedings of the 5th International Workshop on Artificial Intelligence and Cognition, AIC, 2017, pp. 58–71.
[3] A. Matthias, The responsibility gap: Ascribing responsibility for the actions of learning automata, Ethics and information technology 6 (2004) 175–183.
[4] D. Hadfield-Menell, A. D. Dragan, P. Abbeel, S. Russell, Cooperative inverse reinforcement learning, CoRR abs/1606.03137 (2016).
[5] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2019) 765–818.
[6] D. Walton, C. Reed, F. Macagno, Argumentation schemes, Cambridge University Press, 2008.
[7] J. H. M. Wagemans, Constructing a periodic table of arguments, 2016.
[8] V. W. Feng, G. Hirst, Classifying arguments by scheme, in: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 987–996.
[9] J. Lawrence, C. Reed, Argument mining using argumentation scheme structures, in: Proceedings of Computational Models of Argument (COMMA), 2016, pp. 379–390.
[10] E. Musi, D. Ghosh, S. Muresan, Towards feasible guidelines for the annotation of argument schemes, in: Proceedings of the Third Workshop on Argument Mining, ArgMining@ACL, 2016, pp. 82–93.
[11] D. Liga, M. Palmirani, Argumentation schemes as templates? Combining bottom-up and top-down knowledge representation, in: Proceedings of the 20th CMNA workshop, 2020, pp. 51–56.
[12] J. Visser, J. Lawrence, C. Reed, J. Wagemans, D. Walton, Annotating argument schemes, Argumentation 35 (2021) 101–139.
[13] K. Atkinson, T. J. M. Bench-Capon, Action-based alternating transition systems for arguments about action, in: Proceedings of the 32nd AAAI, 2007, pp. 24–29.