Debating Ethics: Enhancing Human & Human-AI Dialogue

Elfia Bezou Vrakatseli
elfia.bezou_vrakatseli@kcl.ac.uk
King's College London
London, UK

ABSTRACT

This research proposes the exploration and analysis of natural language texts, specifically of ethical debates, via tools of argumentation to the end of supporting dialogical exchanges between humans and between humans and AI systems, with respect to transparent and rational reasoning. Specifically, it proposes the deployment of argument schemes and critical questions as a semantically richer manner of argument classification. To optimise this process, this project additionally proposes the generation of a new scheme classification system, which is based on the existing ones but also consists of specialised schemes for ethical reasoning, as well as meta-level schemes. Subsequently, the process of annotating natural language arguments with these argument schemes and critical questions will output datasets, in the form of debates on society's ethical matters, resulting in the creation of a new corpus that will enable the identification and extraction of arguments from texts. This project aims to support the development of technologies to enhance both the natural language processing needed to support human-AI dialogue, as well as for scaffolding human-human dialogue.

CCS CONCEPTS

• **Computing methodologies** → *Nonmonotonic, default reasoning and belief revision.*

KEYWORDS

argumentation, argument schemes, ethics, AI, dialogue

ACM Reference Format:
Elfia Bezou Vrakatseli. 2023. Debating Ethics: Enhancing Human & Human-AI Dialogue. In *10th ACM Celebration of Women in Computing womENcourage.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION & BACKGROUND

Artificial Intelligence (AI) is becoming more and more powerful, efficient, and autonomous. Therefore, the need for safe and trustworthy systems increases as well; systems for which there is assurance about the correctness of their behaviour and that inspire confidence for their users about their decision-making process. Interacting and communicating with AI systems is, thus, of vital importance and at the core of the current research in the field of AI. Dialogue is a powerful tool of interaction and communication and multiple applications of AI make direct use of it (e.g., conversational agents). As dialogues consist of informational exchanges, they enable the understanding of the parties involved, while also facilitating joint reasoning.

This project uses tools from argumentation to enhance dialogues between humans and AI systems. Argumentation, stemming from speech communication and non-monotonic logic [6], focuses on dialogical exchanges and their underlying reasoning. Argumentation can, thus, support AI-human dialogue for joint reasoning, for AI to influence human reasoning (and vice versa), and for explanation (e.g., to explain the decision-making of recommender systems [5]). Overall, the communication and joint reasoning of AI systems and humans leverages the strengths of AI along with those of human reasoning. One of the main challenges for safe and trusted systems is value alignment; i.e., to ensure that the values of current AI systems, and of future AI systems, align with humans [4]. Many state-of-the-art approaches focus on cooperative inverse reinforcement learning (CIRL) [1], which implies requirements for joint deliberation and reasoning. In order to enable those, we need human-AI dialogue, with respect to ethical and moral issues. The focus on ethical debates is crucial for value alignment and it can result in supporting the decision-making process on ethical issues as, for instance, in the case of autonomous cars [3] (ethics for AI).

To the end of enhancing the dialogue between humans and humans and AI systems, natural language (NL) texts, in particular arguments from ethical debates, are analysed with the use of argument schemes and critical questions. Argument schemes represent stereotypical patterns of reasoning and consist of a set of premises and a conclusion and critical questions identify points of possible arguments that attack or support the argument in question. Walton proposed over 60 argument schemes with corresponding sets of critical questions which are used to evaluate the strength of an argument [10]. 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?"*.

Argument schemes and critical questions have been traditionally used in the development of formal philosophy in the area of formal argumentation for individual-agent reasoning and for joint deliberation (dialogue). These same schemes, developed in a more human-oriented informal philosophy can be used to guide the evolution of a framework of arguments that captures the reasoning

womENcourage 2023, September 20–22, 2023, Trondheim, Norway

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

process for individual agents and for agents engaging in dialogue. Schemes are designed for the purpose of dialogical exchange, for instantiation by humans using NL; for AI systems in the machine learning (ML) era, utterances in NL are generated from datasets. I argue, thus, that appropriately annotated datasets will enable the instantiation of schemes by AI systems. The idea is to develop argument schemes and critical questions that achieve to be effectively instantiated by AI systems as well. Therefore, there is a need for schemes that are, firstly, specialised for ethical reasoning because this gives appropriate guidance to humans (as it has been done, for example, in the medical domain [7]). Secondly, for schemes to provide effective guidance, they also need to capture multiple patterns of reasoning, including modes of 'meta-argumentation' [2], which are not currently captured by the existing schemes. This research focuses on developing (1) object-level schemes specialised for ethical reasoning and (2) meta-level schemes as this is necessary for humans to use them effectively (as shown in [7] and [2]) but also because it enhances argument mining (AM) techniques by providing a semantically richer account of the structural relationships currently exploited in AM. In the latter case, as a result of enhancing AM, the use of AI can also be used to support human-human dialogue and ethical reasoning (AI for ethics).

2 METHOD & RESULTS

The question being addressed is, thus: how can argument schemes and critical questions be useful to support ethical debate and dialogue between humans and between humans and AI systems? The first step of addressing it was to examine how existing argument schemes and scheme taxonomies can be used to classify arguments in debates with an ethical dimension. This was done by a multi-leveled analysis, an elaborate labeling process of NL arguments, and ML experiments. The NL arguments were taken from ethical debates on the user-generated platform Kialo[1], using the two most used argument scheme classifications: Walton's argument schemes [10] and Wagemans' Periodic Table of Argument (PTA) [9]. The former is more comprehensive, while the latter is more practically useful and can be seen as an intermediate between the semantic detail of the former and the relation between premise-conclusion used in AM. In order to increase confidence in the described classification process, a second annotator labeled the arguments from three of the debates. The classification task (i.e., the labeling) proved to be a task of multiple difficulties, as many previous works have reported (e.g., [8]), and the inter-annotator agreement was found to be moderate.

The outputs of the labeling process were combined into a dataset that was used for ML experiments. The eight most predominant schemes were chosen as classes, with the annotated arguments constituting the instances. The goal of this process was to further evaluate the usefulness of the existing argument schemes and their taxonomies. The models BERT and RoBERTa were used for a multi-class classification task. Their performance confirmed the insufficiency of the existing classification systems.

The current argument schemes and scheme taxonomies do not suffice, thus, for accurately classifying arguments. So, the following step of the method focuses on building a new classification system specialised for ethical reasoning, by (1) reconciling the two aforementioned taxonomies, (2) generating hybrid schemes that encompass a number of existing schemes, and (3) generating new schemes. These new schemes are (1) object-level schemes specialised for ethical reasoning (e.g., *action scheme appealing to rights*) and (2) meta-level schemes, tailored to make claims about preferences on the object-levels schemes and enable commentary on their reasoning.

Besides the development of said new taxonomy, this research will also output a corpus, annotated with argument schemes, which can be used to train models for AM, and enhance the identification of NL arguments that are used in ethical debates in NL texts. Future steps include incorporating critical questions in the automated methods, developing an online platform to crowdsource annotations based on this new classification system, and devising apposite techniques of consolidating the outputs of intermediate steps to the end of enhancing the human-AI dialogue.

3 CONCLUSION

This paper describes the initial steps towards supporting the development of technologies for human-human and human-AI dialogue via the deployment of argument schemes and critical questions for argument classification. The novelty of this approach lies in going beyond standard annotation approaches for argument mining, proposing a semantically richer approach to argument classification through tools of argumentation. The generation of a well-tailored taxonomy for ethical arguments will enhance ethical debate in order to guide both humans and AIs (ethics for AI and AI for ethics) in building and challenging arguments related to ethical issues.

ACKNOWLEDGMENTS

I would like to thank Oana Cocarascu and Sanjay Modgil for their guidance, valuable insights, and suggestions in this project.

REFERENCES

[1] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. Cooperative Inverse Reinforcement Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3916–3924.

[2] Nadin Kökciyan, Isabel Sassoon, Elizabeth Sklar, Sanjay Modgil, and Simon Parsons. 2021. Applying metalevel argumentation frameworks to support medical decision making. *IEEE Intelligent Systems* 36, 2 (2021), 64–71.

[3] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6, 3 (2004), 175–183.

[4] Sanjay Modgil. 2017. Dialogical Scaffolding for Human and Artificial Agent Reasoning.. In *AIC*. 58–71.

[5] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, and Francesca Toni. 2020. Argumentation as a framework for interactive explanations for recommendations. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, Vol. 17. 805–815.

[6] Iyad Rahwan and Guillermo R Simari. 2009. *Argumentation in artificial intelligence*. Vol. 47. Springer.

[7] Pancho Tolchinsky, Sanjay Modgil, Katie Atkinson, Peter McBurney, and Ulises Cortés. 2012. Deliberation dialogues for reasoning about safety critical actions. *Autonomous Agents and Multi-Agent Systems* 25, 2 (2012), 209–259.

[8] Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2021. Annotating argument schemes. *Argumentation* 35, 1 (2021), 101–139.

[9] Jean Wagemans. 2016. Constructing a periodic table of arguments. In *Argumentation, objectivity, and bias: Proceedings of the 11th international conference of the Ontario Society for the Study of Argumentation (OSSA), Windsor, ON: OSSA*. 1–12.

[10] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

[1] https://www.kialo.com/