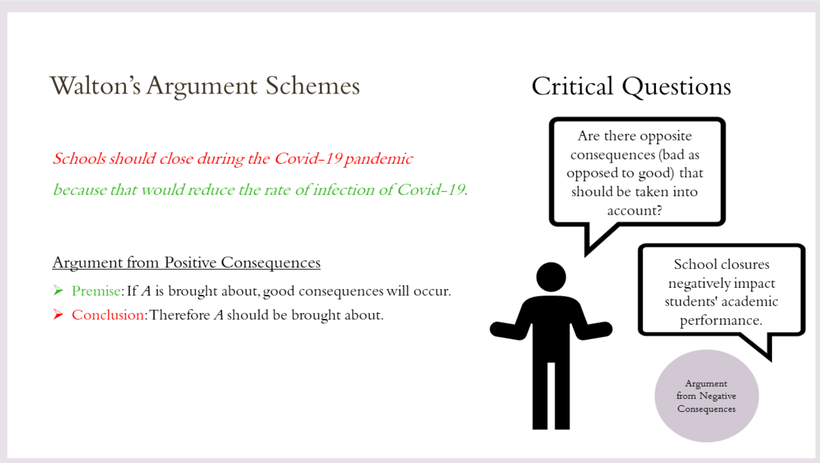
Debating Ethics

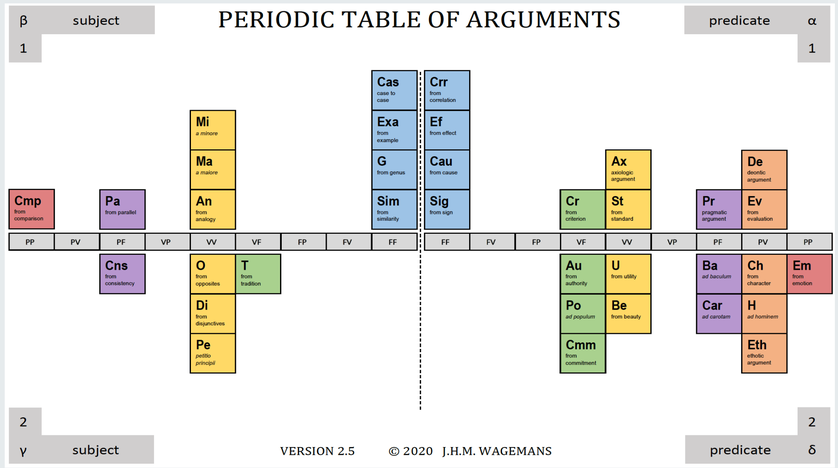
Using Natural Language Datasets to Support Human and AI debate

Elfia Bezou-Vrakatseli, Oana Cocarascu, Sanjay Modgil

SAFE & TRUSTED AI
UKRI CENTRE FOR DOCTORAL TRAINING

KING'S College LONDON





Introduction

The project proposes the exploration and analysis of natural language (NL) texts, specifically of **ethical debates**, via tools of argumentation, in particular: **argument schemes** and **critical questions**. It aims to develop new natural language datasets and AI algorithms through defining semi-automated approaches for the identification and extraction of argument schemes. These new datasets will be in the form of philosophical debates on society's ethical and moral issues and will serve the purpose of creating a new corpus that will enable the automatisation of argument mining from texts. Subsequently, the obtained argument schemes will be used to **support dialogical exchanges** between humans and AI systems with respect to **transparent and rational reasoning**. This project aims to guide both humans and AI systems (ethics for AI and AI for ethics) in building and challenging arguments related to such societal issues.

Research Questions

- How can one reconcile the existing various argument classifications to devise a theoretically well-founded, as well as practically useful, hybrid classification, which can be specialised for application in ethical debates?
- Given such a classification system, what is an effective way to develop a semi-automated way to map NL arguments, used in ethical debates, to argument schemes?
- How can the outputs of the previous techniques be exploited to support the dialogue both between humans and between humans and AI systems, with emphasis on matters of ethics?

Method & Preliminary Results

The initial step consists of annotating user-generated, NL arguments of 22 ethical debates from the *Kialo* platform, using two taxonomies, Walton's argument schemes & Wagemans' Periodic Table (PTA), with the following guidelines.

- **Walton**: Argument Scheme Key (ASK) – An extensive series of disjunctive choices based on the distinctive features of argument schemes which also groups scheme types that share particular characteristics.
  - Most common schemes identified: *argument from example*, *argument from values*, *argument from consequences*, *argument from cause to effect*, *argument from expert opinion*, *argument from alternatives*, *argument from analogy*.
- **Wagemans**: 3 characterizations of the PTA
  1. First vs Second Order Argument
  2. Predicate vs Subject Argument
  3. Argument Substance: policy, value, fact
  - The majority of the arguments identified belong in the first quadrant (i.e., first-order, predicate arguments).

➢ Observing the co-occurrences of schemes in both taxonomies allows to detect correspondences; e.g., Walton's *argument from positive consequences* is often classified as *1-pre-PF* using Wagemans' PTA.

➢ Comparing and contrasting the annotation guidelines of each taxonomy helps reflect on them; e.g., deciding if an argument is first or second order (source-based or not) is a criterion in both taxonomies, which underlines the significance of said distinction.

Discussion

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.

The use of argument schemes and critical questions offers a *semantically* richer approach to inner- and inter- argument classification:

- Premises support a conclusion *by virtue* of instantiating a scheme.
- Support/attack relations are instigated *in response* to critical questions.

The reconciliation of the two predominant taxonomies in a hybrid one leverages the strengths of both, while identifying the schemes particular to ethical reasoning.

- Walton: more comprehensive
- Wagemans: more practically useful and an intermediate between the semantic detail of Walton and the relation between premise-conclusion used in argument mining approaches

Following steps

- Critical questions will be incorporated.
- Scheme groupings will be formed: 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.
- New schemes will be generated.
  - Object-level schemes specialised in ethical reasoning
  - Meta-level schemes: schemes of another level that enable commentary on object-level reasoning

Conclusion

This poster describes the initial step towards realising the long-term research goal of supporting dialogue between humans and between humans and AI systems. The first step consists of developing a new taxonomy, as well as new argument schemes, specialised in ethical reasoning. To this end, arguments from ethical debates were annotated using Walton's and Wagemans' taxonomies. This is an important step in identifying schemes and taxonomies specialised for ethical reasoning. The use of argument schemes and critical questions goes beyond standard annotation approaches for argument mining and proposes the use of argumentation to achieve a semantically richer approach to argument annotation.