Ethical-related issues of AI have affected the safety and fairness of machine applications in recent years. Most ethical thought is captured in natural language, from regulatory and legal codes to philosophical frameworks. Thus, reasoning ethically can largely increase the transparency of a machine system. However, it requires the coordination and integration of complex inference steps. There is also a gap between symbolic and NLI that needs models which are specialised to ethical reasoning.

This project introduces a new dataset and provides multi-fact explanations for commonsense ethical statements that combine core moral foundation knowledge and world knowledge. This project also presents a new data annotation method. Instead of a costly, fully human annotation process, this project applies a hybrid system that combines retrieving model (Valentino et al. 2021) and differentiable logic solver (Weber et al. 2019) to automatically annotate data from GenericsKB-Best (Bhakthavatsalam et al. 2020) with semantic parsers. These resources provide new training and testing data for developing ethical reasoning ability for multi-hop inference models.

The main contribution of this project would have 1) creation of a new ethical NLI gold-standard with multi-fact explanations. 2) construction of a differentiable NLI solver which integrates logic programming to explanatory inference constraints to support zero-shot ethical reasoning. 3) design of an ethical Natural language Inference (NLI) framework which supports explainable inference over abstract ethical statements.

The method which applied to address this problem is shown in figure 1 as: 1) implementing SCAR(Valentino et al. 2021) to retrieve grounding facts from the knowledge base. 2) apply two parsers to automatically parse statements and facts into first-order logic forms to build the knowledge base for the solver. 3) apply a pre-trained sentence encoder with a weak unification mechanism in a Prolog solver to find the target object.

A screenshot of a computer

Description automatically generated with low confidence

Figure 1

There might be some limitations. The whole process still needs human annotators to validate the result and there is no training data, and the model may be very sensitive to the noise data. Furthermore, GenericsKB-Best contains over 1 million facts, and the solver may be broken when the input data is too large. The current evaluation strategy is by comparing the proof score and proof tree to construct the final success proof. Some examples of a solved proof tree are given in Figure 2. The next step would be evaluating how the model will perform by adding discard information and what will happen to the solver to improve the overall process.

Diagram

Description automatically generated

Figure 2

Reference:

M. Valentino, M. Thayaparan, D. Ferreira, and A. Freitas. 2022. Hybrid autoregressive inference for scalable multi-hop explanation regeneration. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 1403–11411.

S. Bhakthavatsalam, C. Anastasiades, and P. Clark. 2020. GenericsKB: A Knowledge Base of Generic Statements. arXiv preprint arXiv: :2005.00660.

L. Weber, P. Minervini, J. Münchmeyer, U. Leser, and T. Rocktäschel. 2019. NLProlog: Reasoning with Weak Unification for Question Answering in Natural Language. In ACL 2019.