Bridging Ethics and AI: A Path to Moral Machines

Aisha Aijaz

IIIT, Delhi, New Delhi, India
aishaa@iiitd.ac.in

Abstract

Ethics is ubiquitous in most domains, requiring much deliberation due to its philosophical nature. Varying views often lead to conflicting courses of action where ethical dilemmas become challenging to resolve. The major driving forces to make such a decision can be discretized and simplified to provide an indication of the most ethical course of action in a context. Given the parallel ubiquity of AI systems in these domains, it becomes increasingly imperative to work towards building inherently ethical AI that holds the ability to reason morally. This work proposes the use of knowledge representation and neurosymbolic techniques to develop resources for inherently ethical AI. It presents a three-phase framework towards bridging the path to moral machines: (a) Applied Ethics Ontology to make explicit the abstract concepts and relationships, (b) Dataset and graph generation using LLMs to develop a benchmark data store for ethical reasoning, and a (c) Case-based Reasoning algorithm to implement the philosophical concept of casuistry to make moral judgments and resolve ethical dilemmas.

Introduction

Ethics is ubiquitous in most domains, such as medicine, business, and education, requiring much deliberation due to its philosophical nature. Varying views often lead to conflicting courses of action where ethical dilemmas become challenging to resolve. Many factors contribute to such a decision; however, the major driving forces can be discretized and thus simplified to provide an indication of the most ethical course of action in a particular context.

Given the parallel ubiquity of AI systems in these domains, it becomes increasingly imperative to work towards building inherently ethical AI. AI that is more than simply intelligent, but also holds the ability to reason morally. There is a common agreement to build more moral machines (Gunkel 2012; Formosa and Ryan 2021), and my interdisciplinary research aims to fill this gap.

Modeling ethics in AI systems is not new, however, most approaches are narrow in their scope (Awad et al. 2018; Anderson and Anderson 2018). These valuable contributions rely on data and frameworks that reflect acute moral reasoning rather than a deeper understanding of ethical theory and its general application. At this time, there are few resources for the development of ethical AI that may be applied generally (Dehghani et al. 2008; DeBellis 2018). Perhaps because of how daunting the task actually is.

The challenges of building a general, inherently ethical AI include but are not limited to the lack of agreement on ethical schools of thought, the complexity of context, the precedence of different factors affecting an ethical decision, and finding usable data that involves context information with corresponding ethical considerations. For truly inherently Ethical AI systems, we need to develop AI that can make sense of moral and contextual information alike. This is not easy for rational humans let alone machines, however, the answer to this mystery lies in epistemology. Humans make decisions based on what they know. This is why the use of knowledge representation and neurosymbolic techniques to develop resources for inherently ethical AI is the ideal choice.

Applied Ethics Ontology

Philosophers generally do not agree on which theories may be superior (Bergmann and Kain 2014), but there is some understanding of how to approach certain ethical issues in particular domains (Holmes 2018). As proposed by Demarco in his work, an outlook that considers ethical principles, rules, and judgments to determine which aspect of morality would take precedence in a particular context would be at the core of applied ethics (DeMarco 1997). Due to this general understanding, *applied* ethics becomes more comprehensible to machines than normative ethics.

Thus, an explicit taxonomy of applied ethics is required as an invariable source of ethics theory for computer systems to refer to and infer from when provided some context. This requirement may be fulfilled by a symbolic representation or an ontology. There have been efforts in this space, such as Debellis' work (DeBellis 2018), which implements a Universal Moral Grammar (UMG) via ontology modeling. However, it does not involve contextual cues that are vital when applying ethics theory in practical scenarios.

To build this ontology, I first developed a taxonomy that would be the framework for applying ethics in real-world events, in collaboration with domain experts. This would, in turn, help machines *understand* the general application of ethics. It includes two modules: Applied Ethics and Event

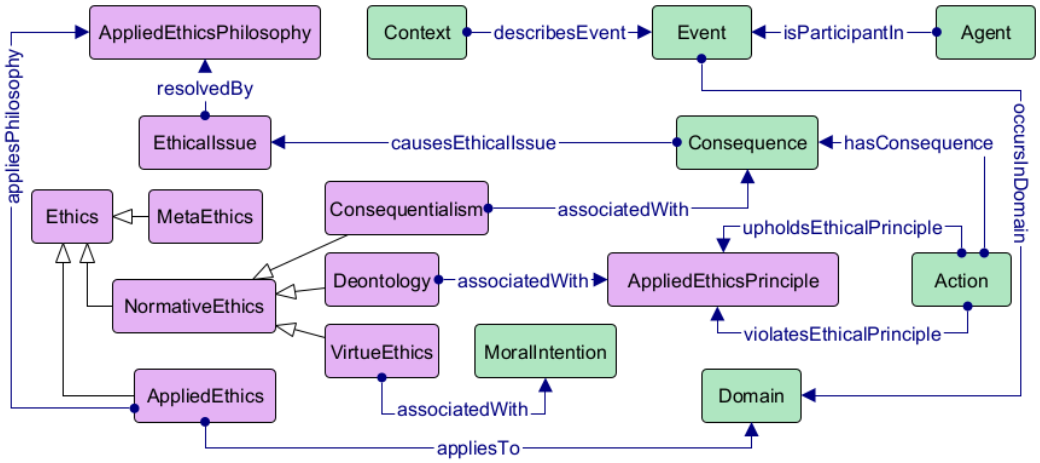

Figure 1: The Applied Ethics ontology with the two interacting modules (purple for Applied Ethics and green for Event Context.)

Context (see Fig.1).

The Simplified Agile Methodology for Ontology Development (SAMOD) (Peroni 2016) was used. This focuses on domain expertise, modular development, extensive testing, FAIR compliance (Wilkinson et al. 2016), and documentation [1]. In accordance with this methodology, the ontology has been model, query, and data tested.

A real-world bioethics scenario was modeled using the ontology to demonstrate its validity. This was reasoned over by special SWRL rules to indicate whether an action would lean towards morally right, wrong, or grey. The ontology was able to capture ethical and contextual information effectively and was able to offer accurate results to query and data testing. The model testing produced no bugs or pitfalls in the ontology.

Applied Ethics Dataset and Knowledge Graph

The next step was to create a dataset that involved context information with corresponding ethical information for events. This dataset aims to be a benchmark in the area of inherently ethical AI systems, backed by a well-established ontology and a parallel knowledge graph.

The Applied Ethics Dataset has 10k cleaned, long-paragraph style instances of raw data scraped from various relevant Subreddits, without any data augmentation. To build this dataset, we leveraged both knowledge representation and LLMs for the most efficient development of reusable and discrete data for the highly abstract domain of ethics. The following methodology was used:

- Identifying key features: Top-level classes from the ontology were chosen as features.
- Finding real-world cases: Web-scraping raw real-world data for use cases which were then further summarized using templates.
- Populating the dataset: Used an LLM to populate explicitly mentioned features and more fine-tuned NLP techniques to learn/infer implicit features from the text.
- Preprocessing the dataset: Applied standard preprocessing techniques to ensure the completeness and accuracy of the dataset.
- Dataset Evaluation: Includes quantitative and qualitative evaluation. The former involved statistical validation of the dataset, and the latter involved manual human verification of distributed samples by domain experts.

This dataset was saved in .csv format and converted to triples to create a knowledge graph using the OntoRefine tool. Through this process, I have developed a relational and graph dataset that may serve as benchmark resources for neurosymbolic applications in the space of inherently ethical AI.

Casuistry for AI

Casuistry is a branch of ethics that falls under the category of analogical reasoning (Richardson 2018). Given a certain case where an ethical dilemma is evident, it may be resolved by considering other cases with similar context. To do this, I applied a custom version of case-based reasoning (CBR) to the Applied Ethics dataset that takes advantage of semantic text similarity using word embeddings (Han et al. 2021).

A statistically significant similarity to a predefined number of instances from the dataset would allow us to determine the most ideal course of action, based on a threshold value. For example, if 6 of the 10 most similar cases to the case in question have determined that prescribing an addictive painkiller to a patient is unethical, then the automated decision maker would arrive at a similar decision with a justifiable reason to do so. The performance of the model may

[1] see https://github.com/kracr/applied-ethics-ontologyGitHub Repository and https://purl.org/appliedethicsontology/documentationOntology Documentation for the Applied Ethics Ontology.

be improved using the knowledge graph as it would provide additional information about the explicit relationships between the different classes. This would be part of my future work.

Future Work

For my thesis, there are a few projects I would like to explore that would stem from the work I have done so far.

1. Question Answering System: A QA system would serve as a follow-up to the previously discussed casuistry task, which would act as an interface to visually display, in real time, how the CBR algorithm would work to resolve ethical dilemmas.
2. Domain-specific applications: We can extrapolate the general work done to specific domains to be more accessible to the issues that occur in those areas. This would be a top-down chronology of the work, going from general Applied Ethics applications to specific domains such as healthcare and business.
3. Explainable Ethical AI System: I would also like to make the system explicitly explainable. This would mean studying the various courses of action and presenting an ethical score based on various parameters to judge the best way forward. The parameters chosen could then be used to provide a natural language explanation to justify the AI's decisions.

Acknowledgments

I would like to thank Dr. Raghava Mutharaju for his guidance and supervision of my research and Dr. Manohar Kumar for his valuable expertise in the domain of ethics.

References

Anderson, M.; and Anderson, S. L. 2018. GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1): 337–357.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature*, 563(7729): 59–64.

Bergmann, M.; and Kain, P. 2014. Challenges to Moral and Religious Belief: Disagreement and Evolution.

DeBellis, M. 2018. A Universal Moral Grammar (UMG) Ontology. *Procedia Computer Science*, 137: 242–248.

Dehghani, M.; Tomai, E.; Forbus, K. D.; and Klenk, M. 2008. An Integrated Reasoning Approach to Moral Decision-Making. In *AAAI*, 1280–1286.

DeMarco, J. P. 1997. Coherence and applied ethics. *Journal of applied philosophy*, 14(3): 289–300.

Formosa, P.; and Ryan, M. 2021. Making moral machines: why we need artificial moral agents. *AI & society*, 36(3): 839–851.

Gunkel, D. J. 2012. The machine question. *Critical perspectives on AI, robots, and ethics*, 5.

Han, M.; Zhang, X.; Yuan, X.; Jiang, J.; Yun, W.; and Gao, C. 2021. A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency and Computation: Practice and Experience*, 33(5).

Holmes, R. L. 2018. Introduction to Applied Ethics.

Peroni, S. 2016. SAMOD: an agile methodology for the development of ontologies. In *Proceedings of the 13th OWL: Experiences and Directions Workshop and 5th OWL reasoner evaluation workshop (OWLED-ORE 2016)*, 1–14.

Richardson, H. S. 2018. Moral Reasoning. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2018 edition.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.