A Formal Language of Ethics and Causation

Vu Le TJ Schlueter (CS Advisor) Marianna B. Ganapini (PHL Advisor)

Computer Science Department, Union College

Question

How can computer DO Ethics?

Introduction and Motivation

Artificial Intelligence has risks. With its increasing application in critical areas such as healthcare, self-driving car, and policing, it is harder for humans to keep them under control. This means that we can't keep chasing around and cleaning up after its mess. What can we do? Well, instead of humans handling the risks, we let the AI handle the risks themselves.

Claim

In order to model ethical scenarios descriptively and objectively, we first need a rigorous, ethics-free model of the causal world.

The overall goal is to build an AI agent that can reason and make ethical decisions by itself. Existing approaches have faced problems: lack of transparency, human biases, oversimplification, subjective, etc. I adopt a rule-based approach, focusing on representation of ethical principles and automated reasoning. Specifically, I aim to design a formal language of ethics and causation, for encoding ethical scenarios unbiasedly and reasoning about them.

The Ideal Ethical Formalism

- Universal
- Expressive enough to describe any ethical scenarios
- Explicit reasoning process
- Meta-ethically neutral

Causation in Ethics

The concept of causation has implications to ethical judgments, such as moral responsibility. A common line of reasoning is the **Counterfactual Theory of Causation**. Event A causes event B if and only if:

- A is necessary for B OR
- The occurrence of A makes a difference for the occurrence of B OR
- Had A not occurred, B would not occur.

But this has problems!

1. Preemptive Causation: The Hitman/Poison

Suppose I poisoned you, but the poison takes 2 days to kill you. In the mean time, a hitman from afar shot you right in the head, killing you instantly. Did the hitman cause you to death? **NO**!

Counter-factually, had the hitman not shot you, you'd still be dead anyways, because of the poison. The hitman made no difference, therefore not responsible for your death.

2. Duplicative Causation: Voting

5 participants A, B, C, D, E in an election. P is elected if 3 out of them voted. A, B, C, D actually voted for P. Who is responsible? **NONE!**

Is A responsible? Had A not voted for P, P would still be elected, because B, C, D is enough. Therefore, A did not make a difference, and not responsible for P's election. Same for B, C, D.

Claim

In these examples, counterfactual reasoning conflates two notions: lawful causality vs actual causation

Actual Causation

We'll introduce another theory of actual causation: **Necessary Element of Sufficient Set** (NESS)

1. Definition: A condition c is a cause for an event E if and only if c is an element of some set S such that

- S is actual
- S is sufficient for E
- S is minimal

Each condition c only has to be necessary for some set S, instead of being necessary for the outcome E like in the counterfactual account.

2. Application: There are four sufficient sets in the voting scenario

- $\{A, B, C\}$
- $\{B, C, D\}$
- $\{A, C, D\}$
- $\{A, B, D\}$

Thus, A, B, C, D are all contributing causes because each is an element of some sufficient set.

3. Quantified Causal Contribution:

$$\mathit{weight}(c) = \frac{\text{Number of sets } S', c \in S'}{\text{Total number of sets } S}$$

In the voting example,

$$\mathit{weight}(A) = \mathit{weight}(B) = \mathit{weight}(C) = \mathit{weight}(D) = \frac{3}{4}$$

If $\mathit{weight}(c) = 1$, c is present in *all* sufficient sets. c is necessary for all the sets and thus the outcome E itself. This collapses into counterfactual reasoning: Had c not occurred, none of the sets would be sufficient, and the outcome E would not occur.

Overview of technologies

- Answer Set Programming: logic programming
- Event Calculus: the logic of change and action

Result

- A formal language of ethics and causation
- A framework for encoding ethical scenarios, running analysis, and testing them
- Sample ethical dilemmas

Language Components

- Causal rules: knowledge of causal world
- Event axioms: allowing the world to evolve according to described causal rules
- Causal axioms: causal inferences, formalized theories of causation
- Ethical rules: normative evaluations, ethical theories.

Future Works

- Expand the scope of the formal language
- Incorporate Inductive Learning for the agent to learn from examples
- Adding more sample dilemmas

References

[1] Johnson, G.
Algorithmic bias: On the implicit biases of social technology, May 2020.

[2] Sarmiento, C., Bourgne, G., Inoue, K., Cavalli, D., and Ganascia, J.-G.
Action languages based actual causality for computational ethics: a sound and complete implementation in asp, 2022.

[3] Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., and Bernstein, A.
Implementations in machine ethics: A survey.
ACM Comput. Surv. 53, 6 (dec 2021).