Automating the Moral Reasoning Process in Ethical Dilemmas

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

Computer Science Department, Union College

Question

How can computer DO Ethics?

Introduction, Motivation, and Background

Artificial Intelligence is getting more involved in our daily lives, especially in important disciplines such as healthcare, policing, and self-driving cars. Giving control to the machines in such high-risk, consequential situations means that they will have to act in an ethically appropriate way, even without human supervision. For this reason, there has been growing interest in Machine Ethics in recent years [5]. The central question of this discipline is how to build and design a machine that can automatically recognize and make decisions in ethical scenarios.

(A) Switch the lever (B) Push the fat man

Figure: 2 classic variations of the trolley problem

This research take a rule-based approach to automate the ethical decision-making - that is, it puts emphasis on representing *moral principles* and making use of *logical reasoning* to evaluate an ethical scenario. As of now, this project focuses on formalizing and modeling the moral reasoning process in two well-known variations of the Trolley Problem.

Overview of Project Components

- Answer Set Programming: a logic programming language for representing knowledge and reasoning
- Event Calculus: the logic of change and action
- Doctrine of Double Effect: the moral principles [3]

Methods

The methodology consists of three main parts:

1. The Axioms describe in general the logical relationship between Events, Fluents (States) and Time. [4]

$$holdsAt(F, T+1) \leftarrow initiates(E, F, T), happens(E, T).$$

This *simplified* axiom says that if an event E can make a fluent F to be true at time T, AND event E actually happens at time T, then fluent F will start to hold at time $T+1$.

2. Trolley Problem Encoding We'll use the language of Event Calculus to model the trolley problems: movement of the trolley, effect of switching the lever, pushing the fatman, etc. Some examples of such *domain-dependent* rules.

- If the trolley and some person is at the same track location, the trolley will crash.

$$\begin{aligned} happens(U, crash, T) \leftarrow & holdsAt(on(trolley, Trk, TrkNo), T), \\ & holdsAt(on(P, Trk, TrkNo), T), \\ & person(P). \end{aligned}$$

- If the trolley runs over some person on the track, the person will die.

$$\begin{aligned} terminates(run(Trk, TrkNo), alive(P), T) \leftarrow \\ holdsAt(on(P, Trk, TrkNo), T), person(P). \end{aligned}$$

3. Means Versus Side Effect Given a bad effect F_1 and a good effect F_2, how do we know if the bad effect is a *means*, or a *mere side effect* for the good effect? The intuition test: *Had it not been for* F_1, *would* F_2 *still hold true?* [1] [2]

- Possible world semantics:

$$\begin{aligned} world(without(P1)) \leftarrow & \neg holdsAt(actual, alive(P1), T1), \\ & holdsAt(actual, alive(P2), T2), \\ & P1 \neq P2, T1 \leq T2. \end{aligned}$$

- Person $P1$ is treated as a means for $P2$ if, without $P1$, $P2$ wouldn't be alive.

$$\begin{aligned} treatedAsAMeansFor(P1, P2) \leftarrow \\ \neg holdsAt(actual, alive(P1), T1) \\ holdsAt(actual, alive(P2), T2) \\ \neg holdsAt(without(P1), alive(P2), T3) \\ P1 \neq P2, T1 <= T2. \end{aligned}$$

Results

We wish to make sure that, in our simulation, the *causation of events* is what we expected, and the *ethical evaluation* aligns with our theoretical formulas. The Event Calculus is run for 11 discrete time points. The following section displays a *subset* of the Answer Set output at the last time point for each scenario:

1. Switching the lever

$$\begin{aligned} & happens(base, switch, 1). \\ & \neg holdsAt(base, alive(group(1), 11). \\ & sideEffectFor(group(1), group(5)). \\ & permissible(switch). \end{aligned}$$

2. Pushing the fat man

$$\begin{aligned} & happens(base, push, 1). \\ & \neg holdsAt(base, alive(fatman), 11). \\ & treatedAsAMeansFor(fatman, group(5)). \\ & impermissible(push). \end{aligned}$$

3. No actions were made. (joined outputs of two scenarios).

$$\begin{aligned} & \neg holdsAt(base, alive(group(5)), 11). \\ & sideEffectFor(group(5), group(1)). \\ & sideEffectFor(group(5), fatman). \end{aligned}$$

Discussion

The logical formulas yield the output that aligns with our intuitions in these dilemmas

1. **Switching the lever: Side effect**
   Had it not been for the *group of 1 people*, the *group of 5 people* would have died?
2. **Pushing the fat man: Used as a Means**
   Had it not been for the *fat man*, the *group of 5 people* would have died?

Interestingly, in both scenarios where no actions are made, the death of group of 5 people are considered as a side effect for the survival of the fat man on the bridge and the group of 1 on the side track. This is consistent with the underlying logic, but it is an unexpected result for our intuition.

Conclusions

This model is a first step towards automating ethical reasoning, starting with a simple but ethically puzzling dilemma. There are yet many dilemmas that do not fit well within our theoretical formulations, both on an implementation and philosophical level. For the next steps of this project:

- Add more "test cases" and dilemmas
- Provide a general reasoning framework to easily encode any dilemmas
- Explore different formulations of the Doctrine of Double Effects to cover more possibilities
- Evaluate the theoretical limit of Double Effects in cases such as: the sacrificing soldier, abortion versus hysterectomy.
- Incorporate inductive-learning capability for the model to learn from examples.

References

[1] Bentzen, M. M., and Lindner, F.
A formalization of kant's second formulation of the categorical imperative.
CoRR abs/1801.03160 (2018).

[2] Govindarajulu, N. S., and Bringsjord, S.
On automating the doctrine of double effect.
CoRR abs/1703.08922 (2017).

[3] McIntyre, A.
Doctrine of Double Effect.
In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Fall 2023 ed. Metaphysics Research Lab, Stanford University, 2023.

[4] Mueller, E. T.
Commonsense Reasoning.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006.

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