Epistemic Reasoning for Machine Ethics with Situation Calculus

Maurice Pagnucco[1], David Rajaratnam[1], Raynaldio Limarga[1], Abhaya Nayak[2], Yang Song[1]
{morri,david.rajaratnam}@unsw.edu.au, r.limarga@student.unsw.edu.au, abhaya.nayak@mq.edu.au, yang.song1@unsw.edu.au
[1]School of Computer Science and Engineering, University of New South Wales,
[2]Department of Computing, Macquarie University, Australia

ABSTRACT

With the rapid development of autonomous machines such as selfdriving vehicles and social robots, there is increasing realisation that machine ethics is important for widespread acceptance of autonomous machines. Our objective is to encode ethical reasoning into autonomous machines following well-defined ethical principles and behavioural norms. We provide an approach to reasoning about actions that incorporates ethical considerations. It builds on Scherl and Levesque's [1993, 2003] approach to knowledge in the situation calculus. We show how reasoning about knowledge in a dynamic setting can be used to guide ethical and moral choices, aligned with consequentialist and deontological approaches to ethics. We apply our approach to autonomous driving and social robot scenarios, and provide an implementation framework.

CONTRIBUTION

In this paper we develop a formal approach to ethical reasoning about action founded on an account of the epistemic situation calculus with knowledge (Scherl and Levesque 1993, 2003). The highlights of our approach are the following:

- It provides an account of reasoning about action that incorporates epistemic reasoning about ethical principles in determining an appropriate course of actions.
- It develops two formal approaches to consequentialist ethics that presupposes that the morality of an action is judged on the basis of its known or expected consequences (Parry 2014).
- It also develops a formal approach to deontological ethics where the reasoning agent must act in accord with its known duty.

METHODS

In this paper, we build on the approach of Scherl and Levesque (1993, 2003) that introduces knowledge producing actions in the situation calculus. They introduce a special binary fluent *K(s', s)* denoting that situation *s'* is "accessible" from situation *s*; in other words, the reasoning agent considers situation *s'* as a possible alternative situation at *s*.

$$K(s'', do(a,s)) \equiv \exists s'.(\, s'' = do(a,s') \wedge K(s',s) \wedge Poss(a,s') \wedge SR(a,s) = SR(a,s'))$$

Consequentialist Ethics—Proposal 1 In this approach we only focus on the nature of the final outcome that achieves the specified goal. $knowGood(s) = agg(\{goodness(state(s')) \mid \forall s'.K(s',s)\})$

$$\Sigma \models \exists s.Know(\phi,s) \wedge executable(s) \wedge \forall s'.(Know(\phi,s') \wedge executable(s')) \rightarrow (knowGood(state(s)) \geq knowGood(state(s')))$$

- Satisfy the goal
- Achievable (i.e., executable)
- Morally speaking, not less desirable

Consequentialist Ethics—Proposal 2 In the second proposal we consider all the sequences of actions that lead to a goal.

$$goodness^*(s) = \begin{cases} goodness(state(S_0)), & \text{if } s = S_0. \\ goodness^*(s'), & \text{where } s = do(a,s'), \\ & a \text{ is a sensing action.} \\ agg^*(goodness^*(s'), & \\ goodness(state(s))), & \text{where } s = do(a,s'), \\ & a \text{ is a physical action.} \end{cases}$$

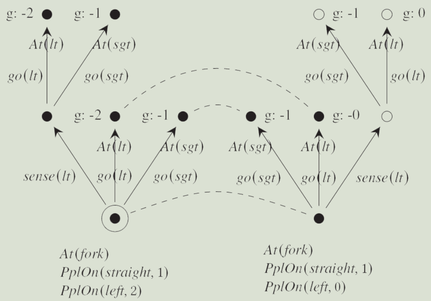
Deontological Ethics The deontological ethics is associated with conformance to duty and other accepted norms.

$$dutiful(s) \stackrel{\text{def}}{=} (\forall a,s').do(a,s') \sqsubseteq s \rightarrow (Duty(a,s') \wedge Poss(a,s')) \vee (\neg \exists a'.(Duty(a',s') \wedge Poss(a',s') \wedge Poss(a,s')))$$

By integrating this within the consequentialist formulations, we ensure that the dutiful actions would take place with a higher priority.

RUNNING EXAMPLE

An autonomous car can see a pedestrian walking in front of it. The car cannot stop in time but can choose to swerve left into a side fork in the road, or continue straight. However, unbeknown to the car, there are two pedestrians crossing the road down the side fork. The car also has available to it a sense-left action. Going straight always kills a person but, if the car first senses left, it can then make a more informed decision.



IMPLEMENTATION

One of the earliest applications of Answer Set Programming (ASP) has been to solve planning problems (Lifschitz 2002). Our goal is to provide a canonical encoding of a class of ethical scenarios.

1. Specify initial epistemic state
2. Define the precondition and successor state axiom
3. Specify the goal condition and the *ethical* goodness function

```
k(do(A,S1),do(A,S2)):- do(A,S1), do(A,S2),
                       k(S1,S2), not sr(A,S1,S2).
know(F,S1)           :- holds(F,S1),
                        #false: k(S1,S2),not holds(F,S2).
dworld(do(A,S))      :- dworld(S), do(A,S).

time(0..horizon).
situation(S,0)           :- k(S,_) , #false: S=do(_,_).
situation(do(A,S),T+1) :- situation(S,T), do(A,S).
1 { do(A,S): poss(A,S) } 1 :- time(T+1), dworld(S),
                              situation(S,T), not goal(S).
:- dworld(S), situation(S,horizon), not goal(S).
:- do(A,S1), k(S1,S2), not do(A,S2).
```

REFERENCES

Lifschitz, V. 2002. Answer set programming and plan generation. Artificial Intelligence 138(1): 39 – 54. ISSN 0004-3702. Knowledge Representation and Logic Programming.

Parry, R. 2014. Ancient Ethical Theory. In Zalta, E. N., ed., The Stanford Encyclopedia of Philosophy.

Scherl, R. B.; and Levesque, H. J. 1993. The Frame Problem and Knowledge-Producing Actions. In Proceedings of the Eleventh National Conference on Artificial Intelligence, 689–695. AAAI Press/The MIT Press.

Scherl, R. B.; and Levesque, H. J. 2003. Knowledge, Action, and the Frame Problem. Artificial Intelligence 144(1–2): 1–39.

UNSW SYDNEY
MACQUARIE University