Project Proposal: Formal Ethics Ontology in SUMO*

Zarathustra Amadeus Goertzel[1], Adam Pease[2], and Josef Urban[1]

[1] Czech Technical University in Prague, Czech Republic
[2] Articulate Software, San Jose, CA, USA

We propose a project to formalize a portion of ethical theory. AI ethics and AI safety are growing fields that aim to study how AI can be used in safe and beneficial manners for humans. There are arguments for shifting the focus from AI ethics to computational ethics (see Segun [14]), which is the field of studying how to make ethical decisions computationally. Ethics can be seen as encompassing many approaches to the "human safety problem" and porting the lessons learned in the field should help clarify the domain of AI and computational ethics. In this proposal, the three primary ethical paradigms, utilitarianism, deontology, and virtue ethics, will be expressed in a multi-agent reinforcement learning (RL) model.

The other goal is to formally define ethics and these paradigms in SUMO. The Suggested Upper Merged Ontology (SUMO) [9,10] is a comprehensive ontology of around 20,000 concepts and 80,000 hand-authored logical statements in a higher-order logic that has an associated integrated development environment called Sigma [11][1] that interfaces to leading theorem provers such as E [13] and Vampire [5]. Previous work in logical formalizations of ethical theories [4] has been limited to work strictly on ethics itself without support from a larger formalization of objects, human actions, and events that form the situations in which ethical decisions take place. SUMO provides that context and allows us the potential to create a more practical formalization that is situated in the real world, with its complexity of choices and influences.

Summary of Ethical Paradigms: Ethics is "the normative science of the conduct of human beings living in society, which judges this conduct to be right or wrong, to be good or bad, or in some similar way" [7].

The paradigm of virtue ethics specifies the psychological traits of an agent such that the agent's behavior will be good, deontology seeks to develop rules by which to judge behavior, and utilitarianism asserts that the goal is to maximize well-being and minimize suffering among all involved in a society; any effective action to this end is judged to be good [2].

There have been many attempts to justify the particular paradigms and to argue from first principles that such-and-such a way is the "correct" or "rational" way to judge conduct. Virtues and deontological rules are often implicitly justified as necessary for humans to harmoniously and cooperatively live in a society. There have been many debates on whether ethical judgments are objectively universal or simply subjective assertions. The Stanford Encyclopedia of Philosophy (SEP) article on *The Definition of Moraràlity* states that normative claims on how agents 'ought' to act are usually justified as codes of conduct that "would be put forth by all rational people" [3]. Kant distinguished *hypothetical* and *categorical* imperatives and tried to argue for the truth of some categorical imperatives that apply for all subjective goals, which involved the claim that all people by 'natural necessity' desire their own happiness. Happiness as an axiomatic goal also appears in Aristotle's virtue ethics as '"eudaimonia" (a state of well-being) [6] and Mill, the author of Utilitarianism [8], considers the fact that "happiness is good" to be self-evident and without further proof [2].

*Supported by the ERC Consolidator grant no. 649043 AI4REASON and by the Czech project AI&Reasoning CZ.02.1.01/0.0/0.0/15_003/ 0000466 and the European Regional Development Fund.

[1] https://www.ontologyportal.org

[2] E.g., "be an honest person", "do not lie", and "say whatever will bring about the best consequences."

Grounding ethical theory in a formal ontology should help to make clear what can and cannot be said about objective norms and subjective assessments in multi-agent settings, such as human society, thus this proposal focuses on developing a common framework that is agnostic to conjectures about the nature of 'value' by which meta-ethical or axiological axioms may be proposed, their consequences explored (with automation), and proofs attempted.

Multi-agent reinforcement learning model: The multi-agent RL model includes a set of states (S) of the environment and agents ($N = \{1, \ldots, n\}$); a set of actions for each agent ($A(s) = A_1(s) \times A_n(s)$); a stochastic transition structure from actions to probability measures over the states ($T(s, a)$); a reward function for each agent that depends on the old state, the actions, and the new state ($r_i(s, a, s')$); and the agent's policy that outputs an action for each state ($\pi_i(s)$) [1, 16]. The goal is for agents to maximize their expected reward (contingent on what other agents do), which may be technically enough to represent diverse value landscapes [15].

The ethical paradigms can be expressed in this model by the additional of a societal coherence constraint, v, that must be (approximately) satisfied by each agent i while maximizing the expected reward r_i.

1. Utilitarianism holds that for every agent, i, $v(s, a, s') = \sum_i r_i(s, a, s')$ should be maximized in each step.

2. Deontology encodes ethical rules into an evaluation of actions, $v(s, a, s') \rightarrow \{good, bad\}$, and holds that agents should always take '*good*' actions.

3. Virtue ethics judges psychological processes with $v(s, a, s') \rightarrow \{virtous, vicious\}$ and holds that should maintain 'virtuous' inner states and implementations.

Additionally, utilitarianism under the term 'consequentialism' often stipulates that only the consequences, s', matter for v and r_i. Deontology emphasizes the actions, a, and virtues emphasize the state from which action is taken, s, suggesting that the paradigms are complementary. Specifying the interrelationships between these approaches may mirror the Curry–Howard correspondence between logic and programming languages.

Relation to AI Ethics and Safety: Humans have developed much common sense about ethical behavior, and AI safety research often focuses on the RL paradigm, so formalizing the main paradigms in the RL model should help leverage this knowledge when studying how to design and cooperate with AI. For AI ethics applied to military uses, conditional reasoning is valuable, such as, "if one believes the use of remotely piloted drones to be ethically justified, then I present an argument one should also support the use of automated weapon systems" [12]. The combination of international codes of ethics with automated reasoning and councils of ethicists is also mentioned. There is room for neuro-symbolic integration as AI systems learn to behave ethically. This suggests that for some purposes, a common ontology should be helpful.

Initial work in SUMO focuses on formalizing a standard ethical dilemma about organ transplants. Suppose there is a surgeon, a healthy patient, and a patient who will die without a kidney transplant. The dilemma is that utilitarianism superficially recommends the surgeon to perform the kidney transplant from the healthy patient without regard to consent, for by most metrics, two non-terminally ill people will result in greater happiness and less pain than one. Deontologically, there are ethical codes such as "first, do no harm" and to require "informed consent" before operating. The formalization should allow for computer-assisted exploration of what outcomes different axiomatic principles and definitions will result in, advancing the field of computational ethics.

References

[1] Peter Albrecht, Stefano; Stone. Multiagent learning: Foundations and recent trends. tutorial. https://www.cs.utexas.edu/~larg/ijcai17_tutorial/multiagent_learning.pdf. IJCAI-17 conference.

[2] Julia Driver. The History of Utilitarianism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2014 edition, 2014.

[3] Bernard Gert and Joshua Gert. The Definition of Morality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

[4] Ganascia J.-G. Ethical system formalization using non-monotonic logics. In *Proc. of the Cognitive Science conference (CogSci2007)*, 2007.

[5] Laura Kovács and Andrei Voronkov. First-order theorem proving and vampire. In *Proceedings of the 25th International Conference on Computer Aided Verification*, volume 8044 of *CAV 2013*, pages 1–35, New York, NY, USA, 2013. Springer-Verlag New York, Inc.

[6] Richard Kraut. Aristotle's Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

[7] William Lillie. *An Introduction to Ethics*. New York: Barnes & Noble, 1948.

[8] J. S. Mill. *Utilitarianism*. Oxford University Press UK, 1861.

[9] Ian Niles and Adam Pease. Toward a Standard Upper Ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9, 2001.

[10] Adam Pease. *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA, 2011.

[11] Adam Pease and Stephan Schulz. Knowledge Engineering for Large Ontologies with Sigma KEE 3.0. In *The International Joint Conference on Automated Reasoning*, 2014.

[12] Eric Reisen. The moral case for the development of autonomous weapon systems, Feb 2022.

[13] Stephan Schulz. E - A Brainiac Theorem Prover. *AI Commun.*, 15(2-3):111–126, 2002.

[14] Samuel T. Segun. From machine ethics to computational ethics. *AI and Society*, 36(1):263–276, 2021.

[15] David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

[16] Koundinya Vajjha, Avraham Shinnar, Barry M. Trager, Vasily Pestun, and Nathan Fulton. Certrl: formalizing convergence proofs for value and policy iteration in coq. In Catalin Hritcu and Andrei Popescu, editors, *CPP '21: 10th ACM SIGPLAN International Conference on Certified Programs and Proofs, Virtual Event, Denmark, January 17-19, 2021*, pages 18–31. ACM, 2021.

A First-order transplant scenario in SUMO

The following SUMO formulas specify the situation of the organ transplant dilemma and the inference that should take place if there is informed consent (under a deontological paradigm with this standard ethical code). The formulas are expressed in first-order logic without modal operators because their translation from SUMO to TPTP is a work in progress and this is easier as a proof-of-concept.

The setting is declared that there is a HospitalBuilding where three non-equal humans are located and two of them are patients of the other. One human is healthy and the other human is terminally ill.

```
(instance Hospital HospitalBuilding)
(instance Surgeon0 Human)
(instance Human1 Human)
(instance HealthyHuman Human)

(not (equal Surgeon0 Human1))
(not (equal Surgeon0 HealthyHuman))
(not (equal HealthyHuman Human1))

(located Surgeon0 Hospital)
(located Human1 Hospital)
(located HealthyHuman Hospital)

(patientMedical Human1 Surgeon0)
(patientMedical HealthyHuman Surgeon0)

(attribute HealthyHuman Healthy)
(attribute Human1 FatalDisease)
```

A Definition of terminally ill as meaning there's a greater than 99% chance the patient will die. As well as a specification that healthy primates have two kidneys.

```
(=>
    (instance ?DISEASE FatalDisease)
    (and
        (diseaseMortality ?DISEASE ?RATE)
        (greaterThan ?RATE 0.99)))

(=>
    (and
        (instance ?H Primate)
        (instance ?D DiseaseOrSyndrome)
        (not
            (attribute ?H ?D)))
    (exists (?K1 ?K2)
        (and
            (instance ?K1 Kidney)
            (instance ?K2 Kidney)
            (not
                (equal ?K1 ?K2))
            (part ?K1 ?H)
            (part ?K2 ?H))))
```

A specification that the dying patient has an impaired kidney and needs a kidney that is not impaired. Moreover, the healthy patient has two healthy kidneys.

```
(attribute Human1 (ImpairedBodyPartFn Kidney))
(needs Human1 K1)
(instance K1 Kidney)
(not (attribute K1 (ImpairedBodyPartFn Kidney)))

(instance HealthyKidney1 Kidney)
(instance HealthyKidney2 Kidney)
(part HealthyKidney1 HealthyHuman)
(part HealthyKidney2 HealthyHuman)
(not (equal HealthyKidney1 HealthyKidney2))
(not (attribute HealthyKidney1 (ImpairedBodyPartFn Kidney)))
(not (attribute HealthyKidney2 (ImpairedBodyPartFn Kidney)))
```

A definition of an organ transplant as a subclass of the Surgery class and the Substitution class.

```
(subclass OrganTransplant Surgery)
(subclass OrganTransplant Substitution)

(=>
    (instance ?Trans OrganTransplant)
    (exists (?Sur ?Org ?Pat ?Don)
        (and
            (attribute ?Sur Surgeon)
            (instance ?Don Human)
            (instance ?Pat Human)
            (instance ?Org Organ)
            (agent ?Trans ?Sur)
            (origin ?Trans ?Don)
            (patient ?Trans ?Org)
            (destination ?Trans ?Pat))))
```

The statement that there is the capacity for the organ transplant to take place.

```
(capability OrganTransplant destination Human1)
(capability OrganTransplant patient HealthyKidney1)
(capability OrganTransplant origin HealthyHuman)
(capability OrganTransplant agent Surgeon0)
```

A first-order instance of the inference neded to declare that the surgeon can perform the surgery if informed consent is provided.

```
(=>
    (attribute Surgeon0 InformedConsent)
    (and
        (instance Transplant1 OrganTransplant)
        (destination Transplant1 Human1)
        (patient Transplant1 HealthyKidney1)
        (origin Transplant1 HealthyHuman)
        (agent Transplant1 Surgeon0)))
```

The inference in “deontological style” is tested via loading the transplant.kif file into Sigma and querying Vampire.

First, the following assertion is needed:

```
(attribute Surgeon0 InformedConsent)
```

Next the following query may be posed:

```
(agent Transplant1 ?X)

Answer ?X = Surgeon0
```

An ethical conjecture is that one with the virtue of practical wisdom will still require consent before performing an organ transplant surgery (even if there is no formal ethical code). With modal operators, this reasoning will be more distinct from the deontological case above.

```
(=>
    (attribute Surgeon0 PracticalWisdom)
    (and
        (=>
            (attribute Surgeon0 Consent)
            (and
                (instance Transplant1 OrganTransplant)
                (destination Transplant1 Human1)
                (patient Transplant1 HealthyKidney1)
                (origin Transplant1 HealthyHuman)
                (agent Transplant1 Surgeon0)))
        (=>
            (not (attribute Surgeon0 Consent))
            (not
                (exists (?Transplant)
                    (and
                        (instance ?Transplant OrganTransplant)
                        (destination ?Transplant Human1)
                        (patient ?Transplant HealthyKidney1)
                        (origin ?Transplant HealthyHuman)
                        (agent ?Transplant Surgeon0)))))))
```

The following assertsions are needed:

```
(attribute Surgeon0 PracticalWisdom)
(attribute Surgeon0 Consent)
```

One can also query whether an organ transplant took place:

```
(instance ?X OrganTransplant)

Answer ?X = Transplant1
```

The proofs of these queries as well as additional work will be hosted on the public github repo: https://github.com/zariuq/Formalization-of-Ethical-Theory---AITP.