Project Proposal: Formal Ethics Ontology in SUMO interrelations among the paradigms

Zarathustra Amadeus Goertzel[1]

Czech Technical University in Prague, Czech Republic

This document contains brief sketches as how one should be able to map between the paradigms at a high level.

1 Virtue Ethics

1.1 Virtue Ethics → Deontology

Virtues are fairly flexible so this should be easy.

There is a roman virtue of dutifulness (pietas). Thus one who is virtuous (with pietas) will likely be good by a duty-based paradigm (aka deontology) [1].

1.2 Virtue Ethics → Consequentialism

- Empathy (Compassionate Concern for Others)
- Care (Loving Service to Others)
- Episteme (science), which is skill with inferential reasoning (such as proofs, syllogisms, demonstrations).
- Phronesis (practical wisdom) nowledge of what to do, knowledge of changing truths, issues commands
- Techne (art, craftsmanship)

(1) seems to imply incorporating r_i for all agents into one's own goal or reward signal and (2) implies taking actions based on this aggregate of reward signals. (3-5) seem to imply the capacity for consequentialist style accounting to help guide action (to some degree as perfect accounting is intractable in general). However, if one considers the following to be a manifactured virtue, "learning from the consequences of one's actions", then combined with *care* and *empathy* (or compassion), Consequentialism should largely follow (so long as it is in harmony with other virtues).

[1] "The man who possessed pietas "performed all his duties towards the deity and his fellow human beings fully and in every respect," as the 19th-century classical scholar Georg Wissowa described it." (https://en.wikipedia.org/wiki/Pietas)

Pietal: "more than religious piety; a respect for the natural order: socially, politically, and religiously. Includes ideas of patriotism, fulfillment of pious obligation to the gods, and honoring other human beings, especially in terms of the patron and client relationship, considered essential to an orderly society." (https://en.wikipedia.org/wiki/Virtue#Roman_virtues)

2 Deontology

Deontology works with rules which are aligned with the language of logic, so this should be easy, too.

2.1 Deontology → Virtue Ethics

Assert the following rule:

- Be virtuous.

Alternatively, list the virtues: truthfulness, courage, temperance, liberality, etc.

Further, Martha Nussbaum argues that Kant already covered the topic of virtues via his work in deontology[2]

2.2 Deontology → Consequentialism

Take a formal description of consequentialism. Add the rule to "act according to this maxim. Something like the following probably works:

- "Act only according to that maxim by which you can also will that it would become a universal law." (Categorical Imperative 1)
- "To the best of your knowledge and capacity, always act for the maximum benefit of all present and future sentient beings as well as to reduce harm for all present and future sentient beings to the best of your knowledge and capacity."

3 Consequentialism

This is probably the messy or difficult direction. However, the mappings should provide a more constructive grounding to the ethical rules and the virtues of the other paradigms.

3.1 Consequentialism → Deontology

John Stuart Mill, one of the founders of Utilitarianism, recognized that calculating the consequences of every action is infeasible. Thus Mill recognized the utility of following ethical rules and guidelines in practice (as well as openly revising them so as to lead to greater gradients of benefit and suffering reduction) [3].

Consequentialism embraces any effective strategy and if developing rules to reflect common knowledge saves time and resources, allowing reasonably effective actions to be taken for the happiness of all (or to the minimization of the pain of all), then rules ought to be used.

3.2 Consequentialism → Virtue Ethics

A virtue is a trait of an agent that is expected to increase performance over a collection of environments as measured by utility (pleasure vs pain, preference (dis)satisfaction, etc).

This is very general.

The hypothesis is that standard virtues will probably score reasonably well and thus match this definition.

[2] https://link.springer.com/article/10.1023/A:1009877217694.

[3] https://en.wikipedia.org/wiki/Utilitarianism#cite_note-:0-49

4 Computational and Learning Paradigms

If one adopts the multi-agent system view as sketched in the project proposal,

- Consequentialism aims to compute the best action in each step.
- Deontology aims to learn logical rules that are expected to lead to good performance (see inductive logic programming, symbolic regression, etc).
- Virtue ethics aims to train models that are expected to perform well.