# An attempt at formalizing UPB-based ethics

Tuukka Pensala

11.3.2016

Version 1.0.0

**Abstract**

Stefan Molyneux presents a framework for ethics in his book *Universally Preferable Behaviour (UPB) - A Rational Proof of Secular Ethics*, in which commonly accepted moral rules are deduced from first principles. This paper is an attempt at outlining a mathematical formalism for a subset of the philosophical reasoning, namely ethics between actions of two people. This isn't a straightforward translation from the book to the language of mathematics, but an independent sequence of thought being the result of trying to understand the logic of UPB and the derived ethics. Applying the somewhat handwavy mathematical machinery to the common moral test cases seems to agree with the results acquired by Stefan.

## Morality function

For simplicity, we'll classify moral status to three categories:

| Status | $M$ |
|---|---|
| Good | $g$ |
| Aesthetic/Personal/Neutral | $n$ |
| Evil | $e$ |

Let's denote the set of all possible human actions as $\mathcal{A}$. This set includes actions like "killing another human being", that is, observable events with no attached meaning or intent. However, the moral status of an action isn't determined solely by the act itself, but depends on the status of the human minds participating in the action. For example, "murder" and "euthanasia" correspond to the same action, "killing another human being", but with differing states of the human minds, and therefore can have completely different moral status. So we need some way to denote intent behind objective acts. This is done via preferences $\mathcal{P}$.

For simplicity we're only considering situations where two people are interacting. In mathematical terms, the still-unknown moral status function can be parameterized as

$$M \equiv M\left(a_1,\, p_1,\, a_2,\, p_2\right),\ a_1, a_2 \in \mathcal{A},\ p_1, p_2 \in \mathcal{P}, \tag{1}$$

where action $a_i$ is performed by the holder of preference $p_i$. The resulting moral value is chosen to correspond to the person with index 1. We can simplify the situation by claiming that an action directly seeks to fulfill the preference $p$ held by the person, so

$$a_i \equiv a_i\left(p_i\right), \tag{2}$$

which implies

$$M\left(a_1\left(p_1\right),\, p_1,\, a_2\left(p_2\right),\, p_2\right) = M\left(p_1,\, p_2\right). \tag{3}$$

This is of course an idealization, and real situations where (3) doesn't apply can be constructed. We'll simply stay away from them, and focus on clear cut situations. Also, (2) implies that there are both capacity and choice involved, which is fine, because coma patients are not the most interesting test subjects for a moral theory. It's noteworthy to mention that the actions have been pushed to background; they, and the moral status are completely determined by the preferences held by the two people. Restrictions posed by environment are largery ignored.

We've skipped over some ambiguity in respect to the preferences $\mathcal{P}$, which needs to be dealt with. For example, is the preference of a murderer

$$p_1 \left(\text{"preference to initiate violence"}\right),$$

$$p_2 \left(\text{"preference to kill other people"}\right),$$

$$p_3 \left(\text{"preference to swing a knife when someone is near enough"}\right),$$

or

$$p_4 \left(\text{"preference to contract muscles in a sequence of (...) triggered by a visual observation (...)"}\right)?$$

It's clear that the more specific a preference gets, the more it resembles the description of the corresponding action $a\left(p\right)$. To keep the theory conceptually clean we choose to use preferences which are furthest away from the physical action, but still arguably identify the resulting action. We define the generalizing map

$$G : \mathcal{P} \to \mathcal{P},$$

where

$$G \circ G = G,$$

and

$$a\left(G\left(p\right)\right) \approx a\left(p\right).$$

Arguably, $G$ would then resolve the ambiguity in the presented preferences as

$$G\left(p_1\right) = p_1,$$

$$G\left(p_2\right) = G\left(p_3\right) = G\left(p_4\right) = p_2.$$

However, what the underlying generalized preferences for some actions or preferences are is somewhat debatable, and not in the realm of the formalism. Because we'll focus on clear cut cases, we can state

$$M = M\left(P_1,\, P_2\right),\ P_1,\, P_2 \in G\left(\mathcal{P}\right).$$

## Inverse morality

It could be tempting to say that sometimes an inverse of an evil act ($e$) is not necessarily a good one ($g$), but neutral ($n$). For example, keeping yourself from murdering doesn't seem particularly heroic. However, if

$$\neg e = n,$$

then by the rules of logic

$$\neg n = e,$$

which is to say, that not performing a neutral action could be an evil action, which in turn would render our classification moot. Therefore it's reasonable to establish

$$\neg e = g,\ \neg g = e,\ \neg n = n,$$

that is, an opposite of good is evil, an opposite of neutral is neutral. By this rather self-evident reasoning we've found a crucial difference between good, bad, and neutral. Formally, for morally neutral preferences

$$M\left(P_1,\, P_2\right) = M\left(\neg P_1,\, P_2\right) = n,$$

and for morality-involving preferences

$$M\left(P_1,\, P_2\right) = \neg M\left(\neg P_1,\, P_2\right) = g.$$

In words, if inverting a preference doesn't lead to changing moral value, we've found a morally neutral preference. In the other case, where it leads to change of moral value, we've found a pair of good and evil behavior.

## Preference universality condition

For a preference pair $(P_1,\, P_2)$ to be considered universal, $a\left(P_1\right)$ shouldn't keep $P_2$ from realizing. Let's take this as the definition for the universality-determining function,

$$W\left(P_1,\, P_2\right) \equiv \neg\left[a\left(P_1\right) \implies \neg a\left(P_2\right)\right],$$

which evaluates to "true" if the preference pair is universal, false otherwise. It's debatable if $W$ should be symmetric or not. We'll settle for the safe assumption, that a preference pair $(P_1,\, P_2)$ is not generally the same as $(P_2,\, P_1)$.

## Objective morality

We've now established a somewhat objective way to classify preference pairs in two categories, universal and non-universal. We also have a bit of framework for working with preferences and morality. The simplest way to join the puzzle pieces together to acquire a testable moral theory is to establish a correspondence between the preference classification and moral categories. With a bit of foresight we choose

$$W\left(P_1,\, P_2\right) = W\left(\neg P_1,\, P_2\right) \iff M\left(P_1,\, P_2\right) = M\left(\neg P_1,\, P_2\right) = n, \tag{4}$$

for classifying a morally neutral preference pair, and

$$W\left(P_1,\, P_2\right) = \neg W\left(\neg P_1,\, P_2\right) = \text{true} \iff M\left(P_1,\, P_2\right) = M\left(\neg P_1,\, P_2\right) = g, \tag{5}$$

for classifying a good/evil pair, where $P_1$ is chosen to correspond to the good case. The chosen equivalences might seem arbitrary, but they're only claiming something along the lines of "what can be preferred, ought to be preferred."

## Tests for the theory

### Rape

$$P_1 = p\left(\text{"preference for intercourse"}\right)$$

$$P_2 = \neg P_1$$

Now, enforcing $P_1$ clearly disallows $P_2$ from realizing;

$$W\left(P_1,\, P_2\right) = \text{false}.$$

For the complementary case, self-evidently

$$W(\neg P_1,\, P_2) = W(P_2,\, P_2) = \text{true}.$$

(5) then dictates

$$M(\text{"rape"}) = M(P_1,\, P_2) = e,$$
$$M(\text{"not-rape"}) = M(\neg P_1,\, P_2) = g,$$

which agrees with the common moral understanding.

## Murder

$$P_1 = G(p(\text{"preference to kill the other person"})) = p(\text{"preference for the other person to be dead"}),$$
$$P_2 = G(p(\text{"preference to fight back or flee"})) = p(\text{"preference to not to die"}).$$

$a(P_1)$ clearly overrides $P_2$ in a murder (otherwise it would be an euthanasia), so

$$W(P_1,\, P_2) = \text{false}.$$

$a(\neg P_1)$ and $P_2$ are compatible, so

$$W(\neg P_1,\, P_2) = \text{true}.$$

Therefore, by (5)

$$M(\text{"murder"}) = M(P_1,\, P_2) = e.$$

## Theft

Theft is evil by the same reasoning as rape or murder.

## Lying

$$P_1 = p(\text{"preference to provide falsehoods"}),$$
$$P_2 = p(\text{"preference to believe the other person"}).$$

Clearly the actions resulting from the preferences don't interfere with each other, therefore

$$W(P_1,\, P_2) = W(\neg P_1,\, P_2) = \text{true},$$
$$\implies M(\text{"lying"}) = M(P_1,\, P_2) = n.$$

## Saving from death

Person 2 is for some reason in the imminent danger of losing his life.

$$P_1 = p(\text{"preference to help other people"}),$$
$$P_2 = p(\text{"preference to stay alive"}).$$

Self-evidently

$$W(P_1,\, P_2) = \text{true},$$

but the $\neg P_1$ case, where the other person is left to die requires a bit more thought. The question is, does neglect cause the violation of $P_2$, that is, death? I'd argue that it doesn't, because we could place e.g. a wall between the test subjects and the action $a(\neg P_1)$ would remain exactly the same, i.e. it doesn't have causal effect to the death of the other person. Therefore it must be that

$$W(\neg P_1,\, P_2) = \text{true},$$
$$\implies M(\text{"helping"}) = M(\text{"neglect"}) = n.$$

This may sound harsh, but one should remember that the category $n$ contains also aesthetically positive and negative behaviors.

## Initiation of the use of force

Because our formalism works with general preferences, and not specific actions themselves, we can try to establish some overarching moral rules.

$$P_1 = p\,(\text{"preference for initiating the use of force"})\,,$$

$$P_2 = p\,(\text{"preference for not being subject to the use of force"})\,.$$

Now,

$$W\,(P_1,\, P_2) = \text{false},$$

$$W\,(\neg P_1,\, P_2) = \text{true},$$

which by (5) results to

$$M\,(\text{"initiation of the use of force"}) = M\,(P_1,\, P_2) = e.$$

So initiating the use of force against a resisting person is evil. What about self-defense?

## Self-defense

Let's just swap $P_1$ and $P_2$ from the previous part and see where the theory leads us. Enforcing $P_2$ would prevent $P_1$ from realizing, so

$$W\,(P_2,\, P_1) = \text{false}.$$

Inverting $P_2$ turns the situation into an agreement, so

$$W\,(\neg P_2,\, P_1) = \text{true}.$$

By (5)

$$M\,(P_2,\, P_1) = e,$$

so "self-defense" is immoral! This apparent problem is resolved by noticing that an initiation of the use of force occurs against an unwilling person, which is the case we already went through. The correct parametrization for a self-defense scene is

$$P_2 = p\,(\text{"preference to initiate the use of force"})\,,$$

$$P_1 = p\,(\text{"preference to respond to the initiation of the use of force"})\,.$$

Now, $W\,(P_1,\, P_2)$ must be true, because being the defender implies that $a\,(P_2)$ has already taken place, so it can't be prevented. $W\,(\neg P_1,\, P_2)$ is also clearly true. Then (4) gives

$$M\,(\text{"self-defense"}) = M\,(P_1,\, P_2) = n.$$

## Voluntary interaction

$$P_1 = p\,(\text{"preference for initiating a voluntary interaction"})\,,$$

$$P_2 = p\,(\text{"preference for not participating in the voluntary interaction"})\,.$$

Because we're talking about voluntary interactions $a\,(P_1)$ can't include any form of coercion, that is, it can't keep $a\,(P_2)$ from happening. Therefore,

$$W\,(P_1,\, P_2) = \text{true},$$

$$W\,(\neg P_1,\, P_2) = \text{true},$$

$$\implies M\,(\text{"initiating voluntary interaction"}) = M\,(P_1,\, P_2) = n.$$

## Further study

- actually trying to find cases where this logic breaks down
- generalization to more than two people
- separation of aesthetical, personal and neutral categories
- identifying grey areas
- how to manage complex preferences, like $p = p_1 \wedge p_2$, or $p = p_1 \vee p_2$