Ethicophysics I: Conservation Laws

Eric Purdy

May 29, 2018

Abstract

What are Good and Evil? How do we explain these concepts to a computer sufficiently well that we can be assured that the computer will understand them in the same sense as humans understand them? These are hard questions, and people have often despaired of finding any answers to the AI safety problem.

In this paper, we lay out a theory of ethics modeled on the laws of physics. The theory has two key advantages: it squares nicely with most human moral intuitions, and it is amenable to rather straightforward computations that a computer could easily perform if told to. It therefore forms an ideal foundation for solving the AI safety problem.

1 Introduction

In this document, we lay out the beginnings of a new theory of ethics and human nature that we term *ethicophysics*. This is intended to be a complete and scientifically accurate account of the nature of Good and Evil, and other such ethical riddles that have haunted humanity since the beginning of our species. We term it ethicophysics to suggest that there are certain natural laws in the ethical sphere that cannot be violated any more than the laws of physics can be violated.

Since such a project is ambitious to the point of madness, we ask the reader's indulgence in following along with what must seem a quixotic quest to end all quixotic quests. Nevertheless, we hold that some things are true and some things are false, that some actions are good and some are evil. Ultimately, words mean things, not because the universe says they must, but because we choose to use them in a certain way and not in other ways.

We consider an *actor network*, which is a set of actors who act in the same physical space and communicate with one another. The minds of the actors are presumed to be *non-physical*, i.e., they are powered by computational devices which are not modeled by the laws of physics used to reason about the rest of reality. This is obviously a weird assumption - all really existing computational devices (brains, computers, abacuses, etc.) are physical and obey the physical laws of reality. The goal here is to separate reality out into the *naive* physical reality modeled by traditional physics and the *ethical* physical reality modeled

by the ethicophysics. Since computational devices exist in reality and have the properties they have because of the laws of physical reality, the ethicophysics is in some sense a proper subset of "real" physics; thus ethicophysics and traditional physics coexist as partners in describing the laws of reality, rather than fighting one another.

It is presumed that actors can communicate ideas to one another at will through non-physical means; this is again a strange assumption, but we make it for similar reasons as above.

It is not presumed that actors are virtuous, ethical, truthful, etc. In fact, the predominant motivating question in the ethicophysics is why people aren't significantly more evil than they appear to be.

2 On God and Souls

We use the term "God" to refer to a potential omniscient observer of the universe. We make no claims as to the ontological status of such a being. Note, in particular, that we do not assume that God is omnipotent or omnibenevolent, which allows us to avoid the classic Epicurean trilemma [7]:

> God, he says, either wishes to take away evils, and is unable; or He is able, and is unwilling; or He is neither willing nor able, or He is both willing and able. If He is willing and is unable, He is feeble, which is not in accordance with the character of God; if He is able and unwilling, He is envious, which is equally at variance with God; if He is neither willing nor able, He is both envious and feeble, and therefore not God; if He is both willing and able, which alone is suitable to God, from what source then are evils? Or why does He not remove them?

We note, however, that the content of the ethicophysics suggests that such an entity, if it did exist, would be reasonably omnibenevolent, and as omnipotent as is consistent with the existence of free will. As noted by Dr. Martin Luther King Jr. [6], a God that did not allow for free will would simply be a tyrant:

> I am thankful that we worship a God who is both tough minded and tenderhearted. If God were only tough minded, he would be a cold, passionless despot sitting in some far-off heaven "contemplating all," as Tennyson puts it in "The Palace of Art." He would be Aristotle's "unmoved mover," self-knowing but not other-loving. But if God were only tenderhearted, he would be too soft and sentimental to function when things go wrong and incapable of controlling what he has made. He would be like H.G. Wells's loveable God in *God, the Invisible King*, who is strongly desirous of making a good world but finds himself helpless before the surging powers of evil. God is neither hardhearted nor soft minded. He is tough minded enough to transcend the world; he is tenderhearted enough to live in it. He

> does not leave us alone in our agonies and struggles. He seeks us in dark places and suffers with us and for us in our tragic prodigality.

2.1 Defining the Soul

We define the soul of an individual actor to be *that which is true about the actor*. In religious terms, it is basically God's opinion about the actor.

Note that, in particular, that which is true about the actor includes what opinion every human that ever lived would have of the actor if they were given true knowledge of the events and choices of that actor's existence. This sort of "subjective truth" will be deeply contradictory (presumably e.g. Hitler and Churchill would disagree about a lot), but it is no less real for that.

2.2 On the Equality of Souls

Many have noted that one can choose to view all human beings as fundamentally equal in the context of ethics, e.g.:

> Do to others what you want them to do to you. This is the meaning of the law of Moses and the teaching of the prophets. [1]

> We hold these Truths to be self-evident, that all Men [sic] are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty, and the pursuit of Happiness... [5]

If we expand this slightly to include all actors that have both a soul and a mind, it seems as good a foundation as any for a theory of ethics. In particular, given our definition of the soul, any actor with a mind can be said to have a soul. This includes, in our opinion, animals (see e.g. [9, 2]) and sufficiently advanced computer programs (see [10]).

3 Main Results

In this section, we pursue traditional mathematical proofs of certain propositions in the field of ethics.

3.1 Love and Hate

We define two quantities called love (denoted by $l(a, B)$) and hate (denoted by $h(a, B)$, intended to be interpreted roughly in their standard English senses. (It is presumed that they can be measured precisely in some way via e.g. advanced neuroscientific theories that we do not presume to know. The important point here is just that there will come to exist some rigorous technical definition of the quantities such that their epistemological status is not in question.) An actor a in the actor network can experience love or hate for any subset B of the actor

network. (In particular, the actor can love and/or hate itself.) Love and hate are presumed to be non-negative quantities. Note that love and hate are not mutually exclusive, but are rather orthogonal quantities.

3.2 Conservation of Opinionatedness

We define the quantity *active opinionatedness of* a, which is the sum of the squared love and hate values of a for all subsets of the network:

$$op(a) = \frac{1}{2} \sum_{B} l(a, B)^2 + h(a, B)^2.$$

We also define the *mindshare* of B in a's mind to be

$$ms(a, B) = \frac{l(a, B)^2 + h(a, B)^2}{2 \cdot op(a)}.$$

We define the *active opinionatedness* of the network to be

$$\mathcal{O}_{act} = \sum_{a} op(a).$$

Active opinionatedness serves roughly the same role in the ethicophysics as kinetic energy does in traditional physics. We also need to define *potential opinionatedness* $\mathcal{O}_{pot}$, which serves the same role in the ethicophysics as potential energy does in traditional physics. We do not yet know how to define all possible sources of future love and hate, so we cannot give a rigorous specification of how to compute potential opinionatedness. It can, however be defined rigorously, by requiring that the *total opinionatedness*

$$\mathcal{O}_{act} + \mathcal{O}_{pot}$$

be conserved, and simply watching what $\mathcal{O}_{act}$ does over time.

Total opinionatedness is conserved by definition, as long as the set of network participants does not change. This can be achieved by defining the love and hate values of an absent participant (e.g., a dead person, or a person yet to be born) to be something relatively arbitrary, and then simply considering all participants that ever have or ever will exist. For instance, we could define the love and hate values of a non-alive person to be the average love and hate they experienced or will experience over the course of their lives.

3.3 The Golden Theorem: Actions Have Consequences

Theorem 3.1 (The Golden Theorem)**.** *Actions have consequences. In particular, the consequence of committing an evil act that goes undetected is that one becomes the person that one becomes after such an act, and has as a consequence an unclean conscience.*

Proof. Note that this proof needs to be checked over very thoroughly, as it may contain errors.

Consider the "objective", "physical" Lagrangian $\mathcal{L}(q, \dot{q}, t) = \mathcal{T} - \mathcal{U}$, where $\mathcal{T}$ is the kinetic energy of a system, and $\mathcal{U}$ is the potential energy of that system. Here q is the physical state of the system in generalized coordinates.

Let $\mathcal{S} = \mathcal{O}_{ac} - \mathcal{O}_{pot}$ be the "subjective", "ethical" Lagrangian of the system. This is supposed to depend upon the generalized coordinates $q, \dot{q}, t$ of the system and the "subjective coordinates" $s, \dot{s}, t$ (which are supposed to have no physical realization that is legible to the laws of physics under consideration).

Let $\tau(t)$ (called the "tweak") be a continuous symmetry of the physical system, i.e., for infinitesimal ϵ, the transformation

$$q(t) \to q(t) + \epsilon\tau(t)$$
$$\dot{q}(t) \to \dot{q}(t) + \epsilon\dot{\tau}(t)$$

leaves the Lagrangian unaffected.

Let $\varphi(s)$ (called the "flip") be a discrete non-physical symetry of the subjective energy function at time t_φ, i.e., a function such that, for one brief instant of time,

$$\mathcal{S}(q, \dot{q}, \varphi(s), \frac{d}{dt}\varphi(s), t_\varphi) = \mathcal{S}(q, \dot{q}, s, \dot{s}, t_\varphi).$$

Since $\mathcal{S}$ is a function of network participant love and hate values, it is generally the case that φ will be a permutation of the actors in the actor network.

Define the following quantity (the "God Lagrangian"):

$$\mathcal{G}(q, \dot{q}, s, \dot{s}, t) = \mathcal{L}(q, \dot{q}, t) + \mathcal{S}(q, \dot{q}, s, \dot{s}, t) - \mathcal{S}(q, \dot{q}, \varphi(s), \frac{d}{dt}\varphi(s), t)$$

By Noether's Theorem [8], the following quantity is conserved:

$$\sum_{i=1}^{n} \frac{\partial \mathcal{L}}{\partial \dot{q}_i}\tau_i.$$

By Noether's Theorem applied to the modified Lagrangian, the same is true of the quantity

$$\sum_{i=1}^{n} \frac{\partial \mathcal{G}}{\partial \dot{q}_i}(\varphi \circ \tau)_i + \sum_{j=1}^{m} \frac{\partial \mathcal{G}}{\partial \dot{s}_j}(\varphi \circ \tau)_j.$$

Note that, since the quantities s_j are presumed to be causally upstream from the physical world, it can be safely assumed that φ has no effect on the values of the q's, while τ probably does have an effect on the s's and thus the following quantity is conserved:

$$\sum_{i=1}^{n} \frac{\partial \mathcal{G}}{\partial \dot{q}_i}\tau_i + \sum_{j=1}^{m} \frac{\partial \mathcal{G}}{\partial \dot{s}_j}(\varphi \circ \tau)_j.$$

After doing some algebra, we arrive at the conclusion that the following quantity is conserved by the laws of ethicophysics:

$$\sum_{i=1}^{n} \frac{\partial \mathcal{S}}{\partial \dot{q}_i} \tau_i - \sum_{i=1}^{n} \frac{\partial \mathcal{S}_\varphi}{\partial \dot{q}_i} \tau_i + \sum_{i=1}^{m} \frac{\partial \mathcal{S}}{\partial \dot{s}_i} (\varphi \circ \tau)_i - \sum_{i=1}^{m} \frac{\partial \mathcal{S}_\varphi}{\partial \dot{s}_i} (\varphi \circ \tau)_i,$$

which simplifies to the equation

$$\sum_{i=1}^{n} \frac{\partial \mathcal{S}}{\partial \dot{q}_i} \tau_i + \sum_{i=1}^{m} \frac{\partial \mathcal{S}}{\partial \dot{s}_i} (\varphi \circ \tau)_i = \sum_{i=1}^{n} \frac{\partial \mathcal{S}_\varphi}{\partial \dot{q}_i} \tau_i + \sum_{i=1}^{m} \frac{\partial \mathcal{S}_\varphi}{\partial \dot{s}_i} (\varphi \circ \tau)_i,$$

We are now ready to finish the proof. Consider some binary decision that can be made, and consider the two possible timestreams that will follow making either choice. Let s be the quantity of respect that God feels for one, defined as $l(\text{God}, a) - h(\text{God}, a)$. We can call this the *coherent extrapolated conscience* of the actor. (We note that it has an imperfect relationship to the subjective experience of the conscience with which most humans are familiar.)

Suppose, further, that the decision has no consequences that are perceivable in the external physical world after some time period $t_{\text{hidethebody}}$ has elapsed. Thus, after this point, $\mathcal{S}$ "should" no longer depend on q.

There is then an additional conserved quantity (in addition to $\mathcal{S}$), which is the "coherent extrapolated conscience momentum with respect to φ"

$$CECM_\varphi = \frac{\partial(\mathcal{S} - \mathcal{S}_\varphi)}{\partial \dot{s}} s.$$

We can calculate $CECM_\varphi$, and find it to be equal to

$$\begin{aligned} = \quad & \frac{\partial \mathcal{S}}{\partial \dot{l}(\text{God}, a)} \frac{\partial \dot{l}(\text{God}, a)}{\partial \dot{s}} \\ & + \frac{\partial \mathcal{S}}{\partial \dot{h}(\text{God}, a)} \frac{\partial \dot{h}(\text{God}, a)}{\partial \dot{s}} \\ & - \frac{\partial \mathcal{S}_\varphi}{\partial \dot{l}(\text{God}, a)} \frac{\partial \dot{l}(\text{God}, a)}{\partial \dot{s}} \\ & - \frac{\partial \mathcal{S}_\varphi}{\partial \dot{h}(\text{God}, a)} \frac{\partial \dot{h}(\text{God}, a)}{\partial \dot{s}} \end{aligned}$$

Assuming that the coordinates s capture all meaningful ethical variables (and in particular that $\mathcal{O}_{pot}$ depends only on s and not on $\dot{s}$), this simplifies to

$$\begin{aligned}
= \quad & \frac{\partial \mathcal{O}_{act}}{\partial \dot{l}(\mathrm{God}, a)} \frac{\partial \dot{l}(\mathrm{God}, a)}{\partial \dot{s}} \\
& + \frac{\partial \mathcal{O}_{act}}{\partial \dot{h}(\mathrm{God}, a)} \frac{\partial \dot{h}(\mathrm{God}, a)}{\partial \dot{s}} \\
& - \frac{\partial \mathcal{O}_{act,\varphi}}{\partial \dot{l}(\mathrm{God}, a)} \frac{\partial \dot{l}(\mathrm{God}, a)}{\partial \dot{s}} \\
& - \frac{\partial \mathcal{O}_{act,\varphi}}{\partial \dot{h}(\mathrm{God}, a)} \frac{\partial \dot{h}(\mathrm{God}, a)}{\partial \dot{s}} \\
= \quad & l(\mathrm{God}, a) \\
& - h(\mathrm{God}, a) \\
& - l(\mathrm{God}, \pi_\varphi(a)) \\
& + h(\mathrm{God}, \pi_\varphi(a))
\end{aligned}$$

This yields what is essentially a proof of Newton's third law (every action has an equal and opposite reaction), but in the ethical domain: every action has an equal and opposite *ethical* reaction. In more poetic terms, this is a proof that God thinks "whatsoever you do to the least of my brethren, that you do unto me". God's opinion of the perpetrator of an unconscionable act with a victim changes by an amount equal to and opposite to the change in his opinion of the victim of the same act. (Take π_φ to be the permutation that switches the roles of perpetrator and victim.)

This same principle can be applied to any binary decision. The total coherent extrapolated conscience momentum will be the same in either case (i.e., in both timestreams). But, assuming the decision is one with a clear right answer, the predominant sign of $\frac{\partial \mathcal{S}}{\partial s}$ will generally be the opposite of the predominant sign of $\frac{\partial \mathcal{S}_\varphi}{\partial s}$, assuming that π_φ is a permutation that switches the positions of beneficiaries and victims. Thus, making the wrong decision will have hugely negative consequences for one's conscience, as expected. These consequences are not permanent; one can be forgiven sins, but in general only when one has overcome the sin and made recompense.

□

3.4 Playing Favorites: Weighted Opinionatedness

Let w_a be the weight of a according to some external observer. It is presumed that God does not apply non-even weighting (because of the equality of souls), but there is nothing stopping the rest of us from having favorites.

We define the quantity *weighted active opinionatedness of* a, which is the sum of the squared love and hate values of a for all subsets of the network,

weighted by the weight of a:

$$op_w(a) = \frac{1}{2} w_a \sum_B l(a, B)^2 + h(a, B)^2.$$

As the name suggests, weighting plays a similar role in the ethicophysics as mass plays in traditional physics. More specifically, we can defined a weighted coherent extrapolated conscience, which is sort of the fully considered opinion of the smartest possible version of whatever entity chose the weighting function. The Golden Theorem can then be extended to show that Newton's third law holds, not only for God's opinion, but for the opinion of any omniscient observer.

4 Discussion

4.1 Theodicy

We wish to point out a potential misreading of the theorems in this paper, which is that God will help people who are virtuous in some straightforward way. This is simply untrue, and potentially dangerous for anyone to believe. Consider, e.g., the following piece of vileness due to Hitler [3]:

> I did not want this struggle. Since January, 1933, when Providence entrusted me with the leadership of the German Reich, I had an aim before my eyes which was essentially incorporated in the program of our National Socialist party. I have never been disloyal to this aim and have never abandoned my program... Only when the entire German people become a single community of sacrifice can we expect and hope that Almighty God will help us. The Almighty has never helped a lazy man. He does not help the coward. He does not help a people that cannot help itself. The principle applies here, help yourselves and Almighty God will not deny you his assistance.

This was a vile lie told by a vile man for vile purposes. In reality, bad things can and do happen to good people, and God will do nothing to stop them. Or rather, he will whisper the truth in our minds, and we all of us will do whatever it is that we will do, and that is the only aid that God ever has or ever will provide. Bad things happen to good people because other good people are not able to stop them from happening, and because bad people ignore the whispers of their broken consciences.

5 Epilogue

We find that the following lyrics of Yusuf Islam [4] capture the sort of spirit of what we are trying to accomplish in this paper:

> I wish I knew, I wish I knew
> What makes me, me, and what makes you, you

It's just another point of view, ooo
A state of mind I'm going through, yes
So what I see is never true, ahhh

I wish I could tell, I wish I could tell
What makes a heaven what makes a hell
And do I get to ring my bell, ooo
Or land up in some dusty cell, no
While others reach the big hotel, yeah

I wish I had, I wish I had
The secret of good, and the secret of bad
Why does this question drive me mad? ahhh
'Cause I was taught when but a lad, yes
That bad was good and good was bad, ahhh

I wish I knew the mystery of
That thing called hate, and that thing called love
What makes the in-between so rough? ahhh
Why is it always push and shove? ahhh
I guess I just don't know enough, yes

References

[1] Christ, J. Sermon on the mount. *Matthew 7:12* (33).

[2] Coetzee, J. The lives of animals. *TANNER LECTURES ON HUMAN VALUES 20* (1999), 111–166.

[3] Hitler, A. *Radio broadcast from Berlin* (1941).

[4] Islam, Y. I wish, i wish. *Mona Bone Jakon* (1970).

[5] Jefferson, T. Declaration of independence. *Various Printers* (1776).

[6] King Jr, M. L., King, C. S., and King, C. S. A tough mind and a tender heart.. *Strength to Love* (1963), 13–20.

[7] Lactantius. *De Ira Dei*.

[8] Noether, E. Invariante variationsprobleme. *Nachr. D. Knig. Gesellsch. D. Wiss. Zu Gttingen, Math-phys. Klasse* (1918), 235257.

[9] Singer, P. *Animal liberation*. Random House, 1995.

[10] Turing, A. Computing machinery and intelligence. *Methodos 6* (1954), 195–223.