Ethicophysics I: Conservation Laws

Eric Purdy

November 27, 2023

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.

1.1 Plan of Attack

In this document (Ethicophysics I), we define some key terms and prove the Golden Theorem, which allows us to derive ethicophysical conservation laws. In the second document (Ethicophysics II), we apply these laws to a very simple model of a community of reinforcement learning agents, and give a number of useful ethicophysical conservation laws.

Future work that remains to be done will include solving all of AI, which will take some time.

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 [8]:

> 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. [7], 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 [sic] 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... [6]

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. [11, 2]) and sufficiently advanced computer programs (see [12]).

3 Main Results

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

Concept	Symbol	Definition	Analogue from traditional physics
a loves B	$l(a, B)$	Accumulated positive emotion	Position
a hates B	$h(a, B)$	Accumulated negative emotion	Position
a likes B	$\dot{l}(a, B)$	Positive emotion being actively experienced by a subject (time derivative of love)	Velocity
a dislikes B	$\dot{h}(a, B)$	Negative emotion being actively experienced by a subject (time derivative of hate)	Velocity
B helps a	$\ddot{l}(a, B)$	Positive emotion being actively caused in a subject (second time derivative of love)	Acceleration
B hurts a	$\ddot{h}(a, B)$	Negative emotion being actively caused by a subject (second time derivative of hate)	Acceleration
a's active subjective energy	$\mathcal{A}(a)$	Total amount of subjective energy bound up in the current experience of the subject	Kinetic energy
Potential subjective energy	$\mathcal{P}$	???	Potential energy

Table 1: Some useful concepts

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.

We define a number of related concepts in Table 3.1.

3.2 Conservation of Subjective Energy

We define the quantity *active subjective energy of* a, which is the difference of the squared like and dislike values of a, summed over all subsets of the network:

$$\mathcal{A}(a) = \frac{1}{2} \sum_{B} \dot{l}(a, B)^2 - \dot{h}(a, B)^2.$$

We define the *active subjective energy of the network* to be

$$\mathcal{A} = \sum_{a} \mathcal{A}(a).$$

Active subjective energy serves roughly the same role in the ethicophysics as kinetic energy does in traditional physics. We also need to define *potential subjective energy* $\mathcal{P}$, 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 like and fear, so we cannot give a rigorous specification of how to compute potential subjective energy. It can, however be defined rigorously, by requiring that the *total subjective energy*

$$\mathcal{A} + \mathcal{P}$$

be conserved, and simply watching what $\mathcal{A}$ does over time to deduce the laws of ethicophysics.

Total subjective energy 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, or some other constant value. (Note that this will have the effect that non-alive actors experience no like or dislike, which makes sense.)

Potential subjective energy serves roughly the same role in the ethicophysics as potential energy does in traditional physics. Potential subjective energy generally arises from cost functions and reward functions - objectives deemed desirable by a particular mind will generally be modeled by some cost function, which then drives goal-directed actors to minimize the costs and maximize the reward, thus making the behavior of the actor at least marginally predictable in theory. We thus arrive in a situation where theories from traditional physics such as minimum-action principles can be applied to the freely chosen actions of actors in an actor network.

3.3 Playing Favorites: Weighted Subjective Energy

Let w_a be the weight of actor 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 *emotion*, which is the *weighted active subjective energy as perceived by a*. This is the same as active subjective energy, but weighted by how much a cares about b's opinion:

$$\mathcal{A}_a = \frac{1}{2} \sum_b w_a(b) \sum_C \dot{l}(a,B)^2 - \dot{h}(a,B)^2.$$

As the name suggests, weighting plays a similar role in the ethicophysics as mass plays in traditional physics.

We similarly define $\mathcal{P}_a$. We thus have laws of emotion for each specific actor that determine their behavior in particular. The laws of emotion for ethicophysics are as follows: let $\mathcal{S}_a$ be the subjective Lagrangian $\mathcal{S}_a = \mathcal{A}_a - \mathcal{P}_a$. Then we have

$$\frac{\partial \mathcal{S}}{\partial x} = \frac{d}{dt} \frac{\partial \mathcal{S}}{\partial \dot{x}}.$$

Here x can be either a physical or an ethical variable.

3.4 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 $\mathcal{S} = \sum_a \mathcal{S}_a$ be the "subjective", "ethical" Lagrangian of the system. This is supposed to depend upon the generalized coordinates $q, \dot{q}, t$ of the physical 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). We thus write $\mathcal{S}(q, \dot{q}, s, \dot{s}, t)$.

At every point in time t each actor emits both a *physical action* u and a *speech action* v. The idea is that both the physical action and the speech action are caused by the contents of the actor's mind. Actions are assumed to be a continuous output of the actuators attached to the mind (e.g., motors for a robot, muscles for a human). Note that actions are consistent with the laws of traditional physics, so that e.g. every muscle contraction exerts a balanced set of forces.

Speech acts are assumed to be a point process that is instantaneous, and directed at some subset of the actors in the actor network. We assume for the moment that speech acts can be directed at arbitary subsets of the actor network, since this is basically possible given the current status of information technology.

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 symmetry 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 and their time derivatives, it will often prove useful to use a φ that is a permutation of the actors in the actor network - we call these *empathy transforms*.

The ethicophysical Lagrangian is then

$$\mathcal{E} = \mathcal{L} + \mathcal{S}.$$

The laws of motion and emotion together combine to establish the following:

$$\frac{\partial \mathcal{E}}{\partial x} = \frac{d}{dt}\frac{\partial \mathcal{E}}{\partial \dot{x}}.$$

Define the following quantity (the *Gallifreyan*):

$$\begin{aligned}\mathcal{G}(q, \dot{q}, s, \dot{s}, t) &= \mathcal{E} - \mathcal{E}_\varphi \\ &= \mathcal{S}(q, \dot{q}, s, \dot{s}, t) - \mathcal{S}(q, \dot{q}, \varphi(s), \frac{d}{dt}\varphi(s), t)\end{aligned}$$

By Noether's Theorem [9], the following quantity is conserved:

$$\sum_{i=1}^{k} \frac{\partial \mathcal{E}}{\partial \dot{q}_i}\tau_i + \sum_{j=1}^{n} \frac{\partial \mathcal{E}}{\partial \dot{s}_j}\tau_i.$$

By Noether's Theorem applied to the modified Lagrangian $\mathcal{E}_\varphi$, the same is true of the quantity

$$\sum_{i=1}^{k} \frac{\partial \mathcal{E}_\varphi}{\partial \dot{q}_i}\tau_i + \sum_{j=1}^{n} \frac{\partial \mathcal{E}_\varphi}{\partial \dot{s}_j}\tau_i.$$

Subtracting one from the other, we learn that the following quantity is conserved:

$$\sum_{i=1}^{k} \frac{\partial \mathcal{G}}{\partial \dot{q}_i}\tau_i + \sum_{j=1}^{n} \frac{\partial \mathcal{G}}{\partial \dot{s}_j}\tau_i.$$

Note that the flip φ is nonphysical, so that φ has no effect on the "objective", "physical" q's while the tweak τ probably does have an effect on the s's.

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 C_a be the quantity of self-respect that one feels for oneself at any given moment, defined as $l(a,a) - h(a,a)$. We can call this the *conscience* of the actor. (We note that it is definitionally equivalent to the subjective experience of the conscience with which most humans are familiar.)

Suppose the action taken has some victim b. Then take the flip φ to be the empathy transformation that swaps the subjective variables of a and b.

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, $\mathcal{G}$ should no longer depend on any q or on any s other than $\dot{l}(a,a)$, $\dot{l}(a,b)$, $\dot{l}(b,a)$, $\dot{l}(b,b)$, and the analogous dislike values.

There is then an additional conserved quantity, which is the *karma with respect to the flip* φ

$$K_\varphi = \sum_{x\in\{a,b\}} \sum_{y\in\{a,b\}} \frac{\partial \mathcal{G}}{\partial \dot{l}(x,y)} l(x,y) + \frac{\partial \mathcal{G}}{\partial \dot{h}(x,y)} h(x,y)$$

Let us assume that the potential subjective energy depends only on l and h, rather than on $\dot{l}$ and $\dot{h}$. Then we know that the partial derivatives come only from the active subjective energy terms, which yields the conserved quantity

$$\begin{aligned}
&= w_a \dot{l}(a,b) l(a,b) - w_b \dot{l}(b,a) l(b,a) - w_a \dot{h}(a,b) h(a,b) + w_b \dot{h}(b,a) h(b,a) \\
&= w_a (\dot{l}(a,b) l(a,b) - \dot{h}(a,b) h(a,b)) - w_b (\dot{l}(b,a) l(b,a) - \dot{h}(b,a) h(b,a))
\end{aligned}$$

If we define the quantity *opinion* to be

$$op(a,b) = \dot{l}(a,b) l(a,b) - \dot{h}(a,b) h(a,b),$$

then karma is just

$$K_\varphi = w_a op(a,b) - w_b op(b,a)$$

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. If we take the most obvious interpretation, it seems uninteresting because e.g. if you kill someone, their opinion of you might be thought not to be relevant any more. However, this is a naive reading of the conservation of karma. In reality, we all of us are always imagining what other people think and feel. In the post-Jesus world, most actors are aware enough of something like the Golden Theorem in order to be able to infer the opinion of people they hurt, and thus $op(b,a)$ matters to a even when b is dead.

This same principle can be applied to any binary decision. The total subjective energy 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 karma, as expected. These consequences are not necessarily irreversible; one can be forgiven sins, but in general only when one has overcome the sin and made recompense.

□

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 [4]:

> 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.

4.2 Does the Conscience have Momentum?

MORE TEXT HERE

Consider the following scene from the comic Girl Genius[3]:

> JAGER 1: Anodder shtupid easily-duped MINION! Don't you know dis iz an INSANELY dangerous guy?

AGATHA HETERODYNE: I KNEW THAT!

JAGER 2: Vell, let's just keel her.

GOOD GUY: FIENDS. Kill her and I'll tell the BARON.

JAGER 2: Vell, mebbe ve keel you too, schmot guy.

JAGER 1: Gorb...

JAGER 2: Vat!?

JAGER 1: GORB. Dis is turning into vun of DOSE plans...

JAGER 1: Hyu know - de kind vere ve keel everybody dot notices dot ve's killing people?

JAGER 2: It is?

JAGER 1: Uh huh. And how do dose always end?

JAGER 2: De dirigible is in flames, everybodyz dead an' I've lost my hat

JAGER 1: Dot's RIGHT. Und any plan vere you lose you hat iz?

JAGER 2: A bad plan?

JAGER 1: RIGHT AGAIN!

4.3 What does this have to do with AI Safety?

We present the following dialogue with Tom Silver [10]:

tom 21:24
what's the gist?
epurdy 21:24
um
21:24
i guess i need to write an abstract
21:25
but basically you can write down laws of ethics that are modeled on
 the laws of physics
tom 21:36
maybe you get to this later but why are love and hate defined in terms
of subsets?
21:36
rather than individuals
epurdy 21:36
well
21:36
you can hate british people, for instance
tom 21:36
so you can hate british people but you might not hate a specific few
of them?
epurdy 21:36
i should put in some sort of example like that
21:37
well you might not even know any british people
21:37
like i hate nazis, but i can't really name all of them
tom 21:37
but i mean if you defined love and hate in terms of individuals, one
might hate all british people individually
21:37
so it wouldn't preclude that
epurdy 21:37
right....
21:37
hm
21:38
i think my definition is still better because it respects the
cognitive limitations
21:38
of the human mind
21:38
i can hate british people, but i am not capable of hating them all
individually
tom 21:38
hm interesting

```
epurdy 21:38
a computer could probably hate them all individually though
tom 21:39
and so do you think it is possible to have that you hate a set of
people a certain amount but you e.g. don't hate one or two of them?
epurdy 21:39
hm
tom 21:39
i could see that making sense
epurdy 21:39
there are consistency issues i guess
21:39
if you hate british people with a passion but love your british wife
21:39
where does that leave one
tom 21:40
yeah
21:40
well certainly there are people in situations like that
epurdy 21:40
right but then the subset they hate is not really ``british people''
21:40
it's ``british people in general, not including my wife''
21:40
which is interesting
tom 21:40
yeah that's true
tom 21:46
so what's the connection to AI safety?
epurdy 21:46
\begin{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.
\end{abstract}
tom 21:47
gotcha
```

21:47
cool
epurdy 21:47
it's pretty good, right?
21:47
like i need to check the proofs a hundred more times but there's something there i think
tom 21:48
i'm still digesting
21:48
could you give an example of how this would actually turn into straightforward computations for AI safety?
epurdy 21:49
the AI could be given some mission in life
21:49
for instance, a babysitting AI could be given the mission to protect the relevant child at all costs, without breaking the law
21:49
or something
21:50
the point is that you can escape the evil genie problem
21:50
because you can reason about multiple goals and how the tradeoffs between them work
tom 21:50
what's the evil genie problem?
epurdy 21:50
well, if you say protect the relevant child at all costs
21:51
and someone online says something mean to the child
21:51
a naively programmed AI would find and execute that person
tom 21:51
but if you programmed the AI so that it wasn't allowed to break the law, wouldn't that be avoided in theory?
epurdy 21:52
but what if the law is fuzzy
21:52
or what if a law must be broken?
21:52
like you are running from a bad guy and you need to jaywalk in order to escape
tom 21:52
hmm i see, and how would ethicophysics deal with that scenario?
epurdy 21:53
it would make an easy snap judgment that no one is going to give a

shit about jaywalking if the kid is going to die
21:53
you can reason through shit like that over time, but ideally the AI would know the answer immediately without thinking about it
tom 21:53
how exactly though?
epurdy 21:54
so it is trying to maximize l(God, {robot}) - h(God, {robot}), say
21:55
and then it is doing planning in the RL setting, which we know will work when AI gets to that point
21:55
but it has a preposterous number of heuristics that it can apply
21:55
and the heuristics are actually provable laws of ethics derived from ethicophysical proofs that are mathematically tight
tom 21:56
but how would it calculate l(God, {robot}) - h(God, {robot})?
epurdy 21:56
well there it needs some sort of world model that includes the ethicophysics components
21:57
so we don't know yet exactly what htat would look like
21:57
but writing down a bunch of ethical laws that can be used in simulations can only help make matters better
tom 21:58
isn't the hardest part of the AI safety problem figuring out how to precisely describe what the AI should be optimizing though?
epurdy 21:58
yes!
21:58
which is why this is so exciting
21:58
it's the beginnings of a precise description of that
tom 21:59
but when i say
>precisely describe what the AI should be optimizing
what i mean is something like a complete definition of l(God, {robot}) and h(God, {robot}). like isn't that just a rephrasing of the question basically
epurdy 21:59
right, but now we have intermediate concepts that we are defining
21:59
like coherent extrapolated ethical momentum
22:00

these concepts allow us to build a common vocabulary with the AI
22:00
so we don't have to assume that some deep network somewhere knows what we want it to know
22:00
we can write automated theorem provers around the laws of ethicophysics and then get scrutable ethical proofs of what the right thing to do probably is
22:01
then we can test the shit out of those automated theorem provers
22:01
make sure they work using traditional software practices
22:01
then we should be good to go
tom 22:02
okay well a lot of that is still fuzzy for me but i have to jump off for the night, thanks for sharing! looking forward to talking more about it later
epurdy 22:02
cool
22:02
thanks for not calling me crazy

5 Epilogue

We find that the following lyrics of Yusuf Islam [5] 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

6 Acknowledgments

We would like to thank Tom Silver for helpful discussions, including the dialogue above.

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] Foglio, P., and Foglio, K. One of those plans. *Girl Genius 2* (2003), 69.

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

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

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

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

[8] Lactantius. *De Ira Dei*.

[9] Noether, E. Invariante variationsprobleme. *Nachr. D. König. Gesellsch. D. Wiss. Zu Göttingen, Math-phys. Klasse* (1918), 235–257.

[10] Silver, T., and Purdy, E. Some thoughts on ai safety. *Slack conversation* (2018).

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

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