Ethicophysics II: Affilliation Economics and Naturalistic Game Theory

Eric Purdy

November 28, 2023

Abstract

1 Introduction

In this document, we continue the work begun in Ethicophysics I. In Ethicophysics II we seek to understand the nature of ethics in the setting that actors can learn. There are two cases of interest, the supervised learning setting, and the reinforcement learning setting. We tackle first the supervised learning setting.

1.1 Affiliation Economics

We wish to specify a collection of distinct but intersecting economies. Each economy can be thought of as a game played between two opposing teams. Or, if you would rather, a war fought between two opposing armies. We prefer the game terminology, because it helps to remind us that win-win solutions exist in far more situations than most people would believe. Ultimately, fundamentally, peace is possible.

Classical economics is the study of the game between the haves and the have-nots, between the rich and the poor. I put it to you that it is a tolerably complete science, but that it is sorely lacking in its failure to understand any human value that cannot be quantified in dollars or utils or what-have-you. Affiliation economics is a sort of intellectual trick to try to transport the useful parts of economics to a setting in which the full spectrum of human values can be observed and reasoned about.

Fortunately for us, most of the concepts necessary to understand affiliation economics are already well-understood by various subcultures in society, and by various academic communities. In general, whenever two opposing forces learn of each other's existence, the initial result is rather unpleasant, and involves one of the sides brutally subjugating the other. Such is, unfortunately, the lesson that history has taught us. This dynamic is what I take to be the content of Hegel's treatment of the Master-Slave dialectic [1].

Since this brutal subjugation tends to resolve itself in favor of one side or the other, at least initially, each game can be thought of as having a “natural” “top” side, whose position is stronger. The advantages that accrue to those on the top side are what is generally referred to as “privilege”. The disadvantages that accrue to those on the bottom side include what is generally referred to as “stigma” and “marginalization”. Essentially, the main activity of most of those on the top side is reinforcing their power over the bottom side (“the rich get richer”), while the main activity of most of those on the bottom side is simply trying to get by under a crooked system.

1.2 Continuous Actor Space

The fundamental problem in affiliation economics that separates it from classical economics is the necessity of identifying which players are on which team, and to what extent. (The problem of identifying who has how much money is, as far as I know, not treated in great depth in the existing economics literature. There is, of course, a rich literature on “signaling” in classical game theory that one can draw on.) This is a problem that is so complex and thorny that most people who have studied it despair of ever sorting out the truth from the lies. And this worry is merited, but we can at least postulate a rather simple description of the space of possibilities. We refer to this postulated space as “continuous actor space”, because it describes the space of possible actors that can exist in reality. It is, of course, an incomplete description; ultimately, each human is a beautiful and infinite mystery. But I put it to you that certain large-scale regularities in human nature exist, and are necessary to explain the human capacity for making any sense whatsoever of the complex web in which we live.

Continuous actor space is actually quite a simple thing. We can posit two versions, the bounded and unbounded versions. The bounded version is simply $[-1, 1]^n$, where n is the number of affiliation economies we have chosen to include in our model. The unbounded version is simply $\mathbb{R}^n$. It is, of course, trivial to map back and forth between the two spaces via the tanh map, as is standard in deep learning.

The fundamental problem is then simply the problem of determining someone’s “true” place in continuous actor space given observations of their behavior over time. Since any action taken in public is known to be in view of others, fundamentally all actions taken in public are suspect; they function more as “signals” in the sense of classical game theory, and less as any reliable indication of the contents of an individual soul. We also have extensive evidence that so-called “moles” can survive for decades in intelligence services while performing their roles to an apparently acceptable level; this is extremely disheartening for anybody hoping to assemble any true picture of what people are up to. Ultimately, I put it to you, the only reliable indicators of what is going on inside an individual are as follows:

- Longitudinal observations from a very young age (a player can only be so good at playing a game for which they do not know the “true” rules, and such clumsiness reveals some amount about the inner workings of the soul)

- High-cost actions (i.e., actions which confer no conceivable benefit to the actor; this idea is well understood in classical game theory and evolutionary game theory)
- Total surveillance of every aspect of someone's life, even and especially their most private moments

Given standard American assumptions about civil liberties, the third possibility is probably not acceptable to most people. On the other hand, it is a rather natural result of allowing digital technology into our homes that organizations like the NSA will acquire such capabilities unless something rather radical is changed about how technology is funded and developed.

1.3 Motivating Example

Ultimately what we are curious about is where each agent is in continuous actor space. If we know that to any degree of certainty, we can come down like the fury of God on anyone who accumulates a position of influence who we do not trust to wield it justly. Or perhaps rather like a gentle breeze, that whispers in their ear the way that they can unfuck their soul.

Let us for the sake of exposition consider a 3-d ethicophysics: unskilled-skilled, evil-good, and poor-rich. Every agent thus has a evil-good score, a unskilled-skilled score, and a poor-rich score. (This latter being simply a bank account.) By convention, we place the bottom side of a soul axis on the left when talking about it, and use negative numbers for affiliation to the bottom side, and positive numbers for affiliation to the top side.

By convention, we consider good to be the top side of the evil-good soul axis. This seems justified by the ethicophysics in the modern context (specifically with non-local communication devices, good seems to have a natural advantage), but not in historical contexts before a certain point. At any rate, it feels less depressing to use that particular sign convention.

1.4 Natural Drives as Ethicophysical Fields

Whenever the agent wants something, a desire field comes into play. It is presumed that we can know which way approximately desire will pull the soul. In our simple 3-d ethicophysics, desire for posessions and influence pulls the soul in the evil-skilled-rich direction. Desire to be of service pulls the soul in the good-skilled-rich direction. Desire to be lazy pulls the soul in the poor-skilled direction. Desire to better oneself pulls the soul in the good-skilled direction. Desire to increase one's employability pulls the soul in the rich-skilled direction.

We can identify a number of desire fields that are so fundamentally good and useful and human that any goal-directed agent that doesn't have them is fundamentally incapable of coexisting with humans. This statement sounds a little scary in the era where artificial intelligence seems to be approaching, but it shouldn't causes us too much worry. Because humans are far and away the dominant life form on Earth, we are the primary obstacle standing between any goal-directed actor and its goals. Therefore, any artificial intelligence will at least need to reason about these desire fields to accomplish anything. We thus are no worse off than before, as long as people don't intentionally hook up an AI deliberately engineered to be evil to a powerful weapons system. This seems like a low enough bar for reasonability that even human beings might clear it eventually.

The following desire fields are identifiable and seem key to the human experience:

- Desire for truth
- Desire for goodness
- Desire for beauty
- Desire for status

1.5 Simulating Capitalism

Let us visualize the results of our simulation as follows: we simply project dots in a 3-d square representing bounded continuous actor space and/or a hyperbolic projection of unbounded continuous actor space. Then we perform a simple market simulation of workers seeking employment from firms. Firms have internal politics characterized by a shifting hierarchy DAG that represents who is the "real" boss of who. It is presumed that all workers are self-interested, and know the ethicophysics (and thus can reason quite well about the effect of various choices on their souls and future prospects).

The tricky thing here is that desire pulls the soul, but the position of the soul shapes desires. This is why the phenomenological tapestry is such a complex beast. We can simulate this in three alternating phases: first the position of each soul drives certain actions on the part of each agent. Then each agent gets some reward from the result of all actions taken. Then the gradient update from processing the reward of each agent drives a small update to the position of each soul. What we are postulating is that the gradient update from processing a reward will have a particular, predictable effect on the shape of the soul of the RL agent in question. This is an experimentally testable hypothesis, possibly.

What game should we choose to experiment with? Let us consider the following game. Each worker decides how hard to work, on a scale from -1 (strenuous sabotage) to 0 (apathy, laziness) to +1 (strenuous service). Each effort number is multiplied by the skill of the subject. The contributions from each team member are then passed up the chain to their boss, their boss's boss, etc. We thus wind up with a total output of the system. This output is presumed to be worth that amount of reward. Rewards are divided between bosses and their subordinates using games of Ultimatum, where workers rejecting the proposed split must quit and receive no future salary, while bosses whose employees quit receive no future labor.

2 Main Results

2.1 Defining the Desire Fields

We define a desire field by defining the potential opinionatedness function (which can also be called the potential subjective energy function). Once we have that, we can use standard Lagrangian mechanics to simulate what will happen. In general, the potential subjective energy will be a function

$$\mathcal{O}(q, s, t)$$

that measures how far the desired state of affairs is from holding. This is generally called a cost function in the computer science and control theory literature.

In general, we would like to be able to achieve optimal control over the ethical domain in order to avoid hurting other actors. Because free will exists and actors other than ourselves exist, this can only be achieved to a certain extent, given the limits of mathematics, computer science, and existing computational devices. But it makes sense to start our investigations in a setting where optimal control is possible, which is the Linear-Quadratic Regulator setting: specifically, a dynamical system where the dynamics are linear and the costs are quadratic. (We would also like to extend to the Linear-Quadratic-Gaussian setting, to handle uncertainty about the state of affairs.) Given that things are easiest to reason about when the laws of physics are time-invariant, the cost function should not be time-varying.

We therefore specify a set-point $[q_0 \quad s_0]$ that is desired, define an offset $[\vec{q} \quad \vec{s}] = [(q - q_0) \quad (s - s_0)]$ and examine cost functions of the form

$$[\vec{q} \quad \vec{s}]Q[\vec{q} \quad \vec{s}]^T.$$

2.1.1 The Active Subjective Energy of the Truth Field

Let $f_1, \ldots, f_k$ be potentially true facts about the world and its history. Each $f_i(q, s, t)$ is a function of the state of the physical world, the state of the ethical world, and the time t. Each f_i takes on values between -1 (totally, inarguably false) and 1 (totally, obviously true). Let s_{ai} be the belief of actor a about fact f_i, expressed as a real number between -1 (total disbelief) and 1 (total belief). Let w_i be a weighting function that assigns importance to individual facts.

The active subjective energy of the truth field is called *active bullshit*, and has the following functional form:

$$\mathcal{O}_{pot}^{truth}(q, s, t) = \frac{1}{2} \sum_{a,i} w_i (s_{ai} - f_i(q, s, t))^2.$$

Note that, like most concepts in the ethicophysics, active bullshit can only be measured relative to some subjective assignment of importance.

The most actively bullshit thing of all time might be that Jesus was deemed worthy of crucifixion, which was a death reserved for slaves and dangerous criminals, of which Jesus was neither (well, the historical Jesus may have been a dangerous revolutionary, but the retconned Jesus was neither). Let $f_+(q, s, t)$ be the statement "Jesus is a slave or dangerous criminal". It seems safe to say that God never thought so, so $s_{\mathrm{God},i}$ can be taken to be constant at -1. And yet everyone watching had undeniable evidence that the supposedly legitimate authorities either believed f_+ to a very high level of certainty, or simply did not care about the truth of the matter at all (i.e., the w_+ values is very low in the subjective weighting of importance assigned by the Roman authorities). ("What is truth", Pontius Pilate is reported to have said.) The weighting w_+ that most observers sympathetic to Jesus supplied was immense, because they believed him to be the messiah promised by the Jewish prophetic tradition. The ethicophysics argues, then, that crucifying Jesus was the equivalent of shorting the wires of a circuit with toweringly high voltage differences.

We can formalize this as follows:

Theorem 2.1. *Saying something untrue that hurts people will cause people to turn on you in the long run.*

Proof. Let $f_i(q, s, t)$ be a true fact that was contradicted through some action. Let the tweak τ be the identity, i.e., let us not try to enforce any physical symmetries.

We invoke the Golden Theorem with $\tau(q) = 0$ and

$$\varphi(s_{aj}) = \begin{cases} -s_{aj} & j = i \\ s_{aj} & \text{otherwise} \end{cases}$$

□

3 The Potential Subjective Energy of the Truth Field

We define *potential bullshit* to be the potential energy term that allows total bullshit to be conserved. Specifically, consider consensus reality, and model assertions as having a specific Glicko rating corresponding to how often they are held to be true in public debates with other, incompatible assertions.

We then note that, as the name suggests, things held to be true with high certainty in consensus reality are potentially large sources of "dark bullshit", i.e., bullshit that God knows is bullshit but which humans do not know. We can thus model the potential bullshit of an assertion by treating its Glicko rating as a height above sea level and using the standard $m \cdot g \cdot h$ potential term used in approximating gravity near the Earth's surface. Thus, the potential bullshit of a true assertion, say “the Earth is an irregular oblate spheroid” is extremely high, like the gravitational potential energy of a satellite at apogee, while the potential bullshit of a false assertion, say “the Earth is flat”, is extremely low, like that of the gravitational potential energy of a dead fish resting on an undersea vent in the Marianas trench. This all seems to fit reasonably well with standard human intuitions about the nature of consensus reality.

4 Open Questions, Exercises, and Assigned Reading

We conclude with a set of exercises for readers to expand and test their understanding of the concepts in Ethicophysics I and Ethicophysics II. Remember to make historically and aesthetically grounded ethicophysical claims and validate all answers with worked-out examples using the laws of ethicophysics to analyze the character (=ethicophysical momentum) of various historical and fictional personalities.

If you complete any of these tasks and write up your findings, I would love to see them, and would happily link to any quality findings in my blog. Please send me any writings or other work that you feel addresses these questions. I can be reached on LessWrong as MadHatter, or via email if you know my email address.

1. Read Keat’s Ode on a Grecian Urn, reflect on the story of Dr. Faust by Goethe, and then formulate a short description of Keat’s Fallacy.

2. Read Friedrich Schiller’s “On the Aesthetic Education of Man”. Why would Keat’s Fallacy be relevant to moral actions in the world?

3. Read “Avant-Garde and Kitsch” by Clement Greenberg (1939). Read any honest description of Adolf Hitler’s artistic style and level of skill. Then formulate working definitions of active kitsch, potential kitsch, describe how to measure both, and prove that total kitsch is a conserved quantity. Further, reflect on the question of whether the art school that denied Hitler admission made the right or the wrong decision for German society.

4. Read T. H. White’s “The Sword in the Stone”. Then explain the concept of the “might makes right makes might” cycle. Prove that “might” is a conserved quantity, and come up with an empirical framework for measuring active and potential might. If a war is won by the side that coordinates more effectively, and if any sane nation is sure to propagandize its citizens into believing that it won the last war it won via moral persuasion rather than force of arms, then how true is the statement “good will always prevail over evil”? Give examples and counterexamples to this assertion. Are there any regularities in the temporal distribution of examples and counterexamples? What is the significance of those regularities, if you observe any?

5. Write a python script to scrape Wikipedia's list of wars and generate Glicko ratings for all actors who participated in at least one war in history. What regularities and interesting patterns do you note in the answer? For instance, where do militant Islamic groups tend to fall? Where do large and heavily armed nation states tend to fall? Characterize the top decile and bottom decile of the rankings. Characterize the middle two deciles of the rankings.

6. Read "White Privilege: Unpacking the Invisible Knapsack" by Peggy McIntosh. Then define active and potential privilege, the subjective energy quantities associated with the status field. Provide a critique grounded in Cultural Marxism of this mathematics. Then provide a critique from within Jordan Peterson's point of view of your previous critique. What conclusions do you draw about the proper management and use of privilege? Is privilege a real thing or a bogeyman invented by social justice warriors? What are your thoughts on the concept of noblesse oblige?

7. Derive a simple and efficient affiliation estimation algorithm that you think would provide reasonably correct Bayesian estimates of people's internal felt sense of what team they themselves are on. How robust do you think your algorithm would be in the face of active Goodharting by an adversary that is trying to virtue signal? Can you think of any way to make your algorithm more robust? What conclusions do you draw about the concept of virtue signaling?

8. Listen to Pete Seeger's song "Solidarity Forever" and reflect on the nature of coordination mechanisms in games of Ultimatum between labor and capital. What conclusions do you draw about Marx's labor theory of value? Are these the same as your previously held conclusions about the labor theory of value? Why or why not? If not, why do you think you previously held a different belief? Can you detect the presence of bullshit in your change of opinion, or in your failure to change your opinion?

9. Watch the movie "The Death of Stalin", and reflect on the moral failings of communism, as enumerated in Arthur Koestler's "Darkness at Noon", or "One Day in the Life of Ivan Denisovich", by Aleksandr Solzhenitsyn. Which side of the Cold War (if any) would the ethicophysics label the Good side? How sure are you of this conclusion?

10. Reflect on the nature of optimal ethicophysical strategy in a tripolar conflict, as reflected in George Orwell's 1984. Then explain the significance of Nixon's detente with Communist China. Was Nixon right to go to China? Why or why not?

11. Reflect on the Cultural Revolution and any other events of the period of Mao's dominance in Communist China. Was Mao a Good leader or an Evil leader in your opinion? How do the contents of the Little Red Book read in light of your opinion?

12. Reflect on the concepts of honor (the opinion of the self about the self) and reputation (the opinion of others about the self) in light of Lois McMaster Bujold's novel The Curse of Chalion. Do enough historical reading to figure out the historical sources of The Curse of Chalion's macro plot. Was the historical inspiration for the character of Iselle a Good leader or an Evil leader, in your opinion?

13. Watch the movie "Lincoln". Try to divide the ethicophysics of the American Civil War into at most 4 or 5 teams. Which side of the American Civil War would you consider the Good side? Which of the 4 or 5 teams would you consider to be Good or Evil? How would you rank the teams by virtue?

References

[1] Hegel, G. W. F. *Phänomenologie des Geistes*, vol. 2. Duncker und Humblot, 1841.