Ethicophysics II: Affilliation Economics and Naturalistic Game Theory

Eric Purdy

May 29, 2018

Abstract

1 Introduction

In this brief document, I lay out the bones of a proposed theory of human nature suitable for allowing certain useful computations to be performed. Nothing in this document contradicts the existence of free will, but its main thrust is to argue that people can be manipulated somewhat well through the clever use of microtargeted lies, and even better through the use of misleading microtargeted truths. This is rather frightening to me, as I am at root a rather honest man. I ask that anyone reading this use it to fight for truth and justice, but I realize that this is, at root, a fool's hope. Ultimately, evil wants what it wants, and it will take what it can.

On the plus side, I think you will find that telling the truth can be more effective as a tactic than cynics would have you believe.

2 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 tolearbly 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.

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.

3 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.

4 Main Results

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: good-evil, skilled-unskilled, and rich-poor. Every agent thus has a good-evil score, a skilled-unskilled score, and a rich-poor score. (This latter being simply a bank account.)

4.1 The Desire Field

Whenever the agent wants something, the 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 perhaps identify other potential desires, but these seem like enough for now.

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.