Is it in our self-interest to being good?*

#20387090

June 29, 2015

We explore this problem from the perspective of how a rational agent would act, and their problem solving method. And as a follow up, the diligent reader would notice that there are surely many different methods to solving a problem. Some of those methods would involve being good, and some not. What do we do when it is possible to achieve the same goals through just and unjust means? In response to this, we explore the problem through an example of an unjust man versus a just man, and how it is easier to be good.

The main arguments are drawn from the course textbook[1], with secondary arguments from an artificial intelligent textbook[2], the readings "The Ring of Gyges" and "Collective Action Problems".

Definition of *our*. We define the term *our* in the question "is it in our self interest to be good?" to mean a *rational* agent. Being *rational* is a fundamental assumption we make, in the sense that human intelligence is plagued with emotional conflicts, resulting in irrational decisions. Hence it is in our self-interest to avoid acting upon emotional decisions, and avoid its pitfalls by being rational. We also define *good* later on, once we've established a framework.

From a Rational Agent's Perspective

First, let's look at the reasons why it would be in a rational agent's self-interest to be good, in to reap better rewards in the future, and **secondly**, why it benefits them more as a whole from being good.

Definition of *good*. We're now ready to define what *good* means, in "is it in our self-interest to be good": being *good* means not being selfish and always maximizing the reward for yourself. Being *rational* and not necessarily picking the most seemingly beneficial action at the moment.

Firstly, why it would be in a rational agent's self-interest to be good, due to future rewards. From our definition of what it means to be good, it follows that we shouldn't be *bad* (not good), which is being selfish and always picking the action that maximizes the current reward. In other words, being *bad* is being *always greedy*. Let's start with an important foundation from the course textbook to build upon: the underlying commonalities from different social theories[3]:

> The key to understanding them [social negotiation theories], however, lies in the idea that the contractors are, above all, *rational and self-interested*.
>
> Being self-interested is not the same thing as being selfish. Being self-interested is having a strong concern for how you well you are faring in life. Being selfish is placing far too much importance on your own well-being relative to the interests of others.

*word count: 1953

[1] Shafer-Landau, R. (2014). The Fundamentals of Ethics. Oxford University Press

[2] Russell, S.J. & Norvig, P. (2010). Artificial Intelligence: A Modern Approach. Prentice Hall

[3] page 215 of The Fundamentals of Ethics

What it means is that there are future benefits of being good, at a cost of gaining less benefit at-the-moment. That is, we should be *self-interested* instead of being *selfish*. Being selfish maximizes the reward we get *now* but at a cost in the *future*. This is also seen in the AI textbooks, through the notion of a reward function[4], measuring how much reward was attained from an action. Our self-interest is to choose the action that attains the most reward, i.e. the optimal action to take is to take the action that maximizes the benefit now, and in the future. And "[r]epeated experiments show that the greedy agent very seldom converges to the optimal policy"[5]. Essentially, the AI textbooks also agree with the idea from the philosophy textbook – the idea that you should be *rational and self-interested* but not *selfish*.

Secondly, why it rewards rational agents more as a group from being good. In other words, we'll look at how rewards increase if rational agents cooperate with each other and be good. It is easy to understand that when others have good opinions of you and you have the trust of others, this opens doors, and you get a wider choice of options to choose from. offers more rewards for the collective group than being bad (greedy). We can look at what happens when everyone is good, when everyone is good except you, when everyone is bad except you, and lastly, when everyone is bad. To the diligent reader, this indeed looks like a Prisoner's Dilemma. And indeed we will look at Nash Equilibriums. To explore this, we'll look at why behaving badly is a bad strategy in the form of a collective action problem.

Collective Action Problem In the Collective Action Problem reading[6], people seem to behave irrationally because it seems like the best action, at a cost to others. This problem looks like the Prisoner's Dilemma problem, which we analyze from the rational agent perspective. This reading suggests that we are lazy individuals who don't want to be the one putting in the effort to be good, unless everyone is doing good. And if everyone is doing good, then certainly it is going be in our self-interest to sit back and reap the fruits of other's labour. So this seems to be a catch-22 problem.

While you can reap rewards by being bad, it is only temporarily so. Assuming that people are not fools and stupid (but certainly not always rational agents) is quite reasonable, then it's reasonable to assume others learn eventually of your tactics, and the whole system collapses. There won't be any fruits of labour to reap. Everyone being good is an unstable equilibrium, like the Prisoner's Dilemma.

Thus it is in our self-interest to be good and help maintain a good, stable equilibrium and cooperate, so that the collective benefit is greater than individual greedy-action benefits, which essentially the Nash equilibrium conclusion.

We argued these points through analysis of rewards to a rational agent, but the keen reader will notice that being bad in a Nash Equilibrium is a dominate strategy. This is true. However, the focus of this section is (cleverly) regarding the rewards to rational agents as a group. Hence it is still a bad strategy. Well why should the rational agent care about other rational agents? The answer to this is that other rational agents are also rational, meaning they know your strategy and so you can't take advantage of them.

Results.

The reader should see that it is in a rational agent's self-interest to be good because it is more rewarding, for the above two reasons. First, making bad (not good, greedy) decisions decreases future expected rewards. Second, rational agents in a group forms a Nash Equilibrium, which is inherently unstable, so not being good destroys the cooperation with other rational agents and results in overall less reward.

[4] Also known as an utility function.

[5] pg. 839 AIMA 3rd ed.

[6] p. 366 Constellations Volume 7, Number 3, 2000 Ideology and Irrationality. http://homes.chass.utoronto.ca/~jheath/ideology.pdf Accessed June 3, 2015

Archiving Same Goals Through Just vs. Unjust Means

This section explores the issue of different methods to achieve the same goals, but the methods have a preference order in goodness.

Roughly translated, being *just* translates to being *good*, and *unjust* means up to no good. A reading which fits nicely here is "The Ring of Gyges" by Plato[7], an example of an unjust man vs. a just man, both meeting their goals.

In the reading, it explores a sly, unjust man versus a righteous and just man. Both of whom theoretically can achieve the same goals but through different methods. A man of the uttermost injustice and deceit can live a life as good as a just man achieving the same goals through just acts, through perfect deceit. Intuition tells us that archiving the same end results through deceitful means is certainly worse than archiving it through honorable means. And indeed Plato argues for this, as the course textbook summarizes it well[8].

> Certainly many immoral people are deeply troubled and unhappy. But others are able to sleep well at night, take pride in a job well done (assassination, theft, betrayal), and find friends within a network of like-minded associates. The bad guys sometimes get away with it, having a lot of fun in the meantime, and never regret the harm they have caused.

Certainly based on pure accomplishments, the two man are the same, since both can accomplish their desired goal through just or unjust means. But the primary difference is that there is a chance for the dishonest man to get caught and (heavily) penalized. And sure, certainly the bad guys sometimes get away with it. A major assumption here is that they can get away with it.

Let's break the argument into two pieces: suppose he fails sometimes in lying, and suppose he never fails in lying.

He might fail at deceit. This is the more realistic argument. Realistically speaking, the underlying assumption is unattainable. No man can deceive the entire life from the moment they were born. With such an outrageous assumption, we might as well assume he was born with everything he could ever want in life. Accomplishing the same goal but taking the unjust method to do it translates to taking more risk, but at a chance to gain higher reward.

But here is the kicker: there is no difference in the reward. We had assumed the two man achieved the same goal but with different means. Then obviously it makes no sense to pick the riskier path. *All the deceitful man has done is save some time and effort.* Which may be valuable, but at a heavy price if caught.

He never fails at deceit. Even if we make the (unlikely) assumption the unjust man is capable of maintaining his composure at all times and not fall out of character, it still doesn't make sense why to take the riskier path as it leads to the same rewards. All he has done is save some time and effort. Seems that this assumption is quite difficult to accomplish (if not impossible) for the average layman. Sure, there may exist those of us that are deceptive enough to carry this out, but they mostly exist for the sake of argument. It is likely the assumption is a fallacy in itself, but cannot be proved.

Results.

A lot of the arguments explored here are suppositions, and not as solid in logic analysis as from the rational agent perspective. But nevertheless, the fundamental argument in this section is to show that it's much easier to maintain a righteous and just composure than a deceitful and unjust one, as

[7]"The Ring of Gyges" by Plato - Philosophy Lander.edu http://philosophy.lander.edu/intro/articles/gyges-a.pdf Accessed June 2, 2015.

[8]page 108 of The Fundamentals of Ethics

in most people find it difficult to maintain a deceitful composure and accept its risks. Consequently, if we act just, then it is in our self-interest to be good.

Conclusion

We've now explored this problem from the perspective of how a rational agent would act, and their problem solving method. It is in a rational agent's self-interest to be good because it is more rewarding, because making bad (not good, greedy) decisions decreases future expected rewards, and not being good destroys the cooperation with other rational agents and results in overall less reward.

And we've followed up on what do we do when there are many different methods to solving a problem and some involve being good and some not. It seems that the underlying assumption is reasonably invalid, and even if it was true, there isn't much to be gained.

The common result from the analysis suggest being good is an optimal choice.