Is it in our self-interest to being good?[Is it in our self-interest to being good?[word count: 1834 | word count: 1953]]

20387090 #20387090

July 27, 2015 | June 29, 2015

It is in our self-interest to be good. I w | We explore this problem from the perspecti through two parts. First, being good reaps | agent would act, and their problem solving rewards. Second, being good benefits more | follow up, the diligent reader would notic individually. | many different methods to solving a proble > methods would involve being good, and some > when it is possible to achieve the same go > unjust means? In response to this, we expl > an example of an unjust man versus a just > easier to be good.

The main arguments are drawn from two text | The main arguments are drawn from the cour textbook, The Fundamentals of Ethics [foo < Shafer-Landau, R. (2014). The Fundamentals Shafer-Landau, R. (2014). The Fundamentals University Press University Press], and the textbook used in CS 486 Introdu |], with secondary arguments from an artifi Intelligence, called Artificial Intelligen | textbook[footnote: Russell, S.J. & Norvig, P. (2010). Artific Russell, S.J. & Norvig, P. (2010). Artific Modern Approach. Prentice Hall Modern Approach. Prentice Hall] (referred as AIMA). There are similar id |], the readings "The Ring of Gyges" and "C fields, different approaches, but same con | Problems".

As in the case with mathematical proofs, I | Definition of our. definitions to the thesis. <

Definition of our | We define the term our in the question "is | interest to be good?" to mean a rational a The term our refers to a rational agent. A | a fundamental assumption we make, in the s that always acts to achieve the best outco | intelligence is plagued with emotional con uncertainty, the best expected outcome[foo | irrational decisions. Hence it is in our s Definition from page 4 of AIMA | acting upon emotional decisions, and avoid]. This simplifying assumption is made bec | rational. We also define good later on, on results in irrational decisions. | framework. | The philosophy text defines rationality (a | From a Rational Agent's Perspective theory) slightly differently. Rational mea | on one's preference orderings. | First, let's look at the reasons why it wo | agent's self-interest to be good, in to re Both are similar, but I highlight that the | the future, and secondly, why it benefits includes acting under uncertainty. This wi | from being good. | Definition of good | Definition of good. | Being good means having insight and lookin | We're now ready to define what good means, term. Consequently, as I will explore late | self-interest to be good": being good mean selfish (which is always trying to maximiz | always maximizing the reward for yourself. yourself). | necessarily picking the most seemingly ben | moment. Part I: Long term rewards | | Firstly, why it would be in a rational a The key

insight into gaining more long-ter | be good, due to future rewards. choosing the best action (being selfish) d | the best rewards. Some actions do not seem | From our definition of what it means to be beneficial action at the moment, but it cr | we shouldn't be bad (not good), which is b opportunities. | picking the action that maximizes the curr | words, being bad is being always greedy. L Philosophy Textbook | important foundation from the course textb | underlying commonalities from different so The first argument is from the course text < action isn't to be greedy, but to be self- < < The previous sections talk about different < this quotation below concludes about the u < from those different social theories[footn < page 215 of The Fundamentals of Ethics page 215 of The Fundamentals of Ethics]. I won't go in detail about what exactly |]: theories are, but merely look at the textb | on their shared commonalities: | The key to understanding them [social nego | however, lies in the idea that the contrac The key to understanding [the social negot | rational and self-interested. in the idea that contractors are, above al < self-interested. <

Being self-interested is not the same thin Being self-interested is not the same thin Being self-interested is having a strong c Being self-interested is having a strong c you are faring in life. Being selfish is p you are faring in life. Being selfish is p importance on your own well-being relative importance on your own well-being relative others. others.

While being selfish maximize your own rewa | What it means is that there are future ben theories have similar commonalities in tha | a cost of gaining less benefit at-the-mome isn't being selfish but rather being self- | be self-interested instead of being selfis saying the best action isn't the same as t | maximizes the reward we get now but at a c | is also seen in the AI textbooks, through This supports the idea that there are futu | function[footnote: good, at a cost of gaining less benefit at | Also known as an utility function. selfish maximizes the present-moment rewar |], measuring how much reward was attained future rewards. | self-interest is to choose the action that | reward, i.e. the optimal action to take is Artificial Intelligence Textbook | that maximizes the benefit now, and in the | [r]epeated experiments show that the greed The second argument is from the AIMA textb | converges to the optimal policy"[footnote: reinforcement learning, supporting the ide < isn't always best for yourself. < < I define some terms for readers in this pa < learning agent has a fixed policy that det < whereas an active agent must decide what a < utility function (also known as reward fun < the agent's performance. A state encloses < action. When actions and the search space < utility function can be learned by value i < < The experiment setup is to finding an opti < grid. But the grid is non-deterministic. E < take may or may not result in the desired < choses to move left, but the move happens < < Take it without proof that in this setup, < be extracted by one-step look-ahead to max < utility. After the 276ˆ{th} < value iteration and learning the < optimal policy at each step, the policy co < sticks to using that policy, never learnin < other states. We call this agent the greed < chooses is optimal for each state (plus on < interestingly, the agent does not learn th < true optimal policy. < < Repeated experiments show that the greedy < converges to

the optimal policy[footnote: < pg. 839 AIMA 3rd ed. pg. 839 AIMA 3rd ed.]! |]. Essentially, the AI textbooks also agre | the philosophy textbook – the idea that yo How can it be that choosing the optimal ac | and self-interested but not selfish. suboptimal results? The answer is that the | the same as the true environment. An agent | Secondly, why it rewards rational agents between exploitation (to maximize its rewa | being good. maximize its long-term well-being). Pure e | getting stuck in a rut. Pure exploration t | In other words, we'll look at how rewards knowledge is of no use if one never puts t | agents cooperate with each other and be go practice. | understand that when others have good opin | have the trust of others, this opens doors The greedy agent's selfish strategy to alw | choice of options to choose from. offers m rewarding action doesn't imply the best re | collective group than being bad (greedy). out on other opportunities. The optimal ac | happens when everyone is good, when everyo of benefits from the future, through explo | when everyone is bad except you, and lastl | bad. To the diligent reader, this indeed l Part I: Conclusion | Dilemma. And indeed we will look at Nash E | this, we'll look at why behaving badly is The two textbooks shares the conclusion th | form of a collective action problem. isn't necessarily one that maximizes your < one that factors in the future long term r < < Part II: Collective benefits < < The main idea in this part is it is in our < because being good benefits the collective < Rewards to a group of rational agents incr < < I explore the readings from Collective Act < the problem as a prisoner's dilemma proble < case of Nash equilibrium. Then applying th < learned about the prisoner's dilemma. <

Collective Action Problem Collective Action Problem

In the Collective Action Problem [footnote | In the Collective Action Problem reading[f p. 366 Constellations Volume 7, Number 3, p. 366 Constellations Volume 7, Number 3, Irrationality. Irrationality. http://homes.chass.utoronto.ca/~jheath/ide http://homes.chass.utoronto.ca/~jheath/ide 3, 2015 3, 2015], people behave irrationally because it i |], people seem to behave irrationally beca but at a cost to others. | best action, at a cost to others. This pro | Prisoner's Dilemma problem, which we analy The reading suggests that we are lazy indi | agent perspective. This reading suggests t to be the one putting in the effort to be | individuals who don't want to be the one p is being good. And if everyone is being go | be good, unless everyone is doing good. An is easier to not be good and reap the frui | good, then certainly it is going be in our catch-22 situation. | back and reap the fruits of other's labour | catch-22 problem. This problem can be casted to a prisoner's | well known case of Nash equilibrium, so we | While you can reap rewards by being bad, i learnt from prisoner's dilemma. The cast i | so. Assuming that people are not fools and | not always rational agents) is quite reaso 1. Everyone being lazy and you are lazy be | reasonable to assume others learn eventual both defect. | the whole system collapses. There won't be | to reap. Everyone being good is an unstabl 2. Everyone being good and you are lazy be | Prisoner's Dilemma. friend cooperates and you defect. | | Thus it is in our self-interest to be good 3. Everyone being good and you are good be | good, stable equilibrium

and cooperate, so both cooperate. | benefit is greater than individual greedy- | essentially the Nash equilibrium conclusio 4. Everyone being lazy and you are good be | friend defects and you cooperate. | We argued these points through analysis of | agent, but the keen reader will notice tha Now I apply results learnt from analysis o | Equilibrium is a dominate strategy. This i dilemma. | focus of this section is (cleverly) regard | rational agents as a group. Hence it is st Prisoner's Dilemma | Well why should the rational agent care ab | agents? The answer to this is that other r A key assumption made is the prisoners hav | rational, meaning they know your strategy reward or punish their partner and that th | advantage of them. affect their reputation in the future. | | Results. This assumption is not realistic in the re | punishment is death. Similarly, this appli | The reader should see that it is in a rati the collective action problem that you won | self-interest to be good because it is mor cooperating. Realistically there would be | above two reasons. First, making bad (not from the collective good on your behaviour | decreases future expected rewards. Second, reward, thus it benefits you to cooperate | group forms a Nash Equilibrium, which is i | not being good destroys the cooperation wi But what if the assumption holds? | agents and results in overall less reward. | Since betraying a partner offers a greater | Archiving Same Goals Through Just vs. Un cooperating with him, all purely rational | betray the other, and so the only possible | This section explores the issue of differe prisoners is for them to betray each other | the same goals, but the methods have a pre the time element is added to the considera | goodness. iterated prisoner's dilemma, then cooperat | rational outcome, as well explained in a s | Roughly translated, being just translates version of prisoner's dilemma, in the Mult | unjust means up to no good. A reading whic chapter [footnote: | The Ring of Gyges" by Plato[footnote: p. 118 Multi-agent Interactions, in An Int | "The Ring of Gyges" by Plato - Philosophy Systems. | http://philosophy.lander.edu/intro/article > June 2, 2015. >], an example of an unjust man vs. a just > goals. > > In the reading, it explores a sly, unjust > and just man. Both of whom theoretically c > goals but through different methods. A man > injustice and deceit can live a life as go > achieving the same goals through just acts > deceit. Intuition tells us that archiving > through deceitful means is certainly worse > through honorable means. And indeed Plato > course textbook summarizes it well[footnot > page 108 of The Fundamentals of Ethics].].

The game of the prisoner's dilemma is play | Certainly many immoral people are deeply t Each play is referred to as a round. Criti | But others are able to sleep well at night that each agent can see what the opponent | well done (assassination, theft, betrayal) round: player i | within a network of like-minded associates can see whether j | sometimes get away with it, having a lot o defected or not, and j | and never regret the harm they have caused can | see whether i | Certainly based on pure accomplishments, t defected or not. Now, for the sake of arg | same, since both can accomplish their desi assume that the agents will continue to pl | or unjust means. But the primary differenc every round will be followed by another ro | chance for the dishonest man to get caught

assumptions, what is the rational thing to | penalized. And sure, certainly the bad guy | with it. A major assumption here is that t If you know that you will be meeting the s | it. rounds, the incentive to defect appears to | diminished, for two reasons. | Let's break the argument into two pieces: | sometimes in lying, and suppose he never f Reason 1: If you defect now, your opponent | defecting. Punishment is not possible in t | He might fail at deceit. dilemma. | | This is the more realistic argument. Reali Reason 2: If you 'test the water' by coope | underlying assumption is unattainable. No receive the sucker's payoff on the first r | entire life from the moment they were born are playing the game indefinitely, this lo | outrageous assumption, we might as well as util) can be 'amortized' over the future r | everything he could ever want in life. Acc | goal but taking the unjust method to do it When taken into the context of an infinite | more risk, but at a chance to gain higher long) run, then the loss of a single unit | represent a small percentage of the overal | But here is the kicker: there is no differ if you play the prisoner's dilemma game in | had assumed the two man achieved the same cooperation is a rational outcome. | means. Then obviously it makes no sense to | All the deceitful man has done is save som The summary is that selfish rationality cr | Which may be valuable, but at a heavy pric possible result. Maximizing individual rew | possible reward. There are more benefits, | He never fails at deceit. collectively, if they both cooperated. Thi | that being good benefits more as a whole, | Even if we make the (unlikely) assumption individually. | capable of maintaining his composure at al | out of character, it still doesn't make se Interestingly, humans display a systematic | riskier path as it leads to the same rewar cooperative behavior in this, which goes a | save some time and effort. Seems that this rationales[footnote: | difficult to accomplish (if not impossible Tversky, Amos; Shafir, Eldar (2004). Prefe | layman. Sure, there may exist those of us similarity: selected writings. (PDF). Mass | enough to carry this out, but they mostly Technology Press. Retrieved July 26, 2015. | argument. It is likely the assumption is a]. It makes sense because the idea of rewa | cannot be proved. into play, because the assumption that pri | to reward or punish their partner is not r | Results. | Part II: Conclusion | A lot of the arguments explored here are s | solid in logic analysis as from the ration A rational agent's self-interest is to be | But nevertheless, the fundamental argument rewarding. Rational agents in a group maki | show that it's much easier to maintain a r Nash equilibrium, so being good means coop | composure than a deceitful and unjust one, more rewards. In the face of punishment fr | it difficult to maintain a deceitful compo selfish decisions decreases future expecte | risks. Consequently, if we act just, then other agents are less inclined to cooperat | self-interest to be good. defecting agent. <

Conclusion Conclusion

I conclude that it is in our self-interest | We've now explored this problem from the p analyzed in the above two parts. Part I co | rational agent would act, and their proble reaps more long term rewards, and Part II | in a rational agent's self-interest to be benefits more as a whole. | rewarding, because making bad (not good, g | decreases future expected rewards, and not | the cooperation with

other rational agents | less reward. | | And we've followed up on what do we do whe | different methods to solving a problem and | good and some not. It seems that the under | reasonably invalid, and even if it was tru References | be gained. < [1] < < [2] < < [3] < < [4] <

[5] | The common result from the analysis sugges > optimal choice.