```
Is it in our self-interest to being g   Is it in our self-interest to being g
word count: 1834                      | word count: 1953
]                                       ]

#20387090                               #20387090

July 27, 2015                         | June 29, 2015

It is in our self-interest to be good | We explore this problem from the pers
through two parts. First, being good  | agent would act, and their problem so
rewards. Second, being good benefits  | follow up, the diligent reader would
individually.                         | many different methods to solving a p
                                      > methods would involve being good, and
                                      > when it is possible to achieve the sa
                                      > unjust means? In response to this, we
                                      > an example of an unjust man versus a
                                      > easier to be good.

The main arguments are drawn from two | The main arguments are drawn from the
textbook,  The Fundamentals of Ethics <
Shafer-Landau, R. (2014). The Fundame   Shafer-Landau, R. (2014). The Fundame
University Press                        University Press
], and the textbook used in CS 486 In | ], with secondary arguments from an a
Intelligence, called Artificial Intel | textbook[footnote:
Russell, S.J. & Norvig, P. (2010). Ar   Russell, S.J. & Norvig, P. (2010). Ar
Modern Approach. Prentice Hall          Modern Approach. Prentice Hall
] (referred as AIMA). There are simil | ], the readings “The Ring of Gyges” a
fields, different approaches, but sam | Problems”.

As in the case with mathematical proo |   Definition of our.
definitions to the thesis.            <

  Definition of our                   | We define the term our in the questio
                                      | interest to be good?” to mean a ratio
The term our refers to a rational age | a fundamental assumption we make, in
that always acts to achieve the best  | intelligence is plagued with emotiona
uncertainty, the best expected outcom | irrational decisions. Hence it is in
Definition from page 4 of AIMA        | acting upon emotional decisions, and
]. This simplifying assumption is mad | rational. We also define good later o
results in irrational decisions.      | framework.
                                      |
The philosophy text defines rationali |   From a Rational Agent's Perspective
theory) slightly differently. Rationa |
on one's preference orderings.        | First, let's look at the reasons why
                                      | agent's self-interest to be good, in
Both are similar, but I highlight tha | the future, and secondly, why it bene
includes acting under uncertainty. Th | from being good.
                                      |
  Definition of good                  |   Definition of good.
                                      |
Being good means having insight and l | We're now ready to define what good m
term. Consequently, as I will explore | self-interest to be good”: being good
selfish (which is always trying to ma | always maximizing the reward for your
yourself).                            | necessarily picking the most seemingl
                                      | moment.
  Part I: Long term rewards           |
                                      |   Firstly, why it would be in a ratio
The key insight into gaining more lon |   be good, due to future rewards.
choosing the best action (being selfi |
the best rewards. Some actions do not | From our definition of what it means
beneficial action at the moment, but  | we shouldn't be bad (not good), which
opportunities.                        | picking the action that maximizes the
                                      | words, being bad is being always gree
  Philosophy Textbook                 | important foundation from the course
                                      | underlying commonalities from differe
The first argument is from the course <
action isn't to be greedy, but to be  <
                                      <
The previous sections talk about diff <
this quotation below concludes about  <
from those different social theories[ <
page 215 of The Fundamentals of Ethic   page 215 of The Fundamentals of Ethic
]. I won't go in detail about what ex | ]:
theories are, but merely look at the  |
on their shared commonalities:        | The key to understanding them [social
                                      | however, lies in the idea that the co
The key to understanding [the social  | rational and self-interested.
in the idea that contractors are, abo <
self-interested.                      <

Being self-interested is not the same   Being self-interested is not the same
Being self-interested is having a str   Being self-interested is having a str
```

you are faring in life. Being selfish
importance on your own well-being rel
others.

While being selfish maximize your own
theories have similar commonalities i
isn't being selfish but rather being
saying the best action isn't the same

This supports the idea that there are
good, at a cost of gaining less benef
selfish maximizes the present-moment
future rewards.

  Artificial Intelligence Textbook

The second argument is from the AIMA
reinforcement learning, supporting th
isn't always best for yourself.

I define some terms for readers in th
learning agent has a fixed policy tha
whereas an active agent must decide w
utility function (also known as rewar
the agent's performance. A state encl
action. When actions and the search s
utility function can be learned by va

The experiment setup is to finding an
grid. But the grid is non-determinist
take may or may not result in the des
choses to move left, but the move hap

Take it without proof that in this se
be extracted by one-step look-ahead t
utility. After the 276^{th}
 value iteration and learning the
optimal policy at each step, the poli
sticks to using that policy, never le
other states. We call this agent the
chooses is optimal for each state (pl
interestingly, the agent does not lea
true optimal policy.

Repeated experiments show that the gr
converges to the optimal policy[footn
pg. 839 AIMA 3rd ed.
]!

How can it be that choosing the optim
suboptimal results? The answer is tha
the same as the true environment. An
between exploitation (to maximize its
maximize its long-term well-being). P
getting stuck in a rut. Pure explorat
knowledge is of no use if one never p
practice.

The greedy agent's selfish strategy t
rewarding action doesn't imply the be
out on other opportunities. The optim
of benefits from the future, through

  Part I: Conclusion

The two textbooks shares the conclusi
isn't necessarily one that maximizes
one that factors in the future long t

  Part II: Collective benefits

The main idea in this part is it is i
because being good benefits the colle
Rewards to a group of rational agents

I explore the readings from Collectiv
the problem as a prisoner's dilemma p
case of Nash equilibrium. Then applyi
learned about the prisoner's dilemma.

you are faring in life. Being selfish
importance on your own well-being rel
others.

What it means is that there are futur
a cost of gaining less benefit at-the
be self-interested instead of being s
maximizes the reward we get now but a
is also seen in the AI textbooks, thr
function[footnote:
Also known as an utility function.
], measuring how much reward was atta
self-interest is to choose the action
reward, i.e. the optimal action to ta
that maximizes the benefit now, and i
[r]epeated experiments show that the
converges to the optimal policy"[foot

pg. 839 AIMA 3rd ed.
]. Essentially, the AI textbooks also
the philosophy textbook — the idea th
and self-interested but not selfish.

  Secondly, why it rewards rational a
  being good.

In other words, we'll look at how rew
agents cooperate with each other and
understand that when others have good
have the trust of others, this opens
choice of options to choose from. off
collective group than being bad (gree
happens when everyone is good, when e
when everyone is bad except you, and
bad. To the diligent reader, this ind
Dilemma. And indeed we will look at N
this, we'll look at why behaving badl
form of a collective action problem.

In the Collective Action Problem [foo
p. 366 Constellations Volume 7, Numbe
Irrationality.
http://homes.chass.utoronto.ca/~jheat
3, 2015
], people behave irrationally because
but at a cost to others.

The reading suggests that we are lazy
to be the one putting in the effort t
is being good. And if everyone is bei
is easier to not be good and reap the
catch-22 situation.

This problem can be casted to a priso
well known case of Nash equilibrium,
learnt from prisoner's dilemma. The c

1. Everyone being lazy and you are la
   both defect.

2. Everyone being good and you are la
   friend cooperates and you defect.

3. Everyone being good and you are go
   both cooperate.

4. Everyone being lazy and you are go
   friend defects and you cooperate.

Now I apply results learnt from analy
dilemma.

  Prisoner's Dilemma

A key assumption made is the prisoner
reward or punish their partner and th
affect their reputation in the future

This assumption is not realistic in t
punishment is death. Similarly, this
the collective action problem that yo
cooperating. Realistically there woul
from the collective good on your beha
reward, thus it benefits you to coope

But what if the assumption holds?

Since betraying a partner offers a gr
cooperating with him, all purely rati
betray the other, and so the only pos
prisoners is for them to betray each
the time element is added to the cons
iterated prisoner's dilemma, then coo
rational outcome, as well explained i
version of prisoner's dilemma, in the
chapter [footnote:
p. 118 Multi-agent Interactions, in A
Systems.

].

The game of the prisoner's dilemma is
Each play is referred to as a round.
that each agent can see what the oppo
round: player i
 can see whether j
 defected or not, and j
 can

In the Collective Action Problem read
p. 366 Constellations Volume 7, Numbe
Irrationality.
http://homes.chass.utoronto.ca/~jheat
3, 2015
], people seem to behave irrationally
best action, at a cost to others. Thi
Prisoner's Dilemma problem, which we
agent perspective. This reading sugge
individuals who don't want to be the
be good, unless everyone is doing goo
good, then certainly it is going be i
back and reap the fruits of other's l
catch-22 problem.

While you can reap rewards by being b
so. Assuming that people are not fool
not always rational agents) is quite
reasonable to assume others learn eve
the whole system collapses. There won
to reap. Everyone being good is an un
Prisoner's Dilemma.

Thus it is in our self-interest to be
good, stable equilibrium and cooperat
benefit is greater than individual gr
essentially the Nash equilibrium conc

We argued these points through analys
agent, but the keen reader will notic
Equilibrium is a dominate strategy. T
focus of this section is (cleverly) r
rational agents as a group. Hence it
Well why should the rational agent ca
agents? The answer to this is that ot
rational, meaning they know your stra
advantage of them.

  Results.

The reader should see that it is in a
self-interest to be good because it i
above two reasons. First, making bad
decreases future expected rewards. Se
group forms a Nash Equilibrium, which
not being good destroys the cooperati
agents and results in overall less re

  Archiving Same Goals Through Just v

This section explores the issue of di
the same goals, but the methods have
goodness.

Roughly translated, being just transl
unjust means up to no good. A reading
The Ring of Gyges” by Plato[footnote:
“The Ring of Gyges” by Plato - Philos
http://philosophy.lander.edu/intro/ar
> June 2, 2015.
>], an example of an unjust man vs. a
> goals.
>
> In the reading, it explores a sly, un
> and just man. Both of whom theoretica
> goals but through different methods.
> injustice and deceit can live a life
> achieving the same goals through just
> deceit. Intuition tells us that archi
> through deceitful means is certainly
> through honorable means. And indeed P
> course textbook summarizes it well[fo
> page 108 of The Fundamentals of Ethic
  ].

Certainly many immoral people are dee
But others are able to sleep well at
well done (assassination, theft, betr
within a network of like-minded assoc
sometimes get away with it, having a
and never regret the harm they have c

see whether i
 defected or not. Now, for the sake o
assume that the agents will continue
every round will be followed by anoth
assumptions, what is the rational thi

If you know that you will be meeting
rounds, the incentive to defect appea
diminished, for two reasons.

Reason 1: If you defect now, your opp
defecting. Punishment is not possible
dilemma.

Reason 2: If you 'test the water' by
receive the sucker's payoff on the fi
are playing the game indefinitely, th
util) can be 'amortized' over the fut

When taken into the context of an inf
long) run, then the loss of a single
represent a small percentage of the o
if you play the prisoner's dilemma ga
cooperation is a rational outcome.

The summary is that selfish rationali
possible result. Maximizing individua
possible reward. There are more benef
collectively, if they both cooperated
that being good benefits more as a wh
individually.

Interestingly, humans display a syste
cooperative behavior in this, which g
rationales[footnote:
Tversky, Amos; Shafir, Eldar (2004).
similarity: selected writings. (PDF).
Technology Press. Retrieved July 26,
]. It makes sense because the idea of
into play, because the assumption tha
to reward or punish their partner is

  Part II: Conclusion

A rational agent's self-interest is t
rewarding. Rational agents in a group
Nash equilibrium, so being good means
more rewards. In the face of punishme
selfish decisions decreases future ex
other agents are less inclined to coo
defecting agent.

  Conclusion

I conclude that it is in our self-int
analyzed in the above two parts. Part
reaps more long term rewards, and Par
benefits more as a whole.

References

[1]

[2]

[3]

[4]

[5]

Certainly based on pure accomplishmen
same, since both can accomplish their
or unjust means. But the primary diff
chance for the dishonest man to get c
penalized. And sure, certainly the ba
with it. A major assumption here is t
it.

Let's break the argument into two pie
sometimes in lying, and suppose he ne

  He might fail at deceit.

This is the more realistic argument.
underlying assumption is unattainable
entire life from the moment they were
outrageous assumption, we might as we
everything he could ever want in life
goal but taking the unjust method to
more risk, but at a chance to gain hi

But here is the kicker: there is no d
had assumed the two man achieved the
means. Then obviously it makes no sen
All the deceitful man has done is sav
Which may be valuable, but at a heavy

  He never fails at deceit.

Even if we make the (unlikely) assump
capable of maintaining his composure
out of character, it still doesn't ma
riskier path as it leads to the same
save some time and effort. Seems that
difficult to accomplish (if not impos
layman. Sure, there may exist those o
enough to carry this out, but they mo
argument. It is likely the assumption
cannot be proved.

  Results.

A lot of the arguments explored here
solid in logic analysis as from the r
But nevertheless, the fundamental arg
show that it's much easier to maintai
composure than a deceitful and unjust
it difficult to maintain a deceitful
risks. Consequently, if we act just,
self-interest to be good.

  Conclusion

We've now explored this problem from
rational agent would act, and their p
in a rational agent's self-interest t
rewarding, because making bad (not go
decreases future expected rewards, an
the cooperation with other rational a
less reward.

And we've followed up on what do we d
different methods to solving a proble
good and some not. It seems that the
reasonably invalid, and even if it wa
be gained.

The common result from the analysis s
optimal choice.