A Game-Theoretic Solution to the Inconsistency between Thrasymachus and Glaucon in Plato's Republic

Hun Chung
University of Rochester, USA

ABSTRACT. In Book I of Plato's *Republic*, Thrasymachus contends two major claims: (i) justice is the advantage of the stronger, and (ii) justice is the good of the other, while injustice is to one's own profit and advantage. In the beginning of *Book II*, Glaucon self-proclaims that he will be representing Thrasymachus' claims in a better way, and provides a story of how justice has originated from a state-of-nature situation. However, Glaucon's story of the origin of justice has an implication that justice is the advantage of the weak rather than the stronger. This is inconsistent with Thrasymachus' first claim, which states that justice is the advantage of the stronger. This is a problem for Glaucon since he is supposed to be representing Thrasymachus' original claims in a better way. In the present article, I provide two solutions to this puzzle with the help of elementary game theory.

KEYWORDS. Republic, Thrasymachus, Glaucon, Glaucon's Challenge, Game Theory, Prisoner's Dilemma

I. INTRODUCTION

In Book I of Plato's *Republic*, Thrasymachus challenges Socrates by proposing two rather provocative claims:

i. Justice is the advantage of the stronger (383c).[1]
ii. While justice is really the good of another, injustice is to one's own profit and advantage (343c; 344c).

By the end of Book I, Socrates succeeds in forcing Thrasymachus to retract both of these claims and makes Thrasymachus admit that justice is, rather, the advantage of the weak and that injustice is never more profitable than justice.

ETHICAL PERSPECTIVES 23, no. 3(2016): 383-410.

doi: 10.2143/EP.23.3.3170038

In Book II, Glaucon and Adeimantus decide to stand up against Socrates in order to rescue Thrasymachus from helpless defeat. Glaucon and Adeimantus attempt to do so by providing the best possible representation of Thrasymachus' views. Between the two, my primary focus here will be on the view presented by Glaucon.

Despite Glaucon's initial announcement that he will be reviving and representing a better version of Thrasymachus' views, if one follows Glaucon's reinterpretation of Thrasymachus' claims carefully, then one can discover that Glaucon's own reformulations are not entirely consistent with Thrasymachus' original claims. Our purpose here is to show what this inconsistency is, and argue that it is ultimately solvable. I will utilize tools of elementary game theory along the way.

II. The Origin of Justice: Glaucon's Interpretation of Thrasymachus' Second Claim

One of the two major claims that Thrasymachus puts forth in Book I is that while injustice is to one's own profit and advantage, justice is the good of another (343c; 344c). In the beginning of Book II, Glaucon attempts to give a reinterpretation of this claim by first characterizing justice as a certain type of good, and then by giving an account of how justice has acquired such status from its very origins.

At the beginning of Book II, Glaucon characterizes three kinds of good: (i) something that is good for its own sake, (ii) something that is good both for its own sake and for the results that come with it, and (iii) something that is good merely because of the results that come with it (i.e. something that is good purely for instrumental reasons; 357b-d). Glaucon then asks Socrates where among the three kinds of good he regards 'justice' to fit. Socrates replies that he puts justice among 'the finest kind of goods'; namely, something that is both good for its own sake and good for the results that come with it (358a). I will not focus on the issue of whether a type (ii) good (i.e. something that is good both

for its own sake and for its accruing results) is really a superior or finer type of good than a type (i) good (i.e. something that is good simply for its own sake). The important point is that both type (i) and type (ii) goods (i.e. things that are good for its own sake [whether or not they are also good for their accruing results]) seem to be superior to type (iii) goods (i.e. things that are merely instrumentally good), and according to Glaucon, popular opinion conceives 'justice' to be merely a type (iii) good. According to Glaucon, people in general

> […] say that justice belongs to the onerous kind, and is to be practiced for the sake of the rewards and popularity that come from a reputation for justice, but is to be avoided because of itself as something burdensome (358a).

In other words, according to Glaucon, ordinary people think justice as something similar to taking bad medicine[2]; doing it is unpleasant, but you do it only because of the good results that it is purported to bring, and you will never do it for its own sake. In short, the general opinion is that justice is good for *purely instrumental reasons*.

According to Glaucon, there is a reason why people generally think of justice in this purely instrumental way, and this is because of the specific way that justice initially emerged as a social norm from a pre-societal state-of-nature situation. According to Glaucon,

> They say that to do injustice is naturally good and to suffer injustice bad, but that the badness of suffering it so far exceeds the goodness of doing it that those who have done and suffered injustice and tasted both, but who lack the power to do it and avoid suffering it, decide that it is profitable to come to an agreement with each other neither to do injustice nor to suffer it. As a result, they begin to make laws and covenants, and what the law commands they call lawful and just. This, they say, is the origin and essence of justice. It is intermediate between the best and the worst. The best is to do injustice without paying the penalty; the worst is to suffer it without being able to take revenge. Justice is a mean between these two extremes (358e-359a).

We can see here that Glaucon is presenting a state-of-nature story that resembles that of Thomas Hobbes.[3] I believe that it would be instructive to represent the situation that Glaucon explains with a simple formal model.

Let us consider an interaction between two individuals, 1 and 2, who encounter each other in Glaucon's state of nature. Suppose that if the two individuals treat each other with justice, each person receives a default payoff of 0. In the above passage, Glaucon explains that "[…] to do injustice is *naturally good* and to suffer injustice *bad*" (italics mine). So, let $x > 0$ denote the positive benefit of committing injustice, and let $-z < 0$ be the cost of suffering injustice from another person, where $z > 0$. In the passage, Glaucon explains that "[…] the badness of suffering [injustice] far exceeds the goodness of doing it." Hence, we must have $z > x > 0$.

The assumption that doing injustice is naturally good, which is reflected in the positive value of x, is in accordance with Thrasymachus' claim that "[…] injustice is to one's own profit and advantage" (343c; 344c). The fact that the value of $-z$ is negative implies that suffering injustice is *naturally bad* as Glaucon assumes. Note that we have: $-z < 0 < x$. This is in accordance with Glaucon's contention that justice

> […] is intermediate between the best and the worst. The best is to do injustice without paying the penalty; the worst is to suffer it without being able to take revenge. Justice is a mean between these two extremes (359a).

What would happen when both people simultaneously tried to inflict injustice on the other person? Let $0 \leq p \leq 1$ be the probability that person 1 is able to successfully inflict injustice on person 2. Then, $1 - p$ would the probability that person 1 suffers injustice from person 2. If so, the expected payoff of person 1 when both people decide to inflict injustice is:

$$px - (1 - p)z$$

The situation of person 2 will be the mirror opposite; that is, for person 2, the probability of successfully inflicting justice will be $1 - p$, and the probability of suffering injustice will be p. Hence, the expected payoff of person 2 when both people decide to inflict injustice is:

$$(1 - p)x - pz$$

According to Glaucon, "[…] those who have done and suffered injustice and tasted both, but who lack the power to do it and avoid suffering it, decide that it is profitable to come to an agreement with each other neither to do injustice nor to suffer it." In other words, what Glaucon is basically saying is that whenever it is the case that both $px - (1 - p)z$ and $(1 - p)x - pz$ are below 0, both person 1 and person 2 would profit from a mutual agreement of justice.

The situation can be summarized by the following game matrix:

[Person 2]

[Person 1]	Justice	Injustice
Justice	0,0	$-z, x$
Injustice	$x, -z$	$px - (1 - p)z, (1 - p)x - pz$

Figure 1: Glaucon's State of Nature

Each row represents the available actions to person 1 and each column represents the available actions to person 2. For each combination of actions, each person receives a payoff; the payoff on the left-hand side of the comma represents person 1's payoff, while the payoff on the right-hand side of the comma represents person 2's payoff.

Assuming that $px - (1 - p)z < 0$ holds for person 1 and $(1 - p)x - pz < 0$ holds for person 2, we are now able to order the four possible situations that an individual may experience in Glaucon's state of nature from best to worst:

i. Best: commit injustice while the other person acts justly (Payoff: x)
ii. Good: mutual justice (Payoff: 0)
iii. Bad: mutual injustice (Payoff: $px - (1 - p)z$ for person 1; $(1 - p)x - pz$ for person 2)
iv. Worst: suffer injustice from the other person while one acts justly (Payoff: $-z$)

One might recognize that such preference ordering has the same structure as the well-known game of Prisoner's Dilemma (PD game). The PD game can be represented by the following game-matrix:

[Player 2]

[Player 1]	Cooperate	Defect
Cooperate	Good, Good	Worst, Best
Defect	Best, Worst	Bad, Bad

Figure 2: The Prisoner's Dilemma

In a PD game, defecting is a *strictly dominant* strategy[4] for both players. Regardless of what the other player does, each player can obtain an outcome that is higher-ranked in his or her preference ordering by defecting. Hence, in a PD game, both players will necessarily defect, and as a result, mutual defection is the *unique* Nash equilibrium of the game.[5] What makes the Prisoner's Dilemma a *dilemma* is the fact that its unique Nash equilibrium is *sub-optimal*, i.e. both players could have improved their situation by mutually cooperating instead of defecting. The PD game shows that, in certain situations, allowing each person to act optimally from his or her own individual standpoint does not necessarily lead to a socially optimal state.[6]

Based on figure 2, in Glaucon's state of nature, doing injustice is a strictly dominant strategy for both people. Hence, in Glaucon's state of nature, both people will necessarily defect, and therefore, mutual injustice becomes the *unique* Nash equilibrium of the game.

Furthermore, given $px - (1 - p)z < 0$ for player 1 and $(1 - p)x - pz < 0$ for player 2, this unique Nash equilibrium is likewise *sub-optimal*; both people can improve their situations by mutually refraining from inflicting injustice on the other person. Therefore, in such a situation, both people can profitably "[…] come to an agreement with each other neither to do injustice nor to suffer it, [and] a result, they begin to make laws and covenants" the requirements of which, according to Glaucon, are none other than the requirements of justice.

We can see that this establishes Glaucon's claim that justice is merely a type (iii) good – something that is unpleasant, but good merely because of the good results that come with it. And the main reason why Glaucon regards justice as merely a type (iii) good relates to how justice originally came about from a pre-societal state-of-nature situation, which has the same structure as that of a PD game.

Nicholas Denyer (1983) has called the type of social theory that sees justice as a solution to a Prisoner's Dilemma as 'Prisoner's Dilemma theories of justice', and has explained that its tradition can be found as stretching from the works of the Sophists to the works of contemporary social contract theorists like David Gauthier (1986). We can see that in order to explain why justice is merely a type (iii) good, Glaucon himself is attempting to present a Prisoner's Dilemma theory of justice.

However, the crucial assumption that is required to make mutual injustice not only strictly dominant, but also sub-optimal, and thereby render Glaucon's state of nature a PD game is for *both* $px - (1 - p)z < 0$ to hold for person 1 and $(1 - p)x - pz < 0$ to hold for person 2 respectively. Without this assumption, it will no longer be 'profitable' for both people "[…] to come to an agreement with each other neither to do injustice nor to suffer it" as Glaucon claims. Surely, if $px - (1 - p)z \geq 0$, then person 1 will lack any incentive to subject him or herself to a mutually agreed law that would coerce him or her to refrain from inflicting injustice on other people; i.e. justice will no longer be a type (iii) good for person 1. Similarly, if $(1 - p)x - pz \geq 0$, then person 2 will

lack such incentive; i.e. justice will no longer be a type (iii) good for person 2.

If either $px - (1 - p)z \geq 0$ or $(1 - p)x - pz \geq 0$, then mutual justice no longer *Pareto dominates* mutual injustice, and hence, Glaucon's state of nature no longer becomes an instance of a PD game. The following two figures illustrate the situation when $px - (1 - p)z \geq 0$ and when $(1 - p)x - pz \geq 0$:

[Person 2]

[Person 1]	Justice	Injustice
Justice	Bad, Good	Worst, Best
Injustice	Best, Worst	Good, Bad

Figure 3: Glaucon's State of Nature when $px - (1 - p)z \geq 0$

[Person 2]

[Person 1]	Justice	Injustice
Justice	Good, Bad	Worst, Best
Injustice	Best, Worst	Bad, Good

Figure 4: Glaucon's State of Nature when $(1 - p)x - pz \geq 0$

We can see that neither situation is a PD game anymore. In figure 3, person 1 will not consider it profitable to live under mutual justice. In figure 4, person 2 will not consider it profitable to live under justice. That is, if either $px - (1 - p)z \geq 0$ or $(1 - p)x - pz \geq 0$, then one person is now going to order the four possible situations as,

i. Best: Commit injustice while the other person acts justly (payoff: x)
ii. Good: Mutual injustice (Payoff: $px - (1 - p)z \geq 0$ or $(1 - p)x - pz \geq 0$)

iii. Bad: Mutual justice (Payoff: 0)
iv. Worst: Suffer injustice from the other person while one acts justly (payoff: $-z$)

and such a person will no longer see mutual justice as profitable as injustice is now a strictly dominant strategy for that person.

So, in order for both individuals to have an incentive to voluntarily subject themselves to justice, it is necessary for p and $(1-p)$ to be simultaneously low enough so that both $px-(1-p)z<0$ and $(1-p)x-pz<0$ hold.

We can immediately see a tension here. As p decreases, $(1-p)$ increases, and as $(1-p)$ decreases, p increases. It is thus not easy to make both values sufficiently low to render mutual justice mutually profitable for both individuals. The condition that is required for both $px-(1-p)z$ and $(1-p)x-pz$ to be below zero is:

$$\frac{x}{x+z}<p<\frac{z}{x+z}$$

Let $\underline{p}=\frac{x}{x+z}$ and $\overline{p}=\frac{z}{x+z}$. Then $\underline{p}$ will be the *lower bound* of p that would make justice profitable, and thereby a type (iii) good for both people, while p will be the upper bound of p that would make justice profitable, and thereby a type (iii) good for both people.

Remember that p is the probability that person 1 is able to successfully inflict injustice on person 2, while $(1-p)$ is the probability that person 2 is able to successfully inflict injustice on person 1. Hence, p can be interpreted as representing person 1's *overall strength*, while $(1-p)$ can be interpreted as representing person 2's *overall strength*. The assumption $\underline{p}<p<\overline{p}$ is essentially saying neither person can be particularly stronger than the other; i.e. the assumption implies that both people must be *sufficiently weak*. This means that if we follow Glaucon and interpret justice as a type (iii) good, justice can be profitable, and thereby good – if good at all – only for the naturally weak.

This is what Glaucon was essentially saying when he explained that

> People value [justice] not as a good but because they are *too weak* to do injustice with impunity. Someone who has the power to do this, however, and is a true man wouldn't make an agreement with anyone not to do injustice in order not to suffer it. For him that would be madness (359a-b; italics mine).

The fact that Glaucon thinks that justice is good only for the sufficiently weak is further supported by his thought experiment concerning what would happen when both the just and unjust person wore *Gyges' ring*:

> We can see most clearly that those who practice justice do it unwillingly and *because they lack the power* to do injustice, if in our thoughts we grant to a just and an unjust person the freedom to do whatever they like. […] The freedom I mentioned would be most easily realized if both people had the power they say the ancestor of Gyges of Lydia possessed. […] He took the ring […] became invisible to those sitting near him […] Let's suppose, then, that there were two such rings, one worn by a just and the other by an unjust person. Now, no one, it seems, would be so incorruptible that he would stay on the path of justice or stay away from other people's property, when he could take whatever he wanted from the marketplace with impunity, go into people's houses and have sex with anyone he wished, kill or release from prison anyone he wished, and to all the other things that would make him like a god among humans. (359c-360c; italics mine)

Glaucon explains that no just person will remain just if given the opportunity to wear Gyges' ring, which would allow him or her to commit injustice with absolute impunity. In our formal model, wearing the Gyges' ring will, in effect, determine the value of p in a particular way, depending on who wears it; if person 1 wears the ring, p will become 1, while if person 2 wears the ring p will become 0. In either case, p will fail to be in the range between the lower bound $\underline{p}$ and the upper bound $\overline{p}$, which is necessary for justice to be profitable, and thereby a type (iii) good for both individuals.

So, if we accept Glaucon's account of justice, we now arrive at the conclusion that, according to Glaucon, justice is *exclusively the advantage of the sufficiently weak*, and not the stronger. As a matter of fact, the thought that justice can be an advantage only to those who are sufficiently weak and thereby relative equals to one another is a permeating theme throughout the history of the social contract tradition, which tries to see justice as a mutually advantageous escape from a Prisoner's Dilemma type situation. As Denyer explains,

> […] the Prisoner's Dilemma theory of justice is based on an *egalitarian* conception of human nature. It assumes that all human being are similarly vulnerable, that all human beings are similarly capable of inflicting harm, that all human being are similarly benefitted to some extent by being unjust, and that all human beings are similarly benefitted to an even greater extent by having no injustice done them (1983, 137; italics original).

Similarly, Thrasher points out that "[…] both Epicurus and Hume suggest, there can be no justice between gods and mortals or humans and animals. Justice is a relationship between relative equals" (2012, 429). It is also well known that equality among individuals was one of the key assumptions of Hobbes's social contract theory.[7] Despite the popularity of such thought among social contract theorists, Glaucon now faces a problem.

III. The Problem – Justice is the Advantage of the Weak

Here is where we are so far. According to Thrasymachus:

i. Justice is the advantage of the stronger (383c).
ii. While justice is really the good of another, injustice is to one's own profit and advantage (343c; 344c).

After Thrasymachus gets defeated by Socrates' cunning argument in Book I, Glaucon proclaims he himself will revive and represent Thrasymachus' claims in a much more forceful way.

Glaucon starts with Thrasymachus' second claim – namely that justice is really the good of another, while injustice is to one's own profit and advantage (343c; 344c). If we accept this, then the question is what good, if any, can justice do for those who abide by its requirements. Glaucon argues that justice, if it is in any sense good, is at most a type (iii) good – things that are unpleasant but good only for their good results and never good in themselves.

Glaucon provides a state-of-nature explanation that shows how justice has acquired its status as a type (iii) good from its very origins. According to Glaucon, people living in the state of nature will find it *profitable* to live under the rule of law (the requirements of which Glaucon identifies with justice) as, in the state of nature, people will find that the cost of suffering injustice from others far outweighs the benefit of committing injustice themselves. This, however, would be the case only for those who are naturally weak.

In other words, a mutual arrangement of justice is profitable only for those who are sufficiently weak; the naturally strong will not find the enforcement of justice by law profitable. As Goldsmith similarly points out, "[…] the few who believe they will get away with more than they suffer will be unwilling to contract" (1995, 358). Hence, for the naturally strong, justice is not even a type (iii) good, which means that justice is never the advantage of the stronger. This is apparently inconsistent with Thrasymachus' first claim – namely that justice is the advantage of the stronger.

This is a problem for Glaucon as he maintains he is representing Thrasymachus' original views only in a better way. As Weiss makes it clear, "[…] the fact is that Glaucon's analysis of justice *departs markedly from* Thrasymachus's" (2007, 99; italics mine). Hence, one might think here that Glaucon's repackaging of Thrasymachus' view on justice is actually closer to the view that was presented by Callicles in *Gorgias* rather than Thrasymachus' own views.[8] I will try, nevertheless, to find ways to make Glaucon's reinterpretation of Thrasymachus' second claim *consistent* with what Thrasymachus had said in his first claim.

IV. Solution 1: An Establishment of a Democracy

As we have seen, according to Glaucon justice can only be, at best, (instrumentally) good for the weak, and never good for the stronger. This is apparently inconsistent with Thrasymachus' claim that justice is the advantage of the stronger that was presented in Book I. How can this apparent inconsistency be resolved?

First, we have to be careful concerning what Thrasymachus intended by the term 'the stronger'. When Thrasymachus claimed that justice is the advantage of the stronger in 338c, what he was trying to say was that justice is the advantage of *the rulers*. As Weiss correctly points out, "For Thrasymachus 'stronger' is a political term: the stronger are those who are actually in power – whether the regime be a tyranny, an aristocracy, or a democracy" (2007, 94). So, when Thrasymachus claims that justice is the advantage of the stronger, 'the stronger' here refers specifically to *the ruling class* regardless of the specific nature of the regime. And we can see that there is no reason for us to think that the ruling class of a certain type of society always has to coincide with the naturally strong.

In a democracy, for example, the ruling class would be ordinary people who regard themselves as free and equal democratic citizens. In addition, not all democratic citizens will possess a superior level of overall strength; in other words, many people who rule in a democratic society may not be regarded as those who could inflict injustice with impunity in the state of nature. In fact, there would be many democratic citizens who would be closer to the naturally weak than to the naturally strong. However, since they are the people who *rule* a democratic society, they are, by definition, 'the stronger' according to Thrasymachus.

This shows that people who are naturally weak in the state of nature can still be 'the stronger' after society has been established; this is precisely the case for a democratic society. And given that this can be the case, we have now found a way to make Glaucon's account of the origin

of justice consistent with Thrasymachus' first claim, which says that justice is the advantage of the stronger.

The account runs as follows: in the state of nature, the naturally weak realize that their expected payoff of the state of universal injustice is negative and that they could increase their expected payoff by entering into a society where everybody is enforced to refrain from inflicting injustice on others. The naturally weak thereby establish a society in which justice is enforced. And since nobody among the naturally weak is particularly superior to anyone else, any political structure that gives unequal political power to any particular person would not be acquiesced; consequently, the naturally weak will let everybody take part in the ruling process, and hence: the establishment of a democracy.

Since the establishment of such a democracy was mainly designed to satisfy the naturally weak's need for protection from injustice – a need that the naturally strong do not have – the establishment of such a democracy, and thereby the establishment of justice, is primarily the advantage of the naturally weak. However, we can now say justice is also the advantage of 'the stronger', since in a democratic society the naturally weak, who are the rulers, are now by definition 'the stronger'.

While this solution gets rid of the inconsistency, it nevertheless has a major problem: how can the naturally weak induce the naturally strong to enter into a democratic society in the first place? The expected payoff of the naturally strong interacting with the naturally weak in the state of nature would be positive. By entering into a democratic society in which justice is universally enforced, the expected payoff of the naturally strong would become zero. As a result, there would be no reason for the naturally strong to cooperate with the naturally weak in establishing a democratic society in which everybody shares the same power of political ruling as everybody else.

Obviously, the naturally weak are not able to *force* the naturally strong to enter into a democratic society; if that were possible, then the naturally weak would not be 'the naturally weak'. In order to induce the naturally

strong to enter into a democratic society, therefore, we would have to find a way to make them enter *voluntarily*. However, as we have seen, the naturally strong have no incentive to enter into a democratic society voluntarily, given that their expected payoff in the state of nature is positive. How then can we make this solution actually happen in the state of nature?

I propose the following solution: in the initial stage, the naturally weak will be unable to induce the naturally strong to enter into a democratic society; the naturally strong would prefer the state of nature in which they can exploit the naturally weak without any legal constraints. At this point, not only would the naturally weak have incentives to form a democratic society and refrain from inflicting injustice on one another, but they would also have incentives to unite and defend themselves from the attack inflicted by the naturally stronger. Before inducing the cooperation of the naturally stronger, therefore, the naturally weak must first be able to form a democratic society for themselves.

After a democratic society of the naturally weak has been successfully established, the state of nature will now consist of two types of entities: (1) individuals who are the naturally strong and (2) a whole democratic society united by the naturally weak. Now, it is not the weak individuals that the naturally stronger can prey on; in order to inflict injustice, the naturally strong would have to fight against other naturally strong individuals or they would have to fight against an entire democratic society united by the naturally weak.

This would significantly change the expected payoffs of the naturally stronger in the state of nature. When confronting another naturally strong individual, a typical naturally strong person will be able to inflict injustice on the other half of the time and will suffer injustice from the other half of the time; therefore, his or her expected payoff will be $\frac{x - z}{2} < 0$.

What would then happen if the naturally strong confronted the whole democratic society united by the naturally weak? It is evident that a typical naturally strong individual could quite easily take out a typical naturally weak individual *individually*. However, I believe that it would be

quite difficult for even the strongest human being to take out four or five weak individuals who are fighting in coalition (each perhaps equipped with a deadly weapon). And since Socrates claims that the essential minimum for there to be a city is four or five men (369e), this means that when a typical naturally strong individual confronts an entire democratic society united by the naturally weak, he or she would be confronting a minimum of four or five naturally weak individuals fighting together, and they would mostly likely be confronting many more than four or five.[9] So, whenever a typical naturally strong individual confronts a whole democratic society united by the whole population of the naturally weak, he or she would always lose and get the payoff: $-z$.

After the naturally weak have successfully established a democratic society, the expected payoff of the naturally stronger in the state of nature would be the probability weighted average of $e \cdot (x - z)/2 - (1 - e) \cdot z$ where 'e' denotes the probability of the naturally stronger meeting another naturally strong individual in the state of nature.

Since $\frac{x - z}{2} < 0$ and $-z < 0$, we have $e \cdot \frac{x - z}{2} - (1 - e) \cdot z < 0$; i.e. the expected payoff of staying the state of nature for the stronger is negative. And since the expected payoff of entering into a democratic society is 0 while the expected payoff of remaining in the state of nature is below zero, the naturally stronger will now have incentive to cooperate with the naturally weak and enter into a democratic society. The naturally stronger will now enter into an already established democracy voluntarily, and this completes our first solution to the problem.

Let us verify this solution via a formal model. Suppose $p < \underline{p} = \frac{x}{x + z}$ – i.e. person 2 is *naturally stronger* relative to person 1. Person 1 (i.e. The Weak) makes the first move and decides whether to form a democratic society with other naturally weak people or to stay in the state of nature by him or herself. If person 1 stays in the state of nature, then person 1 (i.e. The Weak) and person 2 (i.e. The Strong) play the state of nature game depicted in figure 2. If person 1 (i.e. The Weak) decides to form a

democratic society with other weak people, then person 2 (i.e. The Strong) decides whether to enter into the democratic society or to stay in the state of nature. The payoffs generated by each person are what I have explained above.[10] The following represents the extensive form of the model:

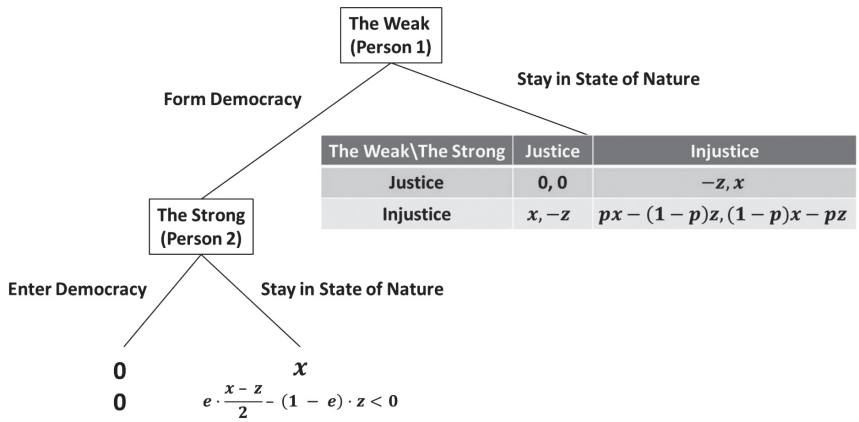


Figure 5: The Game of Democracy *versus* State of Nature

The following proposition formally verifies that the solution I have informally described in this section can be supported as equilibrium of the formal model depicted in figure 5.

Proposition 1

Suppose $p < \underline{p} = \frac{x}{x+z}$. then the following set of strategies:

- The Weak's Strategy = $\begin{cases} \text{Form Democracy} \\ \text{play Injustice in State of Nature subgame} \end{cases}$
- The Strong's Strategy = $\begin{cases} \text{Enter Democracy} \\ \text{play Injustice in State of Nature subgame} \end{cases}$

is *the unique* Subgame Perfect Equilibrium (SPE)[11] of the game.[12]

V. Solution 2: Establishment of a Dictatorship

We have just seen a solution that tries to eliminate the inconsistency between Glaucon's story of the origin of justice and Thrasymachus' first claim – that justice is the advantage of the stronger – by finding a way for the naturally weak to become *the rulers* of a democratic society and thereby making it possible for the establishment of justice to be both the advantage of the naturally weak as well as, by definition, 'the stronger'. The next solution attempts to resolve the inconsistency by finding a way to induce the naturally stronger to form a society in which they occupy the ruling position themselves.

Based on the analysis of our model of Galucon's state of nature (depicted in figure 2), we have seen that the necessary condition for both individuals to find the enforcement of justice profitable is for p to satisfy:

$$\underline{p} = \frac{x}{x+z} < p < \frac{z}{x+z} = \overline{p}$$

Whenever p meets this condition, we can say that both person 1 and person 2 are sufficiently weak to find the enforcement of justice mutually profitable. Consider a situation in which $p > \overline{p}$; i.e. person 1 is *naturally strong* relative to person 2. In such a situation, it will not be profitable for person 1 to live in a legal society in which justice is universally enforced – *unless* there is some *additional benefit* that living under such a legal society would bring besides the protection from injustice it is expected to provide. Can there be such an additional benefit, which even the naturally strong would find attractive? The answer seems to be 'yes'.

In Book II, Socrates describes how a city is initially formulated and then expands throughout time. According to Socrates, "[…] a city comes to be because none of us is self-sufficient, but we all need many things" (369b). In other words, people decide to form a society and cooperate with each other mainly because they judge that doing so will more efficiently satisfy the many needs that they would otherwise not have satisfied.

Note that the need for basic protection is not the only need that people in general have; not only do people have basic needs for food, shelter, and clothes, but they also have many non-basic or luxurious needs, which may not be essential for their basic survival, but are, nonetheless, very important to live a quality life.

Although people may satisfy their most basic needs self-sufficiently in the state of nature, they would be able to satisfy their basic needs much more efficiently if they start to cooperate and live together in a legal society; furthermore, most of the non-basic or luxurious needs that people have would only be satisfied by living in society as these goods or commodities can be produced and supplied only when there is a sufficient amount of security.

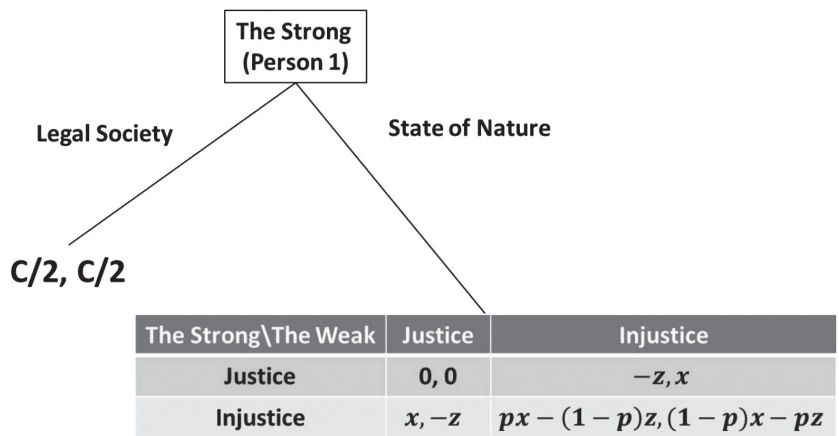
The efficiency of society, according to Socrates, comes from the fact that, after it has been established, everybody can concentrate on what they have *the most aptitude for*; i.e. after society has been established, everybody can concentrate on their own *comparative advantage*. Farmers will be able to concentrate only on producing food; builders will be able to concentrate only on building houses; shoemakers will be able to concentrate only on making shoes, and so on.

> The result, then, is that more plentiful and better-quality goods are more easily produced if each person does one thing for which he is naturally suited, does it at the right time, and is released from having to do any of the others (370c).

By everybody concentrating on producing what they have the most aptitude for, everybody will be able to produce better quality products in much greater quantities than they would have produced in the state of nature; and by everybody exchanging their surplus goods for other types of goods they need that other people have produced, everybody will be able to enjoy a much greater quality and quantity of overall goods than they would have enjoyed in the state of nature, where they relied solely on self-production.

By the reiteration of this process of exchange, the overall prosperity of the society, in principle, will be able to increase indefinitely. This means that the expected payoff of entering into society would be *positive* and *forever increasing*, *not merely zero*, which is the expected payoff of merely mutually refraining from inflicting injustice on one another. And *this,* not protection, would give incentives for the naturally strong to enter into society.

To see this, let $C > 0$ denote the cooperative surplus[13] that a legal society can produce by the protection and specialization of its individual members. We are assuming $\overline{p} < p \leq 1$; i.e. we are assuming that person 1 is naturally strong and person 2 naturally weak relative to each other. Therefore, when the naturally strong (i.e. person 1) contemplate whether or not they should cooperate and enter into society together with the naturally weak (i.e. person 2), the game they will be playing will be the following:



The Strong\The Weak	Justice	Injustice
Justice	0, 0	$-z, x$
Injustice	$x, -z$	$px-(1-p)z, (1-p)x-pz$

Figure 6: Legal Society *versus* State of Nature Game 1

Let me briefly explain the timing of the formal model. In the initial node, The Strong (i.e. Person 1) decides whether to enter a legal society with The Weak (i.e. Person 2) or to remain in the state of nature. If The Strong enters legal society, The Strong and The Weak each receive half of the total cooperative surplus C. If The Strong decides to stay in the state of

nature, then The Strong and The Weak play the state of nature game we have seen previously. The following result summarizes what would happen in such situation.

Proposition 2

A threshold $\underline{C}$ exists such that whenever $C \geq \underline{C}$, it would be profitable for The Strong to enter Legal Society.[14]

Corollary

Suppose $C > \underline{C} = 2[px - (1-p)]$, then the following set of strategies:

- The Strong's Strategy = $\begin{cases} \text{enter Legal Society} \\ \text{play Injustice in State of Nature subgame} \end{cases}$
- The Weak's Strategy = play Injustice in State of Nature subgame

is *the unique* Subgame Perfect Equilibrium (SPE) of the game.[15]

Proposition 2 and Corollary show that as long as the cooperative surplus C produced from a legal society is sufficiently large, entering into a legal society that enforces universal justice is profitable even for the naturally strong. Hence, whenever the cooperative surplus C of a legal society is sufficiently large, justice will be a type (iii) good and thereby profitable for everybody.

This explains how a legal society that enforces universal justice can be established between the naturally strong and the naturally weak from the state of nature. However, this does not entirely resolve the inconsistency between Galucon's story of the origin of justice and Thrasymachus' first claim. This is because, although it is true that justice is the advantage to both the naturally strong as well as the naturally weak according to our current model, it is still much more profitable to the naturally weak than it is to the naturally strong. To see this, define the 'benefit' of living in a legal society as:

Benefit of living in legal society
= Payoff in a legal society – Payoff in the state of nature.

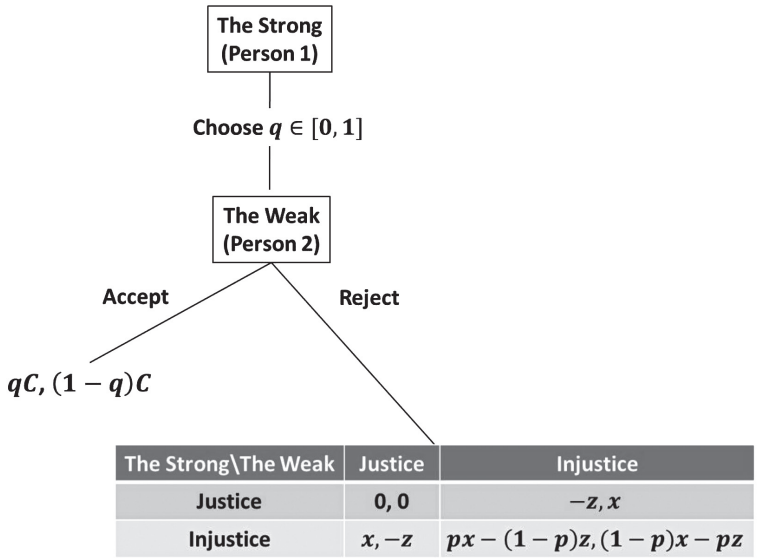
From this formula, the benefit of living in a legal society for The Strong is: $\frac{c}{2} - [px - (1-p)z]$, call this (a); and the benefit of living in a legal society for The Weak is: $\frac{c}{2} - [(1-p)x - pz]$, call this (b). Since $px - (1-p)z > (1-p)x - pz$, we have $(b) > (a)$, which implies that the benefit of living in a legal society is greater for the naturally weak than the naturally strong. Hence, even though the naturally strong will find it profitable to live in a legal society, the naturally weak will find it *more* profitable to live in such a society than the naturally strong. Furthermore, in our current model, in which the cooperative surplus C is divided equally between The Strong and The Weak, it is not obvious who we should consider as being 'the ruler' of the established legal society.

All of this suggests that our current solution to our problem is rather incomplete. We want justice to be a type (iii) good, which is profitable for everybody for instrumental reasons as Glaucon suggests, yet we also want justice to be the advantage primarily for the stronger as Thrasymachus claims. This is not unquestionably true in our current formal model in which living in a legal society and thereby subjecting oneself to universal justice is more profitable for the naturally weak, and when it is not unambiguously determined who the ruling class of such a legal society, once established, would turn out to be.

What we would want is to show that living in a legal system of justice would be much more profitable to the naturally strong than it is to the naturally weak, and we would further want to show that the naturally strong will assume the ruling positions in such legal society once established.

So, let us modify our current formal model as follows. At the initial stage, The Strong proposes $q \in [0,1]$ – the proportion of the cooperative surplus C, that they will appropriate once both The Strong and The Weak enter into a legal society. The Weak can either accept the offer or reject it. If The Weak accept, then they divide the cooperative surplus C accordingly,

which gives The Strong a share of q C and The Weak a share $(1 - q)$ C of the total cooperative surplus C. If The Weak refuse the offer, then both The Strong and The Weak remain in the state of nature playing the state of nature game. The situation can be represented by the following model:



The Strong\The Weak	Justice	Injustice
Justice	0, 0	$-z, x$
Injustice	$x, -z$	$px - (1-p)z, (1-p)x - pz$

Figure 7: Legal Society *versus* State of Nature Game 2

What result will follow from our modified formal model?

Proposition 3

Suppose C > $px - (1 - p)z$ – i.e. suppose that the cooperative surplus of a legal society, C, is greater than what The Strong could solely achieve in the state of nature. Then, The Strong appropriating the entire cooperative surplus C, and The Weak accepting such a legal arrangement will be *the only equilibrium outcome*.[16]

First, note that the assumption C > $px - (1 - p)z$ is plausible; we would expect that once a legal society is established and everyone concentrates

on developing their specialized skill for which they have aptitude, society will prosper and the cooperative surplus thus generated will be at least more than one would expect in a pre-societal state of nature. This is exactly what Socrates anticipates when he explains that, after the establishment of a legal society, "The result, then, is that more plentiful and better-quality goods are more easily produced if each person does one thing for which he is naturally suited, does it at the right time, and is released from having to do any of the others" (370c).

Given such an assumption, what Proposition 2 basically shows is that the naturally weak will agree to live under a legal system in which the naturally strong will appropriate *the entire* cooperative surplus C that has been jointly produced. The society that will be established through such a process will be none other than a *dictatorship* ruled by The Strong. The only service that The Strong offer to The Weak after privately appropriating the entire social wealth is: protection.

The terms of the arrangement are obviously unfair. However, living in such dictatorship is still profitable from the perspective of The Weak as receiving mere protection is still better than remaining in the state of nature, which gives the naturally weak a negative payoff. In other words, The Weak have no choice but to agree to such an unfair arrangement as, by being naturally weak, they lack sufficient bargaining power to make any other social arrangement to which the naturally strong would voluntarily agree. Furthermore, as long as $C > x + z$, becoming the ruler in such a dictatorial society will be *more profitable* to The Strong (for all values of $p \in (\overline{p}\ , 1]$) than it would be for The Weak.

So, as Crombie[17] and Irwin[18] have rightly pointed out, "[…] a system of restrictions is required for social life. Without such a system everyone would be worse off. Glaucon's contract makes the point that most people would be better off with such a system" (Goldsmith 1995, 361). However, neither of these authors properly realized that the only social system that could naturally emerge from the state of nature by a voluntary agreement when there is significant discrepancy of power is *a dictatorship*.

Hence, we have arrived at a solution that removes the apparent inconsistency between Glaucon and Thrasymachus. Justice is a type (iii) good that is profitable to everybody, but it is much more advantageous for the stronger than the weak.

VI. Conclusion

I have tried to resolve the apparent inconsistency between Glaucon's reinterpretation of Thrasymachus' claims and Thrasymachus' original claims. In Book 1, Thrasymachus contends two major claims: (1) justice is the advantage of the stronger, and (2) justice is the good of the other, while injustice is to one's own profit and advantage. In the beginning of Book II, Glaucon states that he will be reviving Thrasymachus' claims and presenting them in a better way. Glaucon provides an account of how justice originated from a state of nature situation. This account was provided in order to explain why justice, according to Glaucon, is commonly assumed to be only the third type of good; that is, something that is unpleasant but good only for purely instrumental reasons.

The account alludes to Thrasymachus' second claim, namely that justice is the good of the other, while injustice is to one's own profit and advantage. However, we have seen that Glaucon's account of the origin of justice has an implication that justice is the advantage of the weak rather than the stronger. This apparently conflicts with Thrasymachus' first claim; namely, that justice is the advantage of the stronger. This is something of a problem for Glaucon as he is supposed to be representing Thrasymachus' original claims in a better way.

I have tried to resolve this inconsistency by providing two broad solutions. One was to show how a democracy can be established from the state of nature. The major strategy of this solution is to make the naturally weak coincide with 'the stronger' by definition and thereby make the advantage of the former automatically the advantage of the latter.

The other was to show how even the naturally stronger could have incentives to establish a society in which they appropriate the entire cooperative surplus and become dictators. It is no surprise that justice (which in the current context means the legal requirements of one's society) is to the advantage of the stronger (rulers). Taking either solution will make all of the following four claims true and consistent to one another:

> Justice is the advantage of the stronger.
> Justice is profitable for everyone.
> Justice is the third type of good – it is good only for purely instrumental reasons.
> Justice is the good of the other, while injustice is to one's own profit and advantage.

In sum, I conclude that Glaucon's reinterpretation of Thrasymachus' claims was really not inconsistent with those of Thrasymachus.

Works Cited

Chung, Hun. 2015. "Hobbes's State of Nature: A Modern Bayesian Game-Theoretic Analysis." *The Journal of the American Philosophical Association* 1/3: 485-508.

Cooper, John M. (ed.). 1997. *Plato – Complete Works*. Indianapolis, IN: Hackett.

Crombie, Ian M. 2012. *An Examination of Plato's Doctrines*: Volume I *Plato on Man and Society*. London: Routledge.

Denyer, Nicholas. 1983. "The Origins of Justice." In *Syzetesis: studi sull'Epicureismo Greco e Romano offerti a Marcello Gigante*, 133-152. Naples: Gaetano Macchiaroli.

Ferrari, Giovanni R. F. (ed.). 2007. *The Cambridge Companion to Plato's Republic*. Cambridge: Cambridge University Press.

Gauthier, David. 1986. *Morals by Agreement*, Oxford: Oxford University Press.

Goldsmith, Maurice M. 1955. "Glaucon's Challenge." *Australasian Journal of Philosophy* 73/3: 356-367.

Hobbes, Thomas. 1994. *Leviathan (with selected variants from the Latin edition of 1668)*. Edited by Edwin Curley. Indianaplis, IN: Hackett.

Irwin, Terence. 1979. *Plato's Moral Theory: The Early and Middle Dialogues*. Oxford: Clarendon.

Irwin, Terence (trans.). 1979b. *Plato: Gorgias*. Oxford: Clarendon.

Kirwan, Christopher. 1965. "Glaucon's Challenge." *Phronesis* 10/2: 162-173.

Kerferd, George B. 1981. *The Sophistic Movement*. Cambridge: Cambridge University Press.

Strauffer, Devin A. 2000. *Plato's Introduction to the Question of Justice*. Albany, NY: SUNY.

Thrasher, John. 2012. "Reconciling Justice and Pleasure in Epicurean Contractarianism." *Ethical Theory and Moral Practice* 16/2: 423-436.

Weiss, Roslyn. 2007. "Wise Guys and Smart Alecks in Republic 1 and 2." In *The Cambridge Companion to Plato's Republic*. Edited by Giovanni R.F. Ferrari, 90-115. Cambridge: Cambridge University Press.

NOTES

1. All references to Plato's Republic are from Cooper (1997).

2. Goldsmith (1995, 356) makes a similar observation.

3. See Hobbes (1994). For a modern Bayesian game-theoretic analysis of Hobbes's state of nature, see Chung (2015).

4. A strategy A is strictly dominant if it generates a greater payoff than any other strategy B in all possible situations.

5. A Nash equilibrium is a situation in which every player is best-responding to every other players' strategies so that no single player has an incentive to unilaterally deviate to another strategy.

6. This has been used as a criticism against Adam Smith's 'invisible hand' argument – namely, that simply allowing members of society to pursue their self-interest will automatically lead, by the guidance of the invisible hand, to a socially optimal state.

7. "For as to the strength of body, the weakest has strength enough to kill the strongest […] As as to the faculties of the mind […] I find yet a greater equality amongst men than that of strength" (Hobbes 1995, XIII, 1,2).

8. As Weiss explains, "For Callicles, [unlike Thrasymachus] the source of conventional justice is the weak, those who fear losing out to the strong and who protect themselves, at the expense of the strong, by dictating what is just and unjust, noble and shameful" (2007, 94-96). "In light of these difference, one might say that it is Glaucon and not Thrasymachus who is a throwback to the Gorgias's Callicles" (Weiss 2007, 100).

9. This is because it is very likely for there to be much more than four or five naturally weak individuals who would agree to enter into a democratic society in the state of nature.

10. For the payoffs written at the bottom of the two left-hand side branches that ramify after 'Form Democracy', the expressions at the top denote person 1's payoffs, while the expressions at the bottom denote person 2's payoffs.

11. A Subgame Perfect Equilibrium (SPE) is a set of strategies that prescribes optimal actions for every player in every subgame. This implies that a SPE is Nash equilibrium in every subgame.

12. Proof. Consider the node right after The Weak form a democracy in which it is The Strong's turn to play. By entering democracy, The Strong receive a payoff of 0, and by staying in the state of nature The Strong receive a payoff of $e.\frac{x-z}{2} - (1-e).z < 0$. Hence, it would be optimal for The Strong to enter democracy. Consider the state of nature subgame that occurs after The Weak decide to stay in nature. Here, playing Injustice is a strictly dominant strategy for both The Weak and The Strong. Hence, (Injustice, Injustice) is the unique Nash equilibrium of

the state of nature subgame. Consider the initial node in which The Weak decide whether to form a democracy or to stay in the state of nature. By forming a democracy, The Weak receive a payoff of 0. By staying in the state of nature, The Weak receive a payoff of $px - (1-p)z$. Since, $p < \underline{p} = \frac{x}{x+z} < \frac{z}{x+z} = \bar{p}$ by assumption, we have $px - (1-p)z < 0$. Hence, forming a democracy would be optimal for The Weak. Hence, the set of strategies described in Proposition 1 is the unique SPE of the game.

13. This does not merely denote economic benefits; it includes the benefit of protection, the sense of security, as well as any other benefit that one may receive by mutual cooperation.

14. Proof. Consider the state of nature subgame played after The Strong decide to stay in the state of nature. Since $\bar{p} < p < 1$, we have $px - (1-p)z > -z$ and $(1-p)x - pz > -z$ and $x > 0$. Hence, Injustice strictly dominates Justice for both The Strong and The Weak. Therefore, (Injustice, Injustice) is the unique Nash equilibrium of the state of nature subgame. In this Nash equilibrium, The Strong's payoff is: $px - (1-p)z > 0$, and The Weak's payoff is: $(1-p)x - pz < 0$. Now, move up to the initial node at which The Strong decide whether to enter legal society or to remain in the state of nature. Entering legal society gives The Strong a payoff: C/2. Therefore, entering legal society will be optimal for The Strong if: $\frac{C}{2} \geq px - (1-p)z \Rightarrow C \geq 2[px - (1-p)]$. Set $\underline{C} = 2[px - (1-p)]$, then whenever $C \geq \underline{C}$, it would be profitable for The Strong to enter legal society.

15. Proof. From the proof of Proposition 1, we know that (Injustice, Injustice) is the unique Nash equilibrium of the State of Nature subgame. By the proof of Proposition 1, it is optimal for The Strong to enter a Legal Society if $C \geq 2[px - (1-p)] = \underline{C}$, which is satisfied by assumption.

16. Proof. Again, since $p > \bar{p}$, remaining in the state of nature gives The Strong a payoff of $px - (1-p)z > 0$, and, The Weak a payoff of $(1-p)x - pz < 0$. Therefore, The Weak will accept any $q \in [0,1]$ such that $(1-q)C \geq 0 \Rightarrow 1 \geq q$ – that is, The Weak will accept any proportion of the cooperative surplus in the legal society. The best offer from the perspective of The Strong will obviously be $q = 1$. So, The Strong will offer $q = 1$ and The Weak will accept it for sure. In other words, The Strong appropriating the entire cooperative surplus, C, and The Weak accepting such legal arrangement will be the only equilibrium outcome.

17. Crombie (2012, 87).

18. Irwin (1979, 186-187).