[Working draft. Please do not circulate or cite without author's permission]

Glaucon's Dilemma.
The origins of social order.

Josiah Ober

Chapter 2 of *The Greeks and the Rational* (book-in-progress, provisional title)
Draft of 2019.09.20 Word count: 17,200.

Abstract: The long Greek tradition of political thought understood that cooperation among multiple individuals was an imperative for human survival. The tradition (here represented by passages from Plato's *Republic*, *Gorgias*, and *Protagoras*, and from Diodorus of Sicily's universal history) also recognized social cooperation as a problem in need of a solution in light of instrumental rationality and self-interest, strategic behavior, and the option of free riding on the cooperation of others. Ancient "anthropological" theories of the origins of human cooperation proposed solutions to the problem of cooperation by varying the assumed motivations of agents and postulating repeated interactions with communication and learning. The ways that Greek writers conceived the origins of social order as a problem of rational cooperation can be modeled as strategic games: as variants of the non-cooperative Prisoners Dilemma and cooperative Stag Hunt games and as repeated games with incomplete information and updating.

In book 2 of the *Republic* Plato's Glaucon offered a carefully crafted philosophical challenge, in the form of a narrative thought experiment, to Socrates' position that justice is supremely choice-worthy, the top-ranked preference of a truly rational person. Seeking to improve the immoralist argument urged by Thrasymachus in *Republic* book 1 (in order to give Socrates the opportunity to refute the best form of that argument), Glaucon told a tale of Gyges and his ring of invisibility.[1] In chapter 1, I suggested that Glaucon's story illustrated a pure form of rational and self-interested behavior, through revealed preferences when the ordinary constraints of uncertainty, enforceable social conventions, and others' strategic choices were absent.

Glaucon's thought experiment posited self-interest, in the sense of egoistic preference-satisfaction, as the driver of human choice and action.[2] It suggested that "primitive" preferences for possession of material goods, access to sex, and power over others were universally top-ranked. And it predicted that a rational individual (one with orderly preferences and beliefs), free to act without constraint to satisfy his top-ranked preferences, would willingly commit acts that both Socrates and ordinary Greeks regarded as unjust, including theft, seduction, and murder.[3]

Following the lead of Andrew Laird (2001), I suggested that Plato's account of Gyges was freely adapted from Herodotus' earlier story of how Gyges became king (*Histories* 1.9-11). Herodotus' tale of Gyges lacked the magic ring and emphatically foregrounded uncertainty and the constraints placed on the options available to self-interested agents in a context in which other agents were likewise seeking to satisfy their preferences. In other words, Plato's Glaucon reduced a non-parametric .strategic situation with multiple choice-makers to a parametric choice situation with only one. In its abstraction from fallible and constrained human beings (e.g. those in Herodotus' story), to hypothetical unerringly rational and self-interested choice-making agents, Glaucon's philosophical construct resembles contemporary choice theory.

2.1. Social order as a problem

An ancient Greek reader of *Republic* book 2, willing to be convinced that humans are motivated in the ways that emerged from Glaucon's thought experiment, confronted a puzzle: How could many persons, each with a primary goal of maximizing his or her share of a limited pool of material goods, sexual access, and power, ever have come to cooperate in the deep and complex ways necessary for the emergence and persistence of social order? How did self-interested individuals become rule-following residents of a sustainable community? The theme of this chapter is how Plato, and others in the classical Greek tradition of social and political thought, addressed that puzzle. I will argue that they did so in narratives that are illuminated by, and seem to anticipate, some of the standard devices of strategic game theory.

For a polis-dwelling Greek in the age of Plato, as for us today, it was self-evident that people do live in rule-bounded communities of one sort or another. For the Greeks these communities prominently included states in the form of *poleis* and rules in the form of *nomoi*: formal laws, social norms, and established customs. Those states had formalized systems of political authority, such that the question of how, and by whom, rules are interpreted and enforced had been at least provisionally answered. Social order meant that people could expect most other people to obey most of the rules most of the time.

But where did the rules – and the authority to make, interpret, and enforce them – come from? What is the origin of law, conceptions of political legitimacy, and the enforceable duty to obey? How did those in authority come to wield power? In communities ruled by kings or tyrants, power might give persons in authority the opportunity to behave more or less like Gyges with the magic ring – taking other people's possessions at will, having superior access to sex, and dominating subjects (see chapter 3). Alternatively, in a civic community, power was more or less equitably distributed among citizens – so that no one enjoyed a Gyges-type freedom of choice and action (see chapters 4 and 5). But why would a rational and self-interested individual ever choose to put him or herself in a position of having to take orders from someone (or everyone) else? Why would I willingly agree to obey your orders, if and when those orders limit my choices or obstruct the course of action that leads most directly to my most preferred outcome? The ordinary, constraining social conditions that Glaucon's narrative thought experiment was designed to strip away may begin to seem strange, once the thought experiment has been performed. How did social order arise in the first place?

The notion that, for humans (unlike other animals), nature does not determine social order, that *nomos* (law, norm, or custom) is both different from and at least potentially in conflict with *phusis* (human nature), was one of the defining ideas of classical Greek political thought. That distinction and the role it played in Greek culture have been the subjects of much scholarly attention. Yet there is, I think, more to say about the implications of the *nomos/phusis* disjunction for Greek explanations of the origins of human social order that began with a premise of rationality and the motivation of self-interest.[4]

In *Republic* book 2, Plato's Glaucon claims that it is commonly supposed that social order, understood as a voluntary agreement on rules by persons seeking to constrain egregious behavior, along with the positive evaluation of justice, arose as a compromise. Each party to that compromise would have preferred to act unjustly in accordance with egoistic self-interest, but each recognized the costs to himself of others doing so:

> By nature, they [ordinary Greeks] say, to commit injustice is [for each individual] a good and to suffer it is an evil, but that the excess of evil in suffering injustice is greater than the excess of good in doing injustice. So that, when men do injustice and suffer it from one another and have experienced both, for those who lack the power at once to avoid the one and choose (*hairein*) the other, it seems profitable (*lusitelein*) to make a compact with one another (*sunthesthai allêlois*) neither to commit nor to suffer injustice; and [they say] that this is the beginning of the establishment of laws (*nomoi*) and covenants (*sunthêkai*) between men and that they name the command of the law "the lawful" and "the just" and [they say] that this is the genesis and essential nature of justice — in between the best, which is to do wrong with impunity, and the worst, which is to be wronged and be impotent to get one's revenge. (Plato, *Republic* 358e-359a).[5]

In this passage, Glaucon sketches a Greek folk theory ("they say") of the origins of social order.[6] It is a companion-idea and logical consequence of the folk theory of instrumental rationality discussed in chapter 1. The social order folk theory is reframed by Thrasymachus' (340e-341a) and Glaucon's (360e-361a) specification that, when seeking to understand rationality and self-interest, one must identify the choices that would be made by hypothetical perfect craftsmen (*demiourgoi*) of self-interest: human agents who unerringly choose the course of action that most benefits themselves. I suggested (chapter 1) that the hypothetical "unerring craftsmen," a notion that goes forward unchallenged by Plato's Socrates, are relevantly similar to the hypothetical fully rational, fully informed, and cognitively unlimited agents assumed in choice theory.

The passage quoted above is immediately followed by the extended thought experiment that includes the Gyges story. As we saw in chapter 1, Glaucon's thought experiment imaginatively places two persons on the same path. One is reputed to be just and the other unjust; each is freed of social constraints. Glaucon later suggests that their freedom could be gained through possession of Gyges-type rings (360b). If we follow the two men, "We should then catch the just man, self-revealed, going along to the same destination as the unjust man, because of the striving to gain more and more (*pleonexia*) which every creature by its nature pursues as a good" (359b-c).[7]

Glaucon predicts that, under conditions of pristine freedom (as it is elucidated in the Gyges story) each person will arrive at the same destination, which is later (360b-c) glossed as acting at others' expense so as to maximize his own access to the goods of material possessions, sex, and rule.[8] By having his "just" and unjust man travel the same path and arrive at the same place, Glaucon has, at least implicitly, placed his two hypothetical craftsmen of self-interest in a situation of strategic interaction and choice. The issue of the origins of social cooperation that he had raised, in the passage quoted above, just before laying out his thought experiment might, therefore, be addressed, in a radically simplified form, by imagining what course of action each of those hypothetical individuals would choose in consideration of the choice likely to be made by the other. In brief, we can transform the parametric Gyges choice situation into a strategic game, by taking into account the two choice makers (the "just" and unjust men) discussed by Glaucon in the extension of his thought experiment.[9]

2.2 Strategic games

We are, I suggest, entitled to think of Glaucon's two self-interested, instrumentally rational individuals as the players in a two-party strategic game: that is, a game in which each player (hypothetical fully rational agents, per above) chooses her move in light of the move she believes will be made by the other. The game may be illustrated in "normal form" as a two-by-two matrix, as in Figure 2.1.[10] Each player (1,2) has two possible choices (a,b), meaning that there are four possible choice pairs (1:a,2:a; 1:a, 2:b; 1:b, 2:a; 1:b, 2:b). Each of the four boxes in the matrix represents a choice pair as a possible outcome and lists a payoff for each player. The payoff for Player 1 (row player) is in the lower left corner of each box; the payoff for Player 2 (column player) is in the upper right corner. Payoffs are descriptively listed in the order: row player, column player. Each player chooses the box with the highest available payoff to herself, expecting the other player to do likewise. In sum, players choose based on their ordered preferences (highest payoff = highest ranked preference) and what they believe the other player will do.

[Figure 2.1: Basic setup, about here]

Players are assumed to be fully informed about the choice situation, that is, they know the payoffs in each box and that each rationally seeks his best payoff. They choose simultaneously; that is, they lack advance knowledge of the other player's move. They may communicate with each other, but a commitment (a promise to make a certain move) made by one player is credible to the other only if it leads to the promise-maker's best available payoff. After each player has chosen, they receive the payoffs in the relevant box. As in chapter 1, given that we are concerned for now only with the basic intuitions behind each game, payoffs indicate only ordinal ranking of preferences over outcomes, rather than the weighting of preferences.

If we assume (for now: *ex hypothesi*) that in *Republic* book 2 Plato has set up something like a strategic two-player strategic game for modeling a Greek folk theory of the origins of social order, the question we must ask is, "what sort of game is it?" Here we consider four possibilities: Pure Coordination, Imperfect Coordination, Chicken, and Prisoner's Dilemma. Only the last will prove a tolerably good fit for the passages in *Republic* Book 2 that we have been considering, but the other games will be useful for thinking about other passages in Greek texts.

Pure Coordination. In the first game, we assume that both players are seeking similar goals – say (recalling Gyges) arriving at the destination of one of two equally nice kingdoms that lie at the opposite ends of a single road. The players are on the road, heading in opposite directions. The goals in this case are non-rival, that is, there is no competition between the players. The right strategy for each player in this game is to coordinate her actions with those of the other player so as to it avoid unnecessary interference in getting to her destination. Suppose that the players must pass each other and that the road is wide enough for them to pass without interference. All that is necessary for each to get to her destination without interference is that each stay either to (her own) right or left as they pass.

[Figure 2.2 about here: Pure Coordination]

As illustrated in Figure 2.2, two outcomes (1:Right, 2:Right and 1:Left, 2:Left), passing without difficulty, are equally good. The other two (1:Right, 2:Left and 1:Left, 2:Right) result in a crash, and so are equally bad. Both players prefer passing to crashing; neither has a preference for Right or Left. So, in this game there are two equilibria – that is, the condition in which neither player has a better move given the best move of the other player – it is just a question of which will be chosen. Assuming that there is some norm about passing (around here, when driving, we stay to the right… or left), the problem is solved. Absent a norm, if the players communicate, one can propose that each play Right (or Left) and the other will rationally agree. These commitments are mutually credible because they are backed by an expectation of best payoffs. So, with either a norm or communication, each player can expect to receive her full-value best outcome. With no norm or communication, the players would have an equal chance of passing or crashing, so the expectation for each would be receiving half of full value.[11]

Imperfect Coordination. In the second game, we assume that the players are headed for some mutually desired destination (say a nice kingdom), which they will gain if and only if they cohabit (one as King, the other as Queen). Otherwise (if they do not cohabit) each gets a low payoff (each lives as a commoner). Player 1 prefers that the capital, the city they will live in if they cohabit, be located on the coast (but rather likes mountains), Player 2 prefers the mountains (but rather likes the coast). In this game, coordinated play will result in one player coming out ahead with a great payoff; the other will get a good payoff. The alternative to coordinating their choices (1:Coast, 2:Coast or 1:Mountain, 2:Mountain) is that both get a low payoff. The situation is illustrated in Figure 2.3. Because they both prefer a good payoff to a low payoff, they will prefer coordination to non-coordination. Since each player does better, by avoiding the risk of the two low payoff outcomes, by coming to an agreement with the other, communication can lead to some fair way (say flipping an honest coin) to decide between living together on the Coast or the Mountain, and thus who gets the great payoff and who must settle for the good payoff. Because of the uncertainty introduced by the coin-flip, the expectation of each player in this game is gaining a part (but not all) of a full-value best outcome.[12]

[Figure 2.3 about here: Imperfect Coordination]

Chicken. In the third game, we assume that the situation is like Pure Coordination, in that the players are traveling in opposite directions on the same road, each headed for a nice kingdom, and must pass each other. But now suppose that the road is so narrow that they cannot pass unless one player pulls off the road, and is therefore delayed. Assume further that if one arrives at his kingdom before the other, he gets an extra payoff, becoming King (a Queen has promised to marry the one who arrives at a kingdom first) while the loser's reputation suffers, so he lives as a commoner. The options for each player are Swerve (chicken out by pulling over) or Straight (aggressively continue down the road). If 1 plays Straight and 2 Swerves, 1 gets a great payoff and 2 gets a low payoff; and vice versa if 2 plays Straight and 1 Swerves. If both play Straight (neither chickens out), both players crash, in which case each gets nothing. If both Swerve they both survive, but are equally delayed; each gets a payoff that is good (arrives at kingdom with no loss of reputation) but not great (the Queen tires of waiting on them and marries someone else). The situation is illustrated in Figure 2.4. In this case communication can take the form of a threat: Either can assert that he will play Straight no matter what; the other player's best move is then Swerve. If both threaten, but also fear the crash, the outcome is mixed (each might privately flip a coin to decide whether to follow through on his threat). Like Imperfect Coordination, the expectation in this game is for each to gain a part of their full-value best outcome.[13]

[Figure 2.4: Chicken about here]

Prisoner's Dilemma. In the fourth game, each player may choose to play "Cooperate," that is, seek a cooperative (in the language of *Republic* book 2: "just") outcome, potentially sacrificing some immediate personal advantage in favor of a longer-term individual *and* social advantage. Or he may defect from any cooperative scheme by playing "Defect" whenever it furthers his immediate interests, acting per Glaucon's specification of self-interested rationality as "injustice," to maximize his payoff without concern for the well-being of the other player.

[Figure 2.5 PD about here]

The usual story of the Prisoner's Dilemma game (hereafter: PD) is that each of two criminals, caught by the police, is offered a choice: rat out his partner, defecting from whatever agreement he had made with her (play Defect), or clam up, cooperatively stick by that agreement (play Cooperate). If one rats out (defects) and the other clams up (cooperates), the rat is let out free (payoff 4) and the clam serves a very long sentence (payoff 0). If each rats the other out (both defect), they each get a moderately long sentence (payoff 1,1). If both clam up (both cooperate), they are convicted on some lesser charge and each serves a short sentence (payoff 3,3). The situation is illustrated in Figure 2.5.

Other stories can be told about this game. For example, for modeling social order, suppose that, if one player defects and the other cooperates, the defector becomes an absolute monarch: she seizes all the goods of the cooperator and enslaves him. If both defect they each live in isolation and poverty. If they both cooperate they live relatively well, sharing returns to social cooperation, as fellow citizens. Each individual's highest payoff is gained by defecting while the other cooperates. Each individual's second-best payoff comes with mutual cooperation. Third-best is mutual defection. Worst is the "sucker's payoff" of cooperating when the other player defects. One other essential feature of this game is that we assume that adding the two quantities in each box determines the aggregate social value of the outcome. The hypothetical community (modeled by the two players) does best (gets the highest aggregate payoff) if both cooperate (3+3=6). The worst aggregate social payoff is when both defect (1+1=2).

In this game, as discussed in more detail below, the dominant strategy for each player is to defect: the predicted outcome is 1: Defect, 2: Defect. That proves to be a unique equilibrium; it dominates all other possible outcomes. This is because it is always in each player's best interest to defect: If 1 defects, 2 must defect or suffer the sucker's payoff. If 1 cooperates, 2 must defect in order to get the highest payoff. And vice versa. So, both defect, and the result is a low payoff for each and a correspondingly low social aggregate value. In Figure 2.5, and all subsequent figures of this kind (2.6, 2.7a, 2,7b, 2.8), there is an equilibrium solution when both quantities in a box are circled, indicating each player's best choice, in light of the predicted choice that will be made by the other player.[14]

If we try to imagine Glaucon's craftsmen of self-interest as players in one of the games sketched above, we can quickly dispense with Perfect Coordination and Imperfect Coordination. Plato's Gyges story is certainly not about coordinating in order to secure mutually advantageous access to non-rivalrous goods: There is only one Queen and one kingdom; Gyges kills the former king to get them. Moreover, when provided with magic rings, each of the "two men" takes goods from others, so goods are regarded as scarce rather than non-rivalrous.[15] Finally, in the origins of order passage, quoted above, Glaucon specifies that the players have suffered and fear suffering injustice, but in cooperation games there is no injustice. We can assume, then, that the game played by Glaucon's craftsmen of self-interest is competitive. Does either Chicken or PD fit Glaucon's specifications?

To decide that question, we can return to Glaucon's description of the folk theory of social cooperation: the hypothesized origins of the contractual agreement that results, "they say," in social order. Glaucon claims that, lacking the advantages that come with being uniquely empowered (as a Gygean ring-holder), each individual is in a compromised position: Her preferences are ranked in the order: (1) "act unjustly with impunity," (2) "avoid the worst," (3) "suffer injustice and be impotent to get revenge." Without a magic ring and assuming strategic play, she has no way to get the best outcome – doing injustice with impunity. Meanwhile, she lives in perpetual fear of the worst – the sucker's payoff of suffering injustice without recourse. *Everyone* is assumed to be in the same position of "lack[ing] the power at once to avoid the [worst] and choose the [best]." In marked contrast to Thrasymachus (in *Republic* book 1) and Callicles (below), Glaucon assumes equality of capability (to benefit from injustice) and liability (to suffer as a victim of injustice).

Glaucon's description of the situation maps quite well onto the PD, but not onto Chicken: According to Glaucon, each of the social contractors fears a sucker's payoff of suffering injustice and being impotent to get revenge (playing Cooperate to the other player's Defect), rather than the consequences of a catastrophe attending mutual acts of aggression (going Straight while the other also goes Straight), or losing out (playing chicken) due to risk aversion. The Defect-Defect equilibrium outcome of the PD does indeed land both players, symmetrically "in between the best and the worst." But, as we have seen, in a PD there is another "between best and worst" outcome (Cooperate, Cooperate) which is foregone and which would be better than Defect, Defect for both players. Glaucon's description of the origins of social order suggests that the contractors have agreed, that is, credibly committed, to cooperate on certain rules. So, if we imagine it as modeled by a PD, one way to think about the problem of the origins of social order is to ask: How did Glaucon's two craftsmen of self-interest get out of the dominant ow-payoff mutual defection equilibrium to the much better (although, for each, suboptimal) situation of mutual cooperation? Of course, Plato does not put that question directly, but if we have set up the problem in the right way, his reader is entitled to pose it.

The outcome of a game (based on the choices made by the players) depends on the players' information and their motivations, that is, their moral psychology. Before further considering the outcome when both players (as in the standard PD) are fully and symmetrically informed and motivated per Glaucon's definition of self-interested rationality, we may consider the results of similar games with players who are assumed to be rational in that they have orderly preferences and beliefs about the state of the world, but are either differently motivated or asymmetrically informed.

Aristotelian Justice. Suppose, first, that the two players are citizens of Aristotle's "polis of our prayers," the practically-achievable ideal community described in the *Politics*, book 7. As illustrated in Figure 2.6, we stipulate that the players are equally virtuous "Aristotelian" citizens. Having been socialized in the rules and the education of the best possible polis, each is fully committed to choosing and acting according to an Aristotelian conception of justice. Each will, therefore, make his choice based on Aristotle's two, compatible, definitions of justice: The first definition is based on a principle of equality: those who are equal, in the relevant sense, ought to receive equal shares of whatever good is being distributed. In the polis of our prayers, distribution of the relevant goods is according to individual virtue. Aristotle's second definition of justice is "the common good" – that is, the advantage of the "whole" (in the first instance, the polis), irrespective of the particular advantage of each of its parts (variously: individuals, families, villages, factions,).[16]

[Figure 2.6 about here: Aristotelian Justice]

Given these stipulations, each player in the Aristotelian Justice game will choose Cooperate. The outcome of the game will be in the upper-left box, so the payoff is 3,3. Both players cooperate because they are equally virtuous (we assume that this is common knowledge), and so they choose to distribute the relevant good (the payoffs in the game) equally. The upper-right (0,4) and lower-left (4,0) boxes are very unequal, and so these outcomes are rejected as unjust. Although defecting would gain an individually higher payoff (say, more time to engage in pure contemplation; see *Nicomachean Ethics* book 10), each player prefers a just equal distribution with a lower personal payoff to a higher personal payoff gained unjustly, that is at the expense of a peer.[17] The lower-right (1,1) box is, however, also an equal distribution. So is the play Defect, Defect equally good, from the perspective of the players' motivations – and thus a second equilibrium? The answer is no, because of the second definition of justice as "the common good." The sum of the quantities in each box, the aggregate social value, defines (in this simple game) a common good. Since 6 > 2, Cooperate, Cooperate is uniquely preferred by both players.

The game models a stable Aristotelian community because the players will choose this outcome over all alternatives: Given their motivations, neither player has a better move to make in this game. So, in the terms of game theory, the Aristotelian Justice game has a unique equilibrium solution. Because the players have aligned preferences, it is a variant on a coordination game rather than a competitive game.

Thrasymachean Injustice. Next, let us assume that one of players in the game has the ability always to get his best payoff at the expense of the other, per Thrasymachus' claim that the unjust man always comes out ahead in his dealings with a just man, to the latter's disadvantage (343d), and that, taken to its extreme (tyranny), the unjust man is completely happy and the unjust completely miserable (344d). This sort of outcome could be achieved if one player holds a Gyges-type invisibility ring and the other player does not. Per Thrasymachus' specifications, we suppose that the player with the ring will act to maximize his own payoff, irrespective of anyone else's welfare, consistently playing Defect. Whatever his motivations, the player without the ring is vulnerable to the choices of the player with the ring because he lacks essential information (regarding the type of the other player, the available moves, and perhaps even that he is in a game) and so, out of ignorance, will unwittingly play Cooperate. And thus, depending which player has the ring and which does not, the outcome to this sort of game ends up in either the lower left (4,0) or in the upper right corner (0,4), per Figures 2.7a and 2.7b.

[Figures 2.7a and 2.7b about here. Thrasymachean Injustice I and II]

Given that "ignorant play" violates the ordinary game-theoretic assumption of full information, we might want to change the background story for this game, abandoning the contrivance of a magical ring, and supposing instead that one player is a skilled rhetorician, as Thrasymachus was reputed to be, with powers of persuasion of the sort that the sophist Gorgias of Leontini, in Plato's dialogue named for him, asserted that he and his students possessed (Plato, *Gorgias* 452d-457c – see further, below).[18] That is, one player has sufficient persuasive power to ensure that the other player's choice will be in line with the preferences of the rhetorician. So the story might go that the row player (in Figure 2.7a) or the column player (in Figure 2.7b) has persuaded the other player to cooperate, perhaps by promising to cooperate himself and emphasizing the superior long-term value to individuals of the cooperative outcome that will ensue – all the while intending to defect, and thus gain his own highest payoff. This result assumes, with Gorgias, that rhetoric actually is a power that is capable of determining others' beliefs, if not of changing their underlying preferences. As we have seen, in strategic game theory, a promise that is not backed up by a credible (rationally self-interested) commitment to act on it is considered "cheap talk" and will be discounted by a rational player, absent some means of ensuring its credibility.[19] Gorgias' account of persuasion in Plato's dialogue seeks to distinguish between the readily discounted cheap talk of the untrained speaker and the irresistible belief-establishing force that is exerted upon an audience by the master of rhetoric.

Assuming that one player actually does have the power to determine the other player's beliefs, impelling her to cooperate and thereby gaining his own best payoff by defecting, each of the games illustrated in Figures 2.7a and 2.7b has a unique equilibrium. If we extrapolate from this game to an actual society, we may say that these two games model a radically unequal society of the sort advocated by Thrasymachus in *Republic* book 1 (and Callicles in Plato's *Gorgias*, below), in which the power (however achieved and maintained) of an individual or a coalition is great enough to ensure the stable domination over the community by that person or group. The aggregate social payoff (4+0) of this "Thrasymachean" society is lower than the cooperative "Aristotelian" society (3+3), but higher than the (1+1) PD society arising from mutual defection.

2.3. Glaucon and Callicles on the emergence of order

Having considered variants on the PD game with players motivated or informed so as to result in a unique equilibrium solution in each of the four possible outcomes, we may return to what I will call "Glaucon's Dilemma" – the situation in which the players are perfect craftsmen of self-interest and playing a PD (Figure 2.5). As we have seen, each player is predicted to defect, resulting in the unique low-payoff equilibrium of the lower right box. This is an equilibrium, because neither player has a better move - if either were to choose to cooperate, he must expect the other player to defect, and thus he would end up with a sucker's payoff.

This result may be described as a dilemma for at least two reasons. First, the mutual defection equilibrium gives each player his third-best (second-worst) payoff. While each avoided the worst case of the sucker's payoff (0), had they played Cooperate, Cooperate (as in the Aristotelian Justice game) each player would get his second-best payoff (3). It is, furthermore, a dilemma because the assumed community in question gets its worst available aggregate social value (2), falling short of the second-best social value (4) that would have resulted from a (Thrasymachean) Cooperate, Defect play, and far short of the best aggregate value (6) that results from (Aristotelian) Cooperate, Cooperate play.

In a one-shot game, in order to avoid the worst, a fully informed rational player must defect, and so each receives the low payoff of the lower right quadrant in Figure 2.5; we end up at a very low-payoff equilibrium lacking cooperation. But the people in Glaucon's "origins of order" story seem to have had repeated interactions in the course of which each has experienced *both* the advantages of doing injustice *and* the harm of suffering it: It is "when men *do wrong* and *are wronged* by one another and *have experienced both*," that they come to the realization of the value of the cooperative contract. So, imagining this as a game that is played more than once and is expected to be played indefinitely (without a specifiable "last play"), we might imagine that each of the players began by playing a mixed strategy (randomly playing Cooperate and Defect), so that each received at least one fairly large payoff (Defect, Cooperate: did wrong, and benefited substantially by it) and sustained at least one very large loss (Cooperate, Defect: suffered wrong and was badly harmed by it). As contemporary work in game theory has established, repeated play games have multiple possible equilibria, and so it is quite possible for *us* to imagine an alternative to the low payoff of the one-shot PD.[20] The question is whether Plato could have been driving at something similar.

Glaucon specifies that the *costs* of suffering wrong *exceed* the *benefits* of doing it (for example, in the PD: the payoff of 0 is further below the Cooperate, Cooperate payoff of 3 than the Defect, Cooperate payoff of 4 is above it), so that each has suffered net losses (relative to Cooperate, Cooperate) as result of mixed play.[21] Without a contract, each can limit future losses by playing a Defect strategy. Mutual injustice (Defect, Defect) avoids the least-preferred outcome of being wronged without recourse, but it results in the lowest aggregate social payoff. So we might press the passage by imagining that, having each played Defect, Cooperate; Cooperate, Defect; and Defect, Defect, the people referred to by Glaucon have learned that "the [individual *and* social] excess of evil in [each *and* all] being wronged is greater than the excess of good [to each] in doing wrong." And that they all act accordingly: agreeing to establish rules for mutual cooperation.

If that is the right reconstruction, learning from the experience of doing and suffering injustice leads the people in Glaucon's experiment to make an agreement that all will renounce doing injustice, and so, in effect, consistently play Cooperate. If all stick to the agreement, this is mutually beneficial (although not optimal for anyone); everyone not only avoids the worst, but also gets his or her second-best payoff. Moreover, the society as a whole gets its highest (3,3) payoff. And this agreement is, then, "they say," the origin of the concept of a sense of justice as some sort of fairness, one version of which is the Aristotelian definition of justice sketched above. Moreover, and essential to our question of the emergence of social order, this sort of agreement is said to be the origin of authoritative laws with the power to command each individual not to make certain self-aggrandizing choices, not to engage in behavior that would (just so long as others were law-abiding) lead to her most-favored outcome and thereby most fully satisfy her preferences.[22]

So understood, Glaucon's account of what "they say" about the origin of social order in the choices of rational persons, resulting in a compact that creates law, is reminiscent of that of Thomas Hobbes in *Leviathan* (1996 [1651]). Like Glaucon, Hobbes postulates self-interest (fear of harm to self and self-aggrandizing "glory") as the primordial motivation of individuals. Like Glaucon, Hobbes postulates a rough equality among the persons living in prepolitical conditions; given weapons and temporary coalitions, each has the potential to threaten all others. Hobbes' solution, like Glaucon's, is a social contract, in which individuals agree to give up some of their pristine freedom of choice and action because they recognize that a contract is their only way out of the miserable conditions that result from mutual defection.[23]

We need not leap all the way forward to seventeenth-century social contract theory to find comparisons to Glaucon's origin story. In Plato's dialogue *Gorgias*, Callicles is introduced as an aspiring politician in Athens and a student of the master rhetorician Gorgias, from whom he hopes to learn sophistic rhetorical techniques. These will, Gorgias has promised, allow Callicles to dominate others by means of persuasive speech. Thus Callicles expects to have the opportunity to behave without concern for social constraints: in effect, rhetoric will be his magic ring. In his first extended speech in the dialogue (Plato, *Gorgias* 483a-484b), Callicles makes a point of contrasting *phusis* with *nomos*, proclaiming that, by nature, the most formidable of men, those with the power to "have more" (*pleon echein*),[24] use their strength to take more – that is to satisfy their preferences by maximizing payoffs to themselves at others' expense. He illustrates his point by reference to the presumably natural behavior of non-human animals and imperialistic Persian monarchs.[25]

In contrast to the rule-makers in Glaucon's origins of social order story, who were equal in capacity and mutually fearful, according to Callicles, the original authors of laws that constrain the naturally strong, and thereby frustrate naturally self-aggrandizing behavior, are the weak and many.[26] They make laws forbidding anyone to "have more and more" (*pleonektein*) because they fear (*ekphobountes*) the formidable few. Thus the weak establish behavior-constraining laws in their own interest.[27] The many and weak also establish norms that prescribe blaming those who do behave in blatantly self-aggrandizing ways. The many prefer that all have equal shares, because, being inferior and unable to get greater shares, equality offers them the best payoff they can hope for.[28] In terms of the games sketched above, Callicles seems to imagine social order as a variant on the Chicken game, in which the chickens rule and have set up traffic laws to block the option of anyone aggressively driving Straight.

Callicles' argument about the origin of constraining *nomoi* is, however, also in some ways reminiscent of Glaucon's account of the origins of social order. In each case, social order, exemplified in laws and norms, is established because of rational fear of the bad consequences arising from a situation in which individuals do as they please, and take as much as they can get. The difference is that Glaucon identifies the motivation for lawmaking as a universal mutual fear that the cost of suffering injustice is, for each individual, greater than the benefit from the opportunity to do injustice. In Callicles' story, by contrast, the pre-social human population is already sorted into many weak persons and a few strong ones. As in Glaucon's thought experiment, Callicles assumes that self-interest is a universal human motivation. His formidable few and weak many have identical moral psychologies; they differ in their capability and aggressiveness, but not in their motivations. The many and weak fear the few and strong, because, one-on-one, strong individuals have the will and the capability to take more for themselves. Recognizing that they will be losers if the strong are unconstrained, the individually-weak many choose, as their best available outcome, to make laws, imposing an unnaturally egalitarian social order on strong and weak alike.

Callicles' distinction between the natural order, in which the individually formidable few take more, and the egalitarian legal order created by the individually weak many, suggests that the existing social equilibrium is unstable. He analogizes the means by which the weak constrain the strong to a practice of capturing lions while still cubs, and enslaving them by enchantments.[29] The enchantment consists of telling the strong that it is necessary to have equal shares, and that this is what is fair and just. He also alludes to the possibility of a transition to a new order of things, one in which nature will have reasserted itself. Callicles predicts that a real man will someday emerge, one who has a "sufficient nature" (*phusin hikanên*). That man will rid himself of all constraints, and proceed to "trample on our documents, and magic tricks, and charms, and on all the laws that are contrary to nature. Then our slave will be revealed to be our master, and the true justice of nature will shine forth."[30] Clearly Callicles fancies himself for the role.

2.4. Comparative dynamics

A central problem with Callicles' argument from nature and strength is quickly revealed when Socrates' puts Callicles' conclusions to the test: Since the weak many were, by Callicles' account, able to constrain the strong few, the many are, self-evidently, collectively more powerful than the few. The collectively powerful many do just what the individually powerful few would prefer to do: maximize their own expected payoffs.[31] Egalitarian laws are, therefore, artifacts of (collective) strength used to achieve a preferred (individual and social) outcome. Social order, in the form of *nomos* simply is the will of the strong – that is to say, it is identical to Callicles' conception of *phusis*. It remains to be seen how it is that the many were capable of inaugurating and sustaining coordinated collective action at scale; that question will be taken up in texts considered below. But, assuming (as Callicles does) that many self-interested individuals are in fact capable of collective action, Callicles' "advantage of the strong" argument is self-defeating.

If Callicles-type claims represent the standard sort of argument that was being made by those classical-era Greek intellectuals who concerned themselves with the social and political implications of the *nomos-phusis* distinction, we can see why Plato used Glaucon's argument in *Republic* book 2 to set the challenge that Socrates must meet if he is to demonstrate that justice is a good in itself. Glaucon's account of human motivation and social order does not suffer from the mistake that renders Callicles' argument from natural individual strength self-defeating. But it does raise a fundamental question for which it offers no easy answer. By the premises of Glaucon's thought experiment, humans are rationally self-interested, so it is in each individual's best interest to defect. We expect, therefore a low payoff Defect, Defect equilibrium. But cooperation offers higher individual and social payoffs. Realizing the high costs to each and all is, Glaucon's account of the folk theory suggests, *why* humans agree to contract with one another. The worry is that Glaucon's argument is based on a fallacious teleological functionalism: the good social outcome mysteriously determines the behavior that enables it.[32]

As suggested above, Glaucon hints at repeated play, which (as game theorists now know) allows for multiple equilibria. But *how*, given the motivational premises of the thought experiment, does the society make the transition? How does each rationally self-interested individual avoid the sucker's payoff of suffering injustice without recourse during the transition? This is the problem of comparative dynamics. It is easy enough to model different equilibria (as we did, above, by varying the motivations and information available to the players), but hard to explain just how it is that a society moves from a less to a more favorable equilibrium. Glaucon's Dilemma, as sketched above, is that anyone agreeing in advance to abide by a contract, would open herself to a sucker's payoff. Since there is no third-party enforcement of prior agreements, no one has the right incentive to make that first move, or to enforce the rules at cost to herself, after the rules have been agreed upon.

This is the point of Hobbes' social contract argument in *Leviathan*. Hobbes' pre-social humans agree upon the original contract among themselves specifically to create a sovereign: a lawless third-party maker and enforcer of rules. That solution was indeed considered in the Greek tradition, notably in Herodotus' stories of the origins of Asian monarchies (chapter 3). But Glaucon does not take that step. The contract he alludes to seems not only to be collectively agreed-upon but also self-enforcing. By claiming that, just because a certain state of affairs will be better for each and all once it is in place, it can be readily brought about and sustained, he appears to have invoked a simplistic sort of functionalism, thereby violating the original premise of his own experiment: Given her preference for self-aggrandizement, each self-interested individual will violate rules whenever it is in her interest. Glaucon's account of emergence of social order suggests that repeated interaction, with different payoffs, allows people to learn about relative costs and benefits of defection and cooperation But if we stay within his strictly self-interest-centered premises, we cannot explain how what they learn is operationalized through credible commitments and collective action. As we will see, below, thinkers in the Greek tradition corrected that deficit, showing how repeated interaction could solve "Glaucon's Dilemma."[33]

Of course Plato's characters Glaucon, Adeimantus, and Socrates, *reject* the premise that a truly rational person – one who recognizes reason as a highly valued end in itself, essential to the achievement of true happiness (*eudaimonia*), rather than merely as a means to other (inferior) ends – will have the preferences posited in Glaucon's challenge. Readers of Plato's *Republic* know that the authority problem is solved in the ideal state of Callipolis by the establishment of Philosopher-kings, who are supported by a military coalition of Guardians and equipped with an ideological toolkit of Noble Lies. But for our present purposes – investigating the main lines of Greek thought about the rationality of choice, rather than Plato's highly distinctive ideas about what would constitute an ideally just society – it is more relevant to ask whether the comparative dynamics problem raised by the origins of social order in the face of Glaucon's Dilemma was recognized in other surviving Greek texts, and, if so, how it was addressed.

The rest of this chapter sketches how a few Greek texts addressed the question of the emergence of order, focusing on the problems that order was required to solve, and assessing a mechanism by which coordinated action might be effectuated. Accounts of the pre-social human condition have come down to us in the works of Thucydides (1.2-8), Plato (*Protagoras*, *Statesman*, *Laws* book 3), Polybius (6.5-6), Diodorus Siculus (1.8, 1.90), and a few later writers. These fragments of what was in antiquity a much fuller tradition about the primitive past were collected and skillfully analyzed by Thomas Cole (1967), who sought to make sense of the doxographic tradition (i.e. sorting out whose work influenced whose). Cole suggested that many of the surviving texts concerned with what he called "Greek Anthropology" owed a debt to the mostly-lost work of the fifth-century Athenian writer, Democritus.

Leaving aside the fascinating, but for our purposes, tangential, puzzles of doxography, it is clear enough that the general question of the emergence of social order was raised in a wide range of classical and Hellenistic texts (section 2.5). I will argue that the specific problem I am calling Glaucon's Dilemma was addressed in Plato's *Protagoras* (section 2.6). Plato's Protagoras self-consciously amends the assumption, common to "Thrasymachean" versions of the folk theory, that rational humans are motivated by purely egoistic preferences. He invokes a moral psychology, common to most, but not all persons, that adds to the preference ordering of rational choice-makers an inherent orientation toward justice (*dikaiosunê*), understood as a sense of equity (*dikê*), and toward shame (*aidôs*), understood as a tendency to impose psychological costs upon oneself (internalizing others' blame) when seeking to gain benefits by acting unjustly. The PD is thereby transformed into an assurance game with two equilibria, one of which offers relatively high individual and social payoffs.

Protagoras drives his hypothesized community towards the high-payoff equilibrium by modeling human society as a repeated game with uncertainty (imperfect information) and updating (players make choices in subsequent rounds based on the outcome of previous rounds).This provides an explanation for the existence of dynamically sustainable (if not perfectly just) forms of productive cooperation through credible commitment to constraining social rules. It does so in the face of less-than universal cooperation and without the metaphysics underpinning Plato's own preferred solution to the cooperation problem, as developed in the later books of the *Republic*.

2.5. Towards social order

Although the Greeks were aware of a tradition that there had been a "Golden Age" of ease and plenty sometime in the mythic past, the dominant Greek way of thinking about change over time was a narrative of progress (Cole 1967).[34] That is to say, a well-educated citizen of a Greek polis, living in the age of Plato, would be likely to suppose that the technological, economic, and political level of his contemporary world was the product of prior social development. In the distant past (as, in his own time, on the fringes of what he took as civilization) people had lived at a primitive level: in poverty, without advanced technology, without law or social order beyond the level of close kin. As such, they were continually exposed to existential threats.

Existential threat was a central theme in Greek accounts of early human existence: Mankind in a primitive state of development hovered at the edge of extinction. People were scattered across the landscape, living as individuals or in tiny kinship groups; there were no substantial towns or cities. Depending on the ancient source, technological primitivism might include the lack of knowledge of the use of fire, metals, and agriculture.[35] Even with basic technology to aid them, humans were endemically exposed to attacks by predatory animals.[36] But human predators were also a source of danger, both the naturally formidable individuals alluded to by Plato's Callicles, and the bands of pirates and marauders who are Thucydides' concern in the "Archaeology" (1.5-8; see chapter 6).[37] It was fear of these endemic threats to vulnerable individuals and small bands that motivated attempts to cooperate at greater social scale.[38]

I have suggested that Plato's *Republic* Book 2 account sets up a comparative dynamics dilemma: how did rationally self-interested individuals cooperate well enough to move forward, from the primitive level at which they began, to the relatively flourishing societies known to classical-era Greek readers? In some other Greek texts the issue of the origin of social order is described in sociological terms of a group-level response to danger or divine fiat. Plato, in the *Laws* (3.678c-681a) says that the threat of wild animals led to the creation of communities that surrounded themselves with rough walls. Within these communities aristocratic rulers (originally heads of kinship groups) cooperated on matters of social order. In the *Statesman* (274b-d), Plato says that, after the end of a mythic "age of Kronos," during which a divine shepherd had cared for the human flock, weak and defenseless humans were preyed upon by wild animals. They were saved by the gifts of the gods: fire, technology, and agriculture, and later they came to be ruled by monarchs. Aristotle (*Politics* 1252a-b) invokes an inherent sociability that led humans through several developmental stages from the nuclear family (plus slaves) ruled naturally by the father, to villages and extended kinship groups ruled by king-like clan leaders, and eventually to the polis. We will consider Aristotle's account in more detail in chapter 4.

Other authors offer simple explanations that seem to come down to the level of individual reactions to specific stimuli, and specifically to fear arising to existential threats: Porphyry (*De Abstinentia* – Cole 1967: 71-72), a third century CE philosopher, drawing on a classical-era source, says that people came together in communities motivated by fear of wild animals and men of evil intent, recognizing the advantage of mutual aid. Diodorus Siculus (1.8) goes one step further: "Since they were attacked by the wild beasts, they came to each other's aid, being instructed by expediency (*to sumpheron*), and when gathered together in this way by reason of their fear, they gradually came to recognize their mutual characteristics."39

These "response-to-threat" accounts of the origins of order do not seem to assume the Thrasymachean variant of the folk theory, and may be modeled by some variant on one of the two coordination games discussed above. As we saw, coordination in a condition of non-rivalrous goods is facilitated by communication and by shared norms. Coordination is more difficult, however, when there is a shared desire for order and no advantage to defection, but there is no communication, no established norms, indeed no single path to follow. Thomas Schelling (1980 [1960]) generalized and formalized the solution to the problem of coordination without communication or preset rules by reference to "focal points" (now sometimes called Schelling points): A focal point can be any commonly recognized symbol. Schelling's examples were prominent landmarks: for natives of the New York city region, for example, the information desk at Grand Central Station. Common knowledge of the focal point, among people with a reason to coordinate their actions in order to reach mutually desired result (they want to find each other for a lunch date in a big city), enables people to achieve a mutually desired cooperative outcome (they find each other), even in the absence of a pre-arranged plan of action (they forgot to discuss where they would meet). Each rationally chooses to coordinate on the focal point (shows up at the information desk at noon), so the problem is solved and all involved gain their desired end.

Diodorus Siculus (1.90-1-2), in a passage that may ultimately derive from Democritus, offers something like Schelling's focal point coordination solution in a discussion of the emergence of early social organization in Egypt. Diodorus' Egyptian narrative assumes that humans had, per above, come together in groups for security against wild animals, but he then brings up the Thrasymachus/Callicles issue of the presence of strong and aggressive elements in the extended human ecology within which these groups had formed.

> When men first ceased living like the beasts and gathered into groups, at the outset they kept devouring each other and warring among themselves, the

> more powerful ever prevailing over the weaker; but later those who were deficient in strength, taught by expediency (*to sumpheron*), grouped together, and took for the device on their standard (*sêmeion*) one of the animals which was later made sacred; then, when those who were from time to time in fear flocked to the standard, an organized body (*sustêma*) was formed which was not to be despised by any who attacked it. And when everybody else did the same thing, the whole people came to be divided into organized bodies (*kata sustêmata*).[40]

Here we have a story of primeval misery (including cannibalism) that is compounded by civil strife, and domination by the powerful. The response to these threats is coordination among the weak. What is distinctive in Diodorus' Egyptian narrative is the introduction of a non-verbal focal point as the mechanism facilitating coordination: A banner or plaque with a picture of an animal serves as a device that enables multiple weak individuals to assemble in response to existential threats. Although, as in Diodorus' earlier account of primitive cooperation against wild animals (above), expediency (*to sumpheron*) is the "teacher," it is the focal point of the animal-standard that allows many individuals to act in a coordinated manner, as an organization (*sustêma*). Notably, Diodorus says that the use of the animal-standard was widely recognized as an effective coordination mechanism: It is adopted in a cascade of adaptive imitation by the rest of the Egyptian population, which thereby comes to be organized (*kata systêmata*).

Diodorus' account, like, for example, that of Thucydides (1.3.1-1.6.), emphasizes the high costs of non-cooperation – mutual threat, strife, and self-aggrandizement by the powerful. Diodorus offers a neat mechanism, in the form of a Schelling-type focal point, which explains how coordination was effected in the face of those threats. But he does not solve our puzzle of how the advantageous new order was sustained in equilibrium. Coordinated mass action is readily explicable as a one-off event in a large body of persons (Hardin 1991), as it is in the costless Pure Coordination case. But Diodorus' story concerns what becomes habitual cooperative behavior under conditions of resource scarcity (wars of aggression). That behavior is costly, both to each of the many weak individuals who must confront danger once they have rallied together, and to the predatory powerful whose preference for domination is frustrated by the use of the mechanism.

How was the response-to-threat, focal-point-based social order sustained over time, such that it became well organized and dependable? If we are to address the Thrasymachean motivational assumptions of Glaucon's Dilemma, we still need some reason that a rational individual would not seek to free ride on the cooperation of others – in this case, letting the others take the risks involved in flocking to the standard and opposing the powerful – and why the threat of free riding would not precipitate a cascade of defection.

2.6. Plato’s Protagoras on order and motivation

A well-known passage in Plato’s *Protagoras* addresses Glaucon’s Dilemma head on, by suggesting that sustained cooperation at scale could not be achieved in a population of “Thrasymachean” rationally self-interested agents just by their recognition of the benefits that would accrue if the new order were in place. The dialogue features a long speech by Protagoras of Abdera (322a-328d), a famous sophist who is visiting Athens, offering to teach fee-paying young Athenians an advanced course in the political craft (*politikê technê*) – the effective management of households and states. One of the central points of the speech is that solving what we are calling Glaucon’s Dilemma is impossible until and unless the assumed psychology of the agents is, first, suitably modified and they then enter into a repeated game with updating and communication.

As in every Platonic dialogue, it is important to keep in mind that it is Plato who provides each character with his lines. There certainly was a real sophist named Protagoras; we have a few short quotations from his works preserved, but it would be wrong to think that Plato’s Protagoras is true in every (or any) respect to the thought and expression of the original. In what follows, when I refer to “Protagoras” – I mean (as in the case of, e.g., “Glaucon” or “Callicles”) Plato’s character in the dialogue rather than the historical person. On the other hand, it is reasonable to assume that each of Plato’s characters is carefully drawn to exemplify and articulate some intellectual position that Plato thought worthwhile exploring or exposing. Plato’s Protagoras presumably develops a position that Plato’s original readers would have recognized as relevantly similar to positions being argued either by the real Protagoras or by other sophists in the classical era. Since what we are after is a better understanding of background Greek ideas about rationality and choice, that is good enough for our purposes.[41]

Unlike Plato’s Gorgias, Plato’s Protagoras does not advertise his art as enabling its possessors to dominate others by special powers of persuasion. Rather, he suggests, he has developed a higher-order version of the sort of informal teaching and learning that characterizes (he argues) a reasonably well-functioning Greek polis – here exemplified by contemporary Athens. In order to situate his art in what amounts to the democratic context that has been sketched by Socrates in the course of challenging Protagoras to explain how virtue can be taught, Protagoras tells what he explicitly describes as a fable (*muthos*) of human origins. His fable seems specifically designed to address Glaucon’s Dilemma.

According to Protagoras’ creation story, we humans were initially brought into existence by divine fiat, but the process was mishandled, leaving us without the life-preserving natural capabilities enjoyed by other animals. In order to secure the survival of the human race, Prometheus stole fire and technology from the gods (and, famously, was punished accordingly by Zeus). Thus provided, humans lived apart from one another. But, lacking the craft of war, which Protagoras defines as one part of the "political craft," they found themselves defenseless against wild animals. "So they sought to band themselves together and secure their lives by founding poleis. Yet as often as they were banded together they did injustice to one another through the lack of political craft and thus they began to be scattered again and to perish" (322a-b).[42]

What is notable here is that Protagoras has added a stage to the standard Greek origins story: As in the "response-to-threat" accounts, humans gather together in order to preserve themselves from dangers that threaten their lives. But in Protagoras' story, they cannot sustain cooperation at scale, due to their tendency to wrong one another – to act unjustly, presumably because of their "Thrasymachean" motivations. Here, then, is Glaucon's Dilemma: the dominant strategy of defection makes beneficial, high-payoff cooperation impossible to sustain *even in the face of existential threats* for those with (we must, I think, assume) the narrowly egoistic rationality that Glaucon will attribute to his hypothetical craftsmen of self-interest.

At this point in the story, divinity reenters the picture: To forestall human extinction, Zeus commands that original human moral psychology (presumably something akin to Glaucon's account of egoistical self-interest) be augmented by distributing "shame (*aidôs*) and concern for equity (*dikê*) among men, so that there would be order (*kosmoi*) within poleis and bonds (*desmoi*) of friendship to unite them." The new moral psychology is to be distributed generally: Zeus orders that all (*pantes*) must share in it because "poleis cannot exist if only a few people have a share, as is the case with the other crafts." But Zeus then decrees death, "as a public health hazard" (*hôs noson poleôs*) for anyone who is incapable of sharing in shame and a sense of equity (322c-d).[43] The fable thus acknowledges the possibility that some individuals may retain the previous, unaugmented moral psychology and that their behavioral motivation could, like a contagious disease, spread through the community. Such individuals are, therefore, dangerous enough to social order to require a rule mandating their extermination.

Moving from mythology (*muthos*), to explanation (*logos*), Protagoras demonstrates the analytic value of his myth by pointing out that, now that there is order in poleis, in an ordinary Greek community like Athens, if a man is known to be unjust, and publicly admits to being so, his behavior would be considered evidence of madness (*mania*). And so, "everyone, they say, must claim to be just, whether he is or is not, and whoever does not make some pretension to justice is mad; since it is necessary that all without exception share in it in some way or other, or else not be of human kind" (323b-c).[44] With this normative behavioral standard in mind, Protagoras proceeds to explain the practice of mutual instruction among the residents of the polis in terms of self-interest: "for our neighbors' justice and virtue, I take it, is profitable (*lusitelei*) for us, and consequently we all willingly speak out and teach one another in matters of justice and lawfulness" (327b).[45] Part of this mutual instruction is public punishment for wrongdoing. Punishment is not, according to Protagoras, rightly understood as retrospective vengeance (an attitude he regards as suited only to beasts) but as future-oriented correction and deterrence. Those wrongdoers who do not respond to this sort of correction must be expelled from the polis or put to death (325a-b). This last proviso recalls Zeus's injunction to execute those incapable of sharing in the new moral psychology.

Looking back from *logos* to *mythos*, we can recognize that Protagoras' fanciful story about divine intervention has enabled him to address the problem of comparative dynamics by, in effect, running a cooperation thought experiment twice. In the first run, the premise is that humans are, as Glaucon will specify, egoistically self-interested and therefore, even in the face of existential threats, they will fail to cooperate in ways that can bring about a stable social order: the equilibrium (reminiscent of Hobbes' state of nature) is "scatter and perish." In the second run of the experiment Protagoras assumes a prevailing moral psychology in which egoistic self-interest is augmented by a sense of shame and justice, and thus the scope of self-interest is extended from "just me" to "me and us." In the second run, as in the first, humans remain rational, in that they have orderly preferences, and they are self-interested in that they act on the basis of beliefs in ways that are expected to fulfill their highest-ranked available preferences. But their beliefs about what actions will promote their interests have changed in salient ways.

In describing the expected public response to an individual who admitted to being unjust, Plato's Protagoras has employed the same terminology of "madness" that is used by Glaucon (*Republic* 359b) to describe what an ordinary, honest Greek would think of someone who possessed a Gyges ring and failed to use it to satisfy his preferences. In Protagoras' description of the behavior of members of a society in the experiment run with agents possessing the "shame and justice" psychology, there is no indication that individuals are doing other than what they regard as being in their own (joint and several) interest: It is because our neighbors' justice and virtue is beneficial *to us* that we choose to assume the costs of speaking out and instructing one another. One part of that instruction is punishment of violators: Protagoras conceives of the motivation for costly punishment in the economic sense of anticipated future outcomes, rather than the vendetta sense of retrospective payback.

The shame-justice psychology makes it possible to imagine a transition from the "scatter and perish" equilibrium to the "order in poleis" equilibrium. It does so because *ex post*, each person's best payoff comes with cooperating, just so long a he has reasons to believe that others will act likewise. And he has reason to believe that because cooperating does bring the best payoff to all others who share his moral psychology. Once again, if we remember that Protagoras has explicitly cast his story as a myth, we will realize that there has not actually been a transition. Rather, Protagoras has shown that a purely egoistical rationality, per Thrasymachus' and Glaucon's "craftsmen of self-interest," is implausible as an account of ordinary human motivation: It fails to explain the observable fact of the role of cooperation in social order. Thus, human motivation must be assumed to be other than purely "Thasymachean." When that assumption is operationalized, what happens, in the language of game theory, is that a Prisoner's Dilemma has been transformed into a Stag Hunt, an assurance game in which the highest payoff for each strategic player is gained by rational mutual cooperation, rather than by rational defection,.[46] This is illustrated in Figure 2.8: the Gifts of Zeus game.

[Figure 2.8: Gifts of Zeus about here]

The only difference between the Gifts of Zeus game and the PD is that the payoffs for defecting when the other player cooperates have been lowered from 4 to 2. This is in recognition of the shame-justice moral psychology: Anyone with an internalized sense of shame, as respect for social norms of just cooperation, bears an internal cost when she acts so as to outrage that sensibility. So, rather than the full payoff that the defecting player received in the PD, in the Gifts of Zeus game the Defect, Cooperate player is docked, as it were, part of the original payoff. While enjoying the benefits of actions that help herself at the expense of another, she is simultaneously ashamed at herself for having done so. And so, some part of her happiness is forfeit. The way to think about this is, I think, just in terms of the contents and ordering of her preferences. She may want to acquire material goods, access to sex, and power – but those preferences are now subordinated to a high-order preference to act, and to be acknowledged as acting, according to her sense of shame and equity.

The adjustment to the payoffs has the effect of adding a second equilibrium outcome to the game. Defect, Defect, with its low payoff, remains an equilibrium outcome, because in the event that either player would choose Defect, Defect is the best play for the other. If I believe that the other player is of a type to play Defect (i.e. is one of those with what is now regarded as an impaired moral psychology), I must play Defect in order to avoid the sucker's payoff. But I no longer need to suppose that every rational player will choose Defect, because playing Defect now offers those with the normal shame-justice moral psychology a maximum payoff of 2, while playing Cooperate offers a higher maximum payoff of 3. So, assuming that this is a full information game (like the Aristotelian Justice game, above), that each player is rationally seeking his or her highest payoff, and that all this is common knowledge, each player can play Cooperate. Playing Cooperate, Cooperate (payoff 3,3) is a second equilibrium, one that not only gives each player his highest available payoff, but is also the highest social aggregate payoff. And thus, the Gifts of Zeus game models a society that is has the potential of realizing increased aggregate social value, as well as individual preference satisfaction through having solved the problem of rational cooperation.[47]

2.7. Repeated play: Learning, punishment, and population dynamics.

Upon reflection, it may appear that in transforming the PD-like Glaucon's Dilemma into the Gifts of Zeus assurance game we have overstated the actual social conditions described by Protagoras. The new moral psychology is not universal: some agents remain rational defectors, and, given the majority opinion of their psychology as dangerous (indeed diseased) they have every reason to dissimulate. At least some of those may be willing to take their chances with punishment. They may reason that they will do better for themselves by exploiting the cooperative behavior of their fellows. So, in playing Cooperate with an unknown second party, each member of Protagoras' imagined community risks a sucker's payoff. Moreover, if rationality and a widely distributed shame-justice psychology were adequate, in and of themselves, to get and keep an optimal level of cooperation, there would be no need for anyone to assume the costs of the mutual instruction and corrective punishment that Protagoras highlights. Furthermore, and seemingly fatal to Protagoras business plan, in a perfectly just community there would be no good reason for anyone to pay a sophist his fee for teaching a master's course in political craft.[48]

In light of these concerns it is important to note that in Protagoras' story, the new psychology is not only non-universal, it is a fairly weak disposition: Protagoras' reformed humans are still rationally self-interested and Zeus' gifts of shame and justice are not imagined as doing all the work necessary to create and sustain social order. The base-line moral psychology must, therefore be strengthened by repeated social interactions: Teaching and learning from each other, across the course of our lives, along with the deterrent function of punishment for deviation from the rules, are essential parts of Protagoras' solution to Glaucon's Dilemma.

In the myth, Zeus decrees death for those who lack a sense of shame and justice, but unless and until they are caught and expelled or killed, the community will continue to harbor some individuals whose preference for acting unjustly is undiminished. Some of these unreconstructed egoists may be "mad" and thus will be easily caught out and punished. Others, however, will choose to *pass* as cooperators – as Protagoras points out, it is insane *not* to assume a cooperative public persona. The threat of death or expulsion may be enough to scare a risk-averse crypto-egoist into acting as a consistent cooperator. It will be the role of mutual instruction and punishment to sort out those cooperators who have fallen into the *error* of playing Defect by mistake, from the true egoists who are not making errors but acting on their settled preferences. And to separate egoists who can be deterred from future wrongdoing by credible threats from those who are incorrigible risk-takers and must be removed from the population by expulsion or execution. Protagoras' optimism about the cooperative society suggests that all that can be done, and at a cost that is reasonably borne by the cooperative majority.

Protagoras does not assume that humans are altruistic saints, or that the initial psychological shift would immediately or permanently eliminate the Gyges-type psychology that leads to defection. He supposes, however, that an ordered polis will be composed mostly of cooperators. In game-theoretic terms, he assumes that cooperation will emerge as the most common strategy in a community in which enough share the shame-justice psychology and act accordingly. This cooperative equilibrium will be achieved and sustained by ongoing social interaction. Protagoras' emphasis on mutual instruction over time and on punishment of deviants implies that the cooperation game that sustains social order cannot be one-off; it must be indefinitely repeated. The players in the repeated game are expected to communicate and to learn from each round of play.

Repeated games that randomly pair a large number of players allow for agent-based modeling of populations that evolve over time: potentially into a stable equilibrium.[49] Depending on how agents are motivated (what strategy each plays), how their experience in a given round affects their play in the next round (how they learn), and what each communicates and with which others (how learning spreads), the population evolves in a process known as "replicator dynamics."[50] "Replicators" are entities capable of making copies of themselves. In our case, the replicators are strategies that are relatively successful, in that those who play them receive higher payoffs over the course of repeated play. Replicators multiply at the expense of unsuccessful strategies, and thereby come to dominate the population. So, by transforming a, one-shot game into an indefinitely repeated game, it is possible to predict, based on the setup of the game (motivation and learning, per above), how a population that includes both egoists and conditional cooperators will evolve over time. The question is whether repeated play in a population including egoists (some of whom learn from punishment) and conditional cooperators (willing to punish defection with defection) will result in a population that is all egoists, all cooperators, or some mix of the two types.

As Protagoras' mythic Zeus realized ("poleis cannot exist if only a few people have a share"), if there are only a few cooperators, the population is likely quickly to devolve to all-egoist and the equilibrium will be "scatter and perish." Moreover, if each round of the repeated game is played as a standard PD (as in Glaucon's Dilemma), egoists will consistently win while cooperators consistently lose. In this case the replicator dynamics predicts that the presence of even a few egoists in the population will, over time, drive out all cooperators (Binmore 2007: 123-124 with figure 29) and we will end up back at an “all egoist” uncooperative society. This regression to non-cooperation is what Protagoras describes in the mythic stage preceding Zeus’s gifts, when attempts at social order collapsed in the face of strife.[51]

Protagoras’ story about the origins of order stipulates a condition of imperfect information: given the incentive of egoists to dissimulate their psychology, Player A must always take into account the possibility that Player B will defect. If loss in any one round is fatal (that is, the player does not survive to play in the next round and cannot inform other players of the results of the first round), then we are back at Glaucon’s Dilemma. But if we assume that A can survive a loss, if play of the cooperators is based on “retaliation” (i.e. if Player B plays Defect in round 1, then Player A retaliates with Defect in a subsequent round played against B – regardless of whether A’s motivation is correction or vengeance), and if punishment is meted out often enough (which will depend on payoffs and the number of egoists in the original population), then retaliation will eliminate egoists or lead them to change their play to cooperation. Thought of in spatial terms, repeated play with reliable retaliation play creates a substantial “basin of attraction.” If we start the play of repeated games within that basin (i.e. with the right number of conditional cooperators and the right payoffs), the eventual result will be an all-cooperator population (Binmore 2007: 136-37). Introducing communication (so that A adjusts her strategy against B in round 2, based on the experience of C playing against B in round 1) hastens the rate of evolution toward an equilibrium.

The general point is that a view of human psychology that rejects the assumption of universal pure egoism, a view introduced in Protagoras’ myth through the device of “Zeus’s gifts,” can produce, first, a high-order preference for cooperation as a result of a sense of *shame* that reduces the subjective value of the injustice payoff. Next, in repeated play a sense of *justice* leads ordinary citizens to respond to instances of unjust behavior, as punishers. They ought to do so, according to Protagoras, not out of a “beastly” vengeful impulse, but because it is in someone’s interest. They may regard just punishment of non-cooperators – those who act unjustly –as benefitting the corrigible and/or the society as a whole. An interest in punishment may include the satisfaction of acting out of righteous anger at acts of injustice and its perpetrators.[52] The point is that punishment is reliably meted out, regardless of whether the motive of the punishers is forward-looking correction (as Protagoras urges it ought to be: punishers seek to correct the error of the perpetrator) or retrospective vengeance (the punishers vent their righteous anger upon those who choose to act unjustly). With the behaviors associated with the shame-justice psychology driving the system toward a the high-payoff, mutually cooperative equilibrium, Protagoras’ “Gifts of Zeus” community can come into being despite the initial presence of a few egoists and it is robust to their periodic reemergence. It can retain the individual and social benefits incumbent upon a norm of conditional cooperation based on expectations: a belief that most others are likely to cooperate in turn, and that those who defect will be punished.

2.8. The limits of egoistic self-interest.

With its recognition of the endemic presence of egoists in the population, Protagoras' fable has reintroduced Thrasymachus' and Callicles' formidable few. But Protagoras has shown how an extensive scheme of cooperation can come about despite them and why it need not be vulnerable to them. He has answered the comparative dynamics problem that (so I have claimed) sets up Glaucon's Dilemma: how did social order first emerge in a population of rationally self-interested agent. The solution has turned out to require, first, an adjustment to the starting assumption that ordinary human motivation, in a state of nature, can be reduced to purely egoistical self-interest. And, next, it required repeated play.

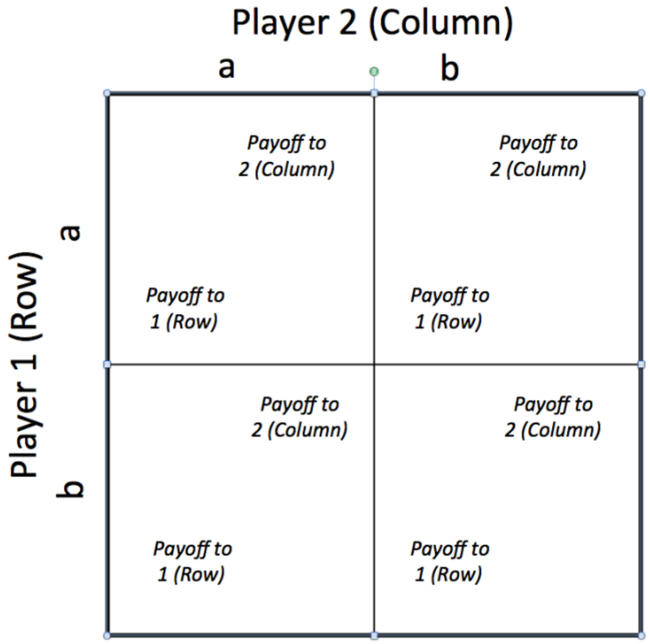
As we saw, Plato's Protagoras ran the origins of social order thought experiment a first time along "Thrasymachean" lines, and concluded that it must end in anarchy and poverty – something akin to Hobbes' state of nature. He ran it a second time with the assumption that self-interest was supplemented by a widely distributed, if relatively weak, moral disposition: shame and justice. That disposition did not transform rascals into saints. It was added on to, rather than simply replacing, self-interest as preference satisfaction and rationality as ordered preferences plus coherent beliefs. Moreover, in the face of incomplete information about the distribution of psychologies in the community, repeated interaction, in the form of mutual instruction and corrective punishment, proved to be essential to the task of translating minimal shared moral intuitions into social order.

Protagoras' imagined "Gifts of Zeus" society is not perfectly just: the tendency of some individuals to self-seeking wrong-doing remains an issue. Some may imitate the attitudes of the just citizen, obeying the rules entirely out of fear of punishment (and of incurring the reputation of being "mad" or "not of the human kind") rather than out of a recognition of the justice of the rules. Severe punishment of incorrigibles, by expulsion or execution, is assumed to be necessary. Protagoras' social order is thus meant to be realistic: emergent, dynamic, and ultimately self-enforcing. As such, it is a long way from Plato's ideal of a fully just, perfectly harmonious, and stable community, as sketched in the later books of the *Republic*. Consideration of the conditions necessary for the existence of that ideal community is Plato's Socrates' ultimate answer to the Glaucon's challenge. And so, it is clear enough why the interlocutors of the *Republic* could not avail themselves of a Protagorean solution to the dilemma that Glaucon's thought experiment poses – and why, after Glaucon poses his challenge, Plato's *Republic* requires eight more books to come to its conclusion.

Our selective survey of Greek thought on the origins of social order has established two main points: First, the classical tradition was fully capable of imagining a population of purely self-interested rational egoists: "unerring craftsmen of self-interest." The problems for social order that arise with that kind of psychological motivation were explored by Greek writers in ways that are reminiscent of some features of contemporary game theory and their proposed solutions can be illustrated by simple games. Next, while it was common among Greek political theorists to suppose that societies could solve cooperation problems by an enlightened recognition of the value of coordination, some Greek thinkers – here represented by Plato's Socrates and his Protagoras – took the challenge of the "Thrasymachean" assumption of human motivation seriously. And certain of them regarded the problem of devising a self-enforcing social order as insoluble under assumptions that reduced human motivation to amoral, self-interested egoism.

This meant, in turn, as in Protagoras' second thought experiment, that the motivation of many (but not necessarily of all) humans must include at least a minimal moral sensibility, capable of taking into account the well-being of others. That assumption recalls the claims of Enlightenment-era moral philosophers, for example, Adam Smith in his *Theory of Moral Sentiments* and *Wealth of Nations*.[53] And, *mutatis mutandis*, it may recall influential work in contemporary political philosophy. John Rawls, for example, posits that the citizens of the presumptively democratic "realistic utopia" that emerges from his contractarian thought experiment have an effective "sense of justice," such that their rational pursuit of self-interest is moderated by a "reasonable" acknowledgement of others' legitimate claims.[54] We need not suppose that the weak shame-justice disposition sketched in Protagoras' "Gifts of Zeus" story provides a thick enough moral sensibility to sustain the kind of just society envisioned by either Smith or Rawls. But the core intuition appears to be similar: Stable human social order rests on motivations that exceed bare egoistic self-interest, but it requires neither a community of saints nor the non-existence of knaves.

Figure 2.1. Basic setup of four-box game form.



Note: Payoffs to Player 1 (row) in this and subsequent 4-box games is in the lower left of each box; payoffs to Player 2 (column) is in the upper right of each box. Figure

Figure 2.2 Pure Cooperation

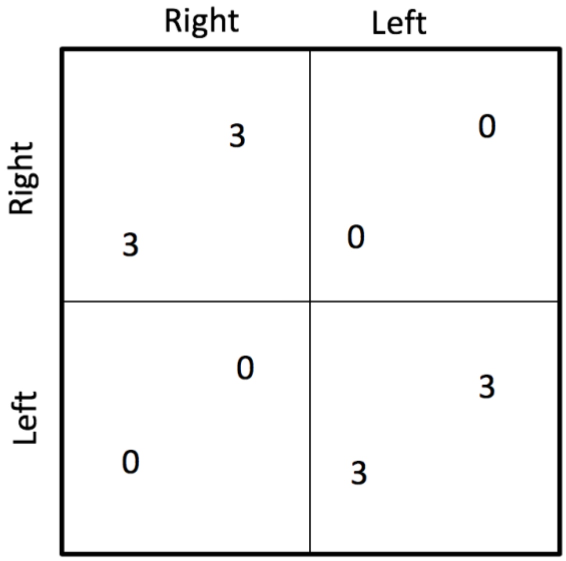


Figure 2.3 Imperfect Cooperation

	Coast	Mountain
Coast	2 3	1 1
Mountain	1 1	3 2

Figure 2.4 Chicken

	Straight	Swerve
Straight	0 0	1 3
Swerve	3 1	2 2

Figure 2.5 Prisoner's Dilemma

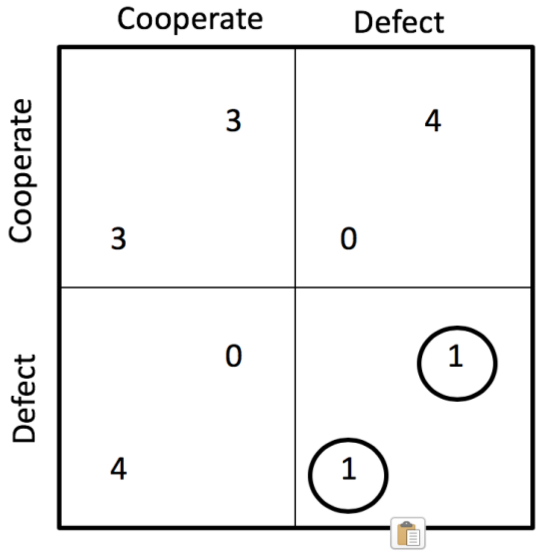


Figure 2.6 Aristotelian Justice.

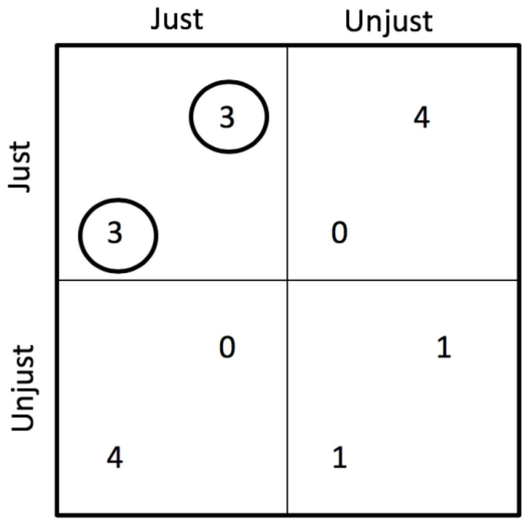


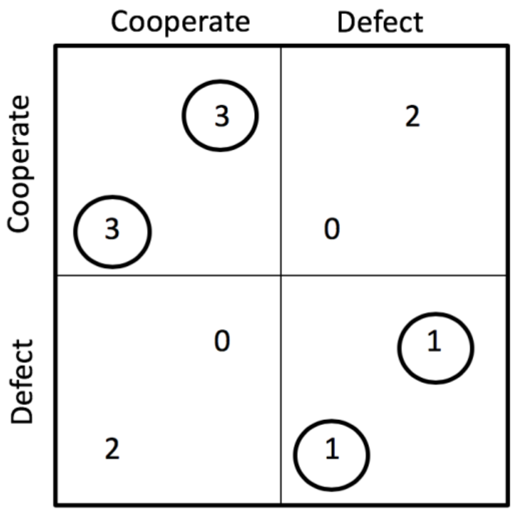
Figure 2.7a. Thrasymachean Injustice - I

		w/o ring	
		Just	Unjust
w ring	Just	3 3	4 0
	Unjust	$\textcircled{0}$ $\textcircled{4}$	1 1

Figure 2.7b Thrasymachean Injustice - II

		w/ ring	
		Just	Unjust
w/o ring	Just	3 3	$\textcircled{4}$ $\textcircled{0}$
	Unjust	0 4	1 1

Figure 2.8. Gifts of Zeus.



	Cooperate	Defect
Cooperate	③ ③	2 0
Defect	0 2	① ①

2. Glaucon. References.

Adam, James and D.A. Rees. 1963. *Plato, The Republic [with critical notes and commentary]*. Cambridge: Cambridge University Press.

Allen, Danielle S. 2000. *The world of Prometheus: Politics of punishing in democratic Athens*. Princeton, N.J.: Princeton University Press.

Amadae, S. M. 2016. *Prisoners of reason: Game theory and neoliberal political economy*. Cambridge: Cambridge University Press.

Anderson, Elizabeth. 2000. "Beyond Homo Economicus: New Developments in Theories of Social Norms." *Philosophy and Public Affairs.* (29):170-200.

Annas, Julia. 1981. *An introduction to Plato's Republic*. Oxford: Clarendon Press.

Austen-Smith, David. 1990. "Information Transmission in Debate." *American Journal of Political Science.* 34:124-152.

Axelrod, Robert M. 1984. *The evolution of cooperation*. New York: Basic Books.

Banting, Keith and Will Kymlicka. 2018. "Theories of Justice, the Strains of Commitment, and Realistic Utopias*"* Presentation. Conference on Political Theory in/and/as Political Science. McGill University: Montreal.

Barney, Rachel. 2017. "Callicles and Thrasymachus." in *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Barry, Brian. 1978. *Sociologists, economists, and democracy*. Chicago: University of Chicago Press.

Binmore, K. G. 2007. *Game theory: A very short introduction.* Vol. Very short introductions ; 173. Oxford and New York: Oxford University Press.

Boehm, Christopher. 2012. *Moral Origins: Social Selection and the Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Bowles, Samuel and Herbert Gintis. 2011. *A cooperative species. Human reciprocity and its evolution.* Princeton: Princeton University Press.

Cairns, Douglas L. 1993. *Aidôs: The psychology and ethics of honour and shame in ancient Greek literature*. Oxford and New York: Clarendon Press.

Calvert, Randall L. 1995. "The rational choice theory of social institutions: cooperation, coordination, and communication." Pp. 216-268 in *Modern Political Economy: Old Topics, New Directions*, edited by Jeffrey S. Banks and Eric Allen Hanuschek. Cambridge: Cambridge University Press.

Choi, Jung-Kyoo and Samuel Bowles. 2007. "The Coevolution of Parochial Altruism and War." *Science.* 318:636-640.

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

Chung, Hun. 2016. "A Game-Theory Solution to the Inconsistency between Thrasymachus and Glaucon in Plato's Republic." *Ethical Perspectives.* 23 (3):383-410.

Cole, Thomas. 1967. *Democritus and the sources of Greek anthropology.* Vol. Philological monographs no. 25. Cleveland: Published for the American Philological Association by the Press of Western Reserve University.

D'Angour, Armand. 2011. *The Greeks and the new: Novelty in ancient Greek imagination and experience*. Cambridge and New York: Cambridge University Press.

Danzig, Gabriel. 2008. "Rhetoric and the Ring: Herodotus and Plato on the Story of Gyges as a Politically Expendient Tale." *Greece & Rome.* 55 (2):169-192.

Davis, Morton D. 1983. *Game theory: A nontechnical introduction*. New York: Basic Books.

Farrar, Cynthia. 1988. *The origins of democratic thinking: The invention of politics in classical Athens*. Cambridge and New York: Cambridge University Press.

Hardin, Russell. 1995. "Self-interest, group identity." Pp. 14-42 in *Nationalism and Rationality,* edited by Albert Breton. Cambridge: Cambridge University Press.

------. 1991. "Acting Together, Contributing Together." *Rationality and Society.* 3:365-380.

Harvey, F. D. 1965. "Two kinds of equality." *Classica et Mediaevalia.* 26:101-146.

Hobbes, Thomas. 1996 [1651]. *Leviathan*. Cambridge: Cambridge University Press.

Joyce, Richard. 2006. *The evolution of morality*. Cambridge, Mass.: MIT Press.

Laird, Andrew. 2001. "Ringing the Changes on Gyges: Philosophy and the Formation of Fiction in Plato's *Republic*." *Journal of Hellenic Studies.* 121:12-29.

Liu, Glory and Barry Weingast. Forthcoming. "Deriving 'General Principles' in Adam Smith: The ubiquity of equilibrium and comparative statics analysis throughout his works." *Adam Smith Review.* 12.

McClennen, Edward F. 2001. "The strategy of cooperation." Pp. 189-208 in *Practical rationality and preference: Essays for David Gauthier*, edited by Christopher W. Morris, Arthur Ripstein, and David P. Gauthier. Cambridge, U.K.; New York: Cambridge University Press.

Morris, Christopher W. and Arthur Ripstein. 2001. "Practical reason and preference." Pp. 1-10 in *Practical rationality and preference : essays for David Gauthier*, edited by Christopher W. Morris, Arthur Ripstein, and David P. Gauthier. Cambridge, U.K.; New York: Cambridge University Press.

Ober, Josiah. 2017. *Demopolis: Democracy before Liberalism in Theory and Practice*. Cambridge: Cambridge University Press.

------. 2017. "Equality, legitimacy, interests, and preferences. Historical notes on Quadratic Voting in a political setting." *Public Choice.* 172 (1-2):223-232.

Ostwald, Martin. 1996. "Shares and Rights: "Citizenship" Greek Style and American Style." Pp. 49-61 in *Dêmokratia*, edited by J. Ober and C.W Hedrick. Princeton: Princeton University Press.

Pettit, Philip. 2014. *Just Freedom: A Moral Compass for a Complex World (Norton Global Ethics)*: W. W. Norton & Company.

Rawls, John. 2001. *Justice as fairness: A restatement*. Cambridge, Mass.: Harvard University Press.

------. 2005. *Political Liberalism (Expanded Edition)*. New York: Columbia University Press.

Reeve, C. D. C. 2008. "Glaucon's Challenge and Thrasymacheanism." *Oxford Studies in Ancient Philosophy.* 34:69-104.

Schelling, Thomas C. 1980. *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.

Segvic, Heda. 2009. *From Protagoras to Aristotle: Essays in ancient moral philosophy*. Edited by Myles Burnyeat. Princeton, NJ: Princeton University Press.

Shields, C. 2006. "Plato's Challenge : The Case Against Justice in Republic II." Pp. 000-

000 in *The Blackwell Guide to Plato's Republic*. Malden, MA: Blackwell.

Skyrms, Brian. 2001. "The Stag Hunt." *Proceedings and Addresses of the American Philosophical Association.* 75 (2):31-41.

Smith, Adam. 1976 [1759]. *The theory of moral sentiments*. Oxford: Clarendon Press.

------. 1981 [1776]. *An Inquiry into the Nature and Causes of the Weath of Nations*. Indianapolis: Liberty Fund.

Taylor, C.C.W. 2007. "*Nomos* and *Phusis* in Democritus and Plato." *Social Philosophy & Policy.* 24:1-20.

Weiss, Roslyn. 2007. "Wise Guys and Smart Alecks in *Republic* 1 and 2." Pp. 000-000 in *The Cambridge Companion to Plato's Republic,* edited by G.R.F. Ferrari.

Wiens, David. 2017. "From Saints to Scoundrels: How Motivations Matter for Ideal Justice (Version 2.2)." *Working Paper (Stanford Political Theory Workshop).*

2. Glaucon. Notes (still only vestigial).

[1] Glaucon’s story concerns “an ancestor of Gyges the Lydian,” Following a convention long-standing among commentators on this passage, I simply call him Gyges. See further chapter 1.

[2] Glaucon’s argument is philosophically distinctive, but its basic premise is common within the classical tradition: a list of passages in which Greek writers assert that self-interest is a (if not the) dominant motivator (for all humans, or most individuals, or powerful persons, or states) would be very long. It would prominent include, e.g., Pseudo-Xenophon, *Athenaion Politeia*, Thrasymachus in Plato, *Republic* book 1, Thucydides’ Athenians in Book 1 (ambassadors at Sparta) and 5 (Melian Dialogue), Aristotle *Politics* (on the interested “parts” in the commonly existing corrupted regimes, Demosthenes 21 *Against Meidias*, as well as many passages in lyric and iambic poetry, tragedy, comedy, satire.

[3] Literature on the challenge offered by Glaucon and Adeimantus and the philosophical purposes to which the story of Gyges is put by Plato: Chapter 1 note XX.

[4] *Nomos-phusis*: Taylor 2007 with literature cited.

[5] πεφυκέναι γὰρ δή φασιν τὸ μὲν ἀδικεῖν ἀγαθόν, τὸ δὲ ἀδικεῖσθαι κακόν, πλέονι δὲ κακῷ ὑπερβάλλειν τὸ ἀδικεῖσθαι ἢ ἀγαθῷ τὸ ἀδικεῖν, ὥστ΄ ἐπειδὰν ἀλλήλους ἀδικῶσί τε καὶ ἀδικῶνται καὶ ἀμφοτέρων γεύωνται, τοῖς μὴ δυναμένοις τὸ [359α] μὲν ἐκφεύγειν τὸ δὲ αἱρεῖν δοκεῖ λυσιτελεῖν συνθέσθαι ἀλλήλοις μήτ΄ ἀδικεῖν μήτ΄ ἀδικεῖσθαι: καὶ ἐντεῦθεν δὴ ἄρξασθαι νόμους τίθεσθαι καὶ συνθήκας αὐτῶν, καὶ ὀνομάσαι τὸ ὑπὸ τοῦ νόμου ἐπίταγμα νόμιμόν τε καὶ δίκαιον: καὶ εἶναι δὴ ταύτην γένεσίν τε καὶ οὐσίαν δικαιοσύνης, μεταξὺ οὖσαν τοῦ μὲν ἀρίστου ὄντος, ἐὰν ἀδικῶν μὴ διδῷ δίκην, τοῦ δὲ κακίστου, ἐὰν ἀδικούμενος τιμωρεῖσθαι ἀδύνατος ᾖ.

[6] The passage below is cited by Barry 1989: 6 as an example of a theory of a contractual theory of justice that “continues to be a live option, and is one of the two theories around which” his own book is constructed. Per the Introduction, Barry sees the theory sketched by Glaucon as an early version of the kind of line of thought later developed by Hobbes, Hume, and contemporary game theory.

[7] ὡς δὲ καὶ οἱ ἐπιτηδεύοντες ἀδυναμίᾳ τοῦ ἀδικεῖν ἄκοντες αὐτὸ ἐπιτηδεύουσι, μάλιστ΄ ἂν αἰσθοίμεθα, εἰ τοιόνδε ποιήσαιμεν τῇ διανοίᾳ: δόντες ἐξουσίαν ἑκατέρῳ ποιεῖν ὅτι ἂν βούληται, τῷ τε δικαίῳ καὶ τῷ ἀδίκῳ, εἶτ΄ ἐπακολουθήσαιμεν θεώμενοι ποῖ ἡ ἐπιθυμία ἑκάτερον ἄξει. ἐπ΄ αὐτοφώρῳ οὖν λάβοιμεν ἂν τὸν δίκαιον τῷ ἀδίκῳ εἰς ταὐτὸν ἰόντα διὰ τὴν πλεονεξίαν, ὃ πᾶσα φύσις διώκειν πέφυκεν ὡς ἀγαθόν.

[8] On the road/path metaphor in Plato and Herodotus and in game theory, see chapter 1, note XX.

[9] Chung 2016 seeks to square Thrasymachus' and Glaucon's accounts of justice with reference to simple games, focusing, as I do below, on the Prisoner's Dilemma. He argues that there are two solutions to resolving inconsistencies between the two: a democracy in which the many weak are the rulers (thus by definition strong), and a dictatorship in which the strong appropriate all social surplus, leaving the weak with only protection as an improvement in their pre-political, state of nature condition. Chung does not draw attention to the co-presence of the "two men" at the end of the path of injustice or on the abstraction of ordinary to "perfect craftsmen" of self-interest. I find Chung's conclusions problematic for three reasons: He takes Thrasymachus' two definitions of justice as the advantage of the strong and the good of the another seriously, rather than seeing them as examples of sophistic reversal (contrasting "so-called justice – the good of another" with "true justice – the advantage of the strong"). He fails to explain what I am calling the comparative dynamics problem, e.g. how the weak many *become* the ruler. He cannot account for how Glaucon's challenge sets up the solution of the *Republic* – which is neither a democracy nor a surplus-expropriating dictatorship. Anderson 2000 is critical of all rational-choice based attempts to explain cooperative behavior, on the grounds that that they fail to explain the actual scope of observed cooperation and norm-following. As we will see in the course of this book, the Greek tradition is very attentive to the incompleteness of rational explanations for social behavior. But I will argue that the Greek thinkers, like contemporary choice theorists, recognized instrumental rationality as an essential foundation on which to build richer and more descriptively satisfactory accounts of behavior under constraint.

[10] Two person games of this sort may be visualized as an "extensive form" game tree. Plato *Meno* on 4-box matrix illustration.

[11] Full Coordination: xx

[12] Coordination in game theory: Hardin 1991, 1995; Calvert 1995.

[13] Chicken: xx. Cf. chariot race in Iliad book 23.

[14] For a detailed, and critical, account of the Prisoner's Dilemma game, and its role in contemporary social theory, see Amadae 2016.

[15] Likewise, Thrasymachus' account of the behavior of the craftsman of self-interest emphasizes the goods that he seizes, through stealth or force, at the expense of others: 343c-344c.

[16] Aristotle on equality, virtue, justice: *Nicomachean Ethics, Politics*, book 1, 4 with Harvey 1962; Ober 2017.

[17] Note that in the Aristotelian Justice game, I am assuming that the payoffs are objectively measured (e.g. by time spent in prison) rather than, as is usual in PD-type games, measuring the "full information" subjective preference of each player. In order to better reflect full information subjective preferences, this game could be set up with the upper left box payoffs at 5,5, to reflect the added utility that each player gains from the performance of virtuous acts. This results in a "Stag Hunt" type game illustrated as the Gifts of Zeus game in Figure 2.8.

[18] Ancient references to Thrasymachus and his rhetorical art are collected in Laks and Most (2016), vol 8 . sec. 35.

[19] Cheap talk (with special reference to legislative debate): Austen-Smith 1990.

[20] Folk Theorem and multiple equilibria for repeated games: Binmore xx.

[21] If we were to take the reference point as Defect, Defect, we would need to write the payoffs differently in order to accommodate the stipulation that the harm in suffering injustice is a greater than the benefit of doing it. For example, switching from ordinal to cardinal weighting, we might specify that Defect, Cooperate pays off 4, -3; Cooperate, Defect -3/4. This has the effect of lowering the "social payoff" for these strategies from 4 (4,0) to 1, and thus below Defect, Defect (2). If this is a repeated game, per below, time, the Thrasymachean tyrant ends up ruling a community poorer than the "state of nature" default.

[22] This solution is in some ways similar to that of McClennen (2001), who argues for an alternative view of rationality that dispenses with the requirement of equilibrium play in favor of Pareto optimization: McClennen's point is that truly rational individuals will prefer the highest social payoff, turning the PD into a coordination problem. He suggests that this offers a much more obvious explanation for the emergence of social order, which is the concern of the texts in this chapter.

[23] Chung 2015 models Hobbes' state of nature (and the escape from it) as a Bayesian game in which each player has only incomplete information about the" type" of the other players. See, below, on the implicit game set up in Protagoras' Great Myth.

[24] τοὺς ἐρρωμενεστέρους τῶν ἀνθρώπων καὶ δυνατοὺς ὄντας πλέον ἔχειν.

[25] Callicles in the *Gorgias*: Taylor 2007: 8-11, 16; Barney 2017 (with comparison to Glaucon in *Republic* book 2).

[26] οἶμαι οἱ τιθέμενοι τοὺς νόμους οἱ ἀσθενεῖς ἄνθρωποί εἰσιν καὶ οἱ πολλοί.

[27] πρὸς αὑτοὺς οὖν καὶ τὸ αὑτοῖς συμφέρον τούς τε νόμους τίθενται. This is similar to the point made by the Anonymous Iamblichi F 6, in reference to the

vulnerability of an imaginary willfully unjust superman to the collective action of many weaker individuals motivated by a preference for the behavior-regulating conditions of justice.

[28] ἀγαπῶσι γὰρ οἶμαι αὐτοὶ ἂν τὸ ἴσον ἔχωσιν φαυλότεροι ὄντες.

[29] ὥσπερ λέοντας, κατεπᾴδοντές τε καὶ γοητεύοντες καταδουλούμεθα.

[30] ἐὰν δέ γε οἶμαι φύσιν ἱκανὴν γένηται ἔχων ἀνήρ, πάντα ταῦτα ἀποσεισάμενος καὶ διαρρήξας καὶ διαφυγών, καταπατήσας τὰ ἡμέτερα γράμματα καὶ μαγγανεύματα καὶ ἐπῳδὰς καὶ νόμους τοὺς παρὰ φύσιν ἅπαντας, ἐπαναστὰς ἀνεφάνη δεσπότης ἡμέτερος ὁ δοῦλος, καὶ ἐνταῦθα ἐξέλαμψεν τὸ τῆς φύσεως δίκαιον.

[31] This is precisely the argument of Ps-Xenophon, *Ath. Pol*. in which the anonymous author stresses that poor Athenians acted rationally (although wrongly) in their own collective interest by choosing to rule (as a *demos*) rather than to be enslaved by the elite (in an order he describes as *eunomia*).

[32] On fallacious and non-fallacious forms of functionalism, see Barry 1978: 168-73.

[33] Barney (2017) notes that, unlike Callicles (and Thrasymachus), Glaucon does not lean on a putative division of human society into the naturally weak and strong, and that it leaves open the question: "given the merely conventional character of justice and the constraints it places on our pleonectic nature, why should any one of us be just, whenever injustice would be to our advantage?" But she does not pursue the implications of this question for the origins of law and social contract. Chung 2016 asks a related, but different, question: how can Thrasymachus' initial argument for "the interests of the strong" be squared with Glaucon's apparent conclusion that justice is in the interest of the weak?

[34] On the concept of progress in Greek thought, see further D'Angour 2011.

[35] Absent technology: Plato, *Protagoras* 321c-d, *Statesman* 1274b-d; *Laws* 677d ff.

[36] Predatory animals: Plato, *Protagoras* 322b, *Statesman* 1274c; *Laws* 681a; Polybius 6.1.7; Diodorus Siculus 1.8.2, 1.15.5. Porphyry xx.

[37] Powerful individuals, pirates: Thucydides 1.5-8; Plato, *Gorgias* (see discussion above); Diodorus Siculus 1.90; Porphyry *De abstinentia* = Cole 1977: 71-72.

[38] Larger scale cooperation: Plato *Gorgias* 483b-c; Plato *Laws* 680e-681b; Aristotle, *Politics* 1252b24 ff.; Diodorus Siculus 1.8.

[39] ἀλλήλοις βοηθεῖν ὑπὸ τοῦ συμφέροντος διδασκομένους, ἀθροιζομένους δὲ διὰ τὸν φόβον ἐπιγινώσκειν ἐκ τοῦ κατὰ μικρὸν τοὺς ἀλλήλων τύπους

40 συναγομένων γὰρ ἐν ἀρχῇ τῶν ἀνθρώπων ἐκ τοῦ θηριώδους βίου, τὸ μὲν πρῶτον ἀλλήλους κατεσθίειν καὶ πολεμεῖν, ἀεὶ τοῦ πλέον δυναμένου τὸν ἀσθενέστερον κατισχύοντος: μετὰ δὲ ταῦτα τοὺς τῇ ῥώμῃ λειπομένους ὑπὸ τοῦ συμφέροντος διδαχθέντας ἀθροίζεσθαι καὶ ποιῆσαι σημεῖον ἑαυτοῖς ἐκ τῶν ὕστερον καθιερωθέντων ζῴων: πρὸς δὲ τοῦτο τὸ σημεῖον τῶν ἀεὶ δεδιότων συντρεχόντων, οὐκ εὐκαταφρόνητον τοῖς ἐπιτιθεμένοις γίνεσθαι τὸ σύστημα: [2] τὸ δ᾽ αὐτὸ καὶ τῶν ἄλλων ποιούντων διαστῆναι μὲν τὰ πλήθη κατὰ συστήματα

41 Farrar 1988: ch. 3 offers a thoughtful and sympathetic treatment of Protagoras, and emphasizes that Plato's Protagoras (a character she calls "Platagoras") must not be mistaken for the actual Sophist. Segvic 2009, ch. 1 (pp. 3-27) analyzes the views of Plato's Protagoras, in respect to democracy, power, competence, and the good life, concluding that Protagoras' rival conception of moral learning, political and civic virtue recurs in Plato's *Republic* and that Aristotle's views were shaped by the dispute between Plato and the Sophists, especially Protagoras.

42 οὕτω δὴ παρεσκευασμένοι κατ᾽ ἀρχὰς ἄνθρωποι ᾤκουν σποράδην, πόλεις δὲ οὐκ ἦσαν: ἀπώλλυντο οὖν ὑπὸ τῶν θηρίων διὰ τὸ πανταχῇ αὐτῶν ἀσθενέστεροι εἶναι, καὶ ἡ δημιουργικὴ τέχνη αὐτοῖς πρὸς μὲν τροφὴν ἱκανὴ βοηθὸς ἦν, πρὸς δὲ τὸν τῶν θηρίων πόλεμον ἐνδεής —πολιτικὴν γὰρ τέχνην οὔπω εἶχον, ἧς μέρος πολεμική— ἐζήτουν δὴ ἁθροίζεσθαι καὶ σῴζεσθαι κτίζοντες πόλεις: ὅτ᾽ οὖν ἁθροισθεῖεν, ἠδίκουν ἀλλήλους ἅτε οὐκ ἔχοντες τὴν πολιτικὴν τέχνην, ὥστε πάλιν σκεδαννύμενοι διεφθείροντο.

43 Ζεὺς οὖν δείσας περὶ τῷ γένει ἡμῶν μὴ ἀπόλοιτο πᾶν, Ἑρμῆν πέμπει ἄγοντα εἰς ἀνθρώπους αἰδῶ τε καὶ δίκην, ἵν᾽ εἶεν πόλεων κόσμοι τε καὶ δεσμοὶ φιλίας συναγωγοί. ἐρωτᾷ οὖν Ἑρμῆς Δία τίνα οὖν τρόπον δοίη δίκην καὶ αἰδῶ ἀνθρώποις: 'πότερον ὡς αἱ τέχναι νενέμηνται, οὕτω καὶ ταύτας νείμω; νενέμηνται δὲ ὧδε: εἷς ἔχων ἰατρικὴν πολλοῖς ἱκανὸς ἰδιώταις, καὶ οἱ ἄλλοι δημιουργοί: καὶ δίκην δὴ καὶ αἰδῶ οὕτω θῶ ἐν τοῖς ἀνθρώποις, ἢ ἐπὶ πάντας νείμω;' 'ἐπὶ πάντας,' ἔφη ὁ Ζεύς, 'καὶ πάντες μετεχόντων: οὐ γὰρ ἂν γένοιντο πόλεις, εἰ ὀλίγοι αὐτῶν μετέχοιεν ὥσπερ ἄλλων τεχνῶν: καὶ νόμον γε θὲς παρ᾽ ἐμοῦ τὸν μὴ δυνάμενον αἰδοῦς καὶ δίκης μετέχειν κτείνειν ὡς νόσον πόλεως.'

44 ὃ ἐκεῖ σωφροσύνην ἡγοῦντο εἶναι, τἀληθῆ λέγειν, ἐνταῦθα μανίαν, καί φασιν πάντας δεῖν φάναι εἶναι δικαίους, ἐάντε ὦσιν ἐάντε μή, ἢ μαίνεσθαι τὸν μὴ προσποιούμενον δικαιοσύνην: ὡς ἀναγκαῖον οὐδένα ὅντιν᾽ οὐχὶ ἁμῶς γέ πως μετέχειν αὐτῆς, ἢ μὴ εἶναι ἐν ἀνθρώποις. On Protagoras' "shame" and its relationship to his "justice" see Cairns 1993: xx; Segvic 2009: 10-11;

45 λυσιτελεῖ γὰρ οἶμαι ἡμῖν ἡ ἀλλήλων δικαιοσύνη καὶ ἀρετή : διὰ ταῦτα πᾶς παντὶ προθύμως λέγει καὶ διδάσκει καὶ τὰ δίκαια καὶ τὰ νόμιμα.

[46] Stag Hunt: Skyrms 2001, with explicit contrast to the Prisoner's Dilemma game.

[47] This is in some ways similar to the constrained maximization" or "resolute choice" in the amended choice theory of David Gauthier; see the succinct account of Gauthier's theory in Morris and Ripstein 2001.

[48] Wiens 2017 fruitfully discusses the problem of core motivation for ideal theory: if all are fully committed to justice, there is no need for coercive institutions, and so political theory is over (solved by a shared ethical standard) before it gets going.

[49] Repeated games: Axelrod 1984, 1987; Calvert 1995; Anderson 2000: 178-181.

[50] This approach is adopted by some game theorists concerned with explaining the evolution of stable strategies of cooperation, and ultimately the emergence of morality: See, for example, Choi and Bowles 2007; Joyce 2006; Bowles and Gintis 2011; Boehm 2012 (among many others).

[51] Chung 2015 helpfully models Hobbes' *Leviathan* state of nature as a static (rather than dynamic) not as a standard Prisoner's Dilemma, but a game in which the population includes both "modest" (non-egoistic) types and violent (egoistic types). Each player is uncertain about the other player's "type" (modest or violent) and highly values his own life. The payoff to cooperating against a defector is "death." Chung's conclusion that, under these conditions, "even when the vast majority… are peace-loving… universal war could still break out," is surely correct. It confirms Hobbes' requirement of a sovereign (whether sole dictator or majority tyrant) as the unique solution. But if the game with which we are concerned is, as Protagoras' appears to be, indefinitely repeated, and if the cost of losing a round is less than fatal, the emergence of a non-Hobbesian alternative of a self-enforcing equilibrium without a lawless sovereign remains available.

[52] Allen 2000 on the role of anger in the politics of punishment in ancient Athens.

[53] Smith 1976 [1759]; 1881 [1776]. On Smith as a normative and positive theorist, whose work on moral psychology is fully compatible with his work on political economy, and readily understood in game theoretic terms, see Liu and Weingast, forthcoming.

[54] Rawls 2001:4, 2005: lx; with discussion of Banting and Kymlicka (2018). I am indebted to Jackie Basu for enlightening discussion of the relationship between the rational and the reasonable in Rawls' political philosophy.