2-7-2014

Time Wounds All Heels: Human Nature and the Rationality of Just Behavior

Timothy Glenn Slattery
University of South Florida, florida_yenny@hotmail.com

Time Wounds All Heels: Human Nature and the Rationality of Just Behavior

by

Timothy G. Slattery

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Philosophy
College of Arts and Sciences
University of South Florida

Major Professor: Douglas Jesseph, Ph.D.
Colin Heydt, Ph.D.
Hugh LaFollette, Ph.D.
Brook Sadler, Ph.D.
Joanne Waugh, Ph.D.

Date of Approval:
February 7, 2014

Keywords: behavioral economics, contractarian, fool, Gauthier, justice

ACKNOWLEDGMENTS

I want to begin by expressing my profound thanks to Professor Doug Jesseph, for all of his valuable insight and for the hours of conversation that allowed me to shape many of the ideas in this dissertation. His interest in the project allowed me to remain motivated during those times when I felt like my progress was stalling. I am grateful also for the thorough and insightful commentary of Professors Colin Heydt and Hugh LaFollette. Their generosity with their time and their knowledge has been invaluable. Professor Joanne Waugh has been an indispensable advisor, not only with her expertise in Ancient Greek philosophy, but also in guiding me through the transition into candidacy. I am especially grateful to Professor Brook Sadler, who has been both an inspiration and a mentor to me at every step of this long journey.

Thank you to the entire faculty and staff of the University of South Florida Philosophy Department. Everything I know about philosophy I learned from one of you.

Thank you to my parents, Francis and Joan Slattery, for providing me with the best possible education, for always being there for me, and for teaching me to be an unapologetic skeptic.

Thank you to Fred MacLean, Jr. and all of my partners and colleagues at Heritage Investment Group for your patience and understanding with respect to this project. I know it has been disruptive at times, but none of you ever had a bad word to say about it. I am eternally grateful.

Thank you to my two wonderful children, Adriana and Declan. The most painful sacrifice that I had to endure in order to produce this dissertation was that I was not able to be

there for some of the small, but very special, moments in your young lives. You are my inspiration, and I hope that someday you will come to love the study of philosophy as I do, and to appreciate the value of knowledge for its own sake. I pledge to do my best to instill this in you.

Finally, and most important, I want to thank my wife, Jenny. Few people possess the patience and understanding that you have demonstrated as I regularly disappeared over the course of many years in pursuit of this goal. You have made my life very special, I love you, and I am deeply indebted to you.

TABLE OF CONTENTS

ABSTRACT

We share our world with many people who ignore the principles of justice and who regularly take advantage of others by breaching trust or breaking agreements. This dissertation is about the irrationality of the actions of these covenant-breakers. A covenant-breaker typically believes that unjust behavior is to his advantage and that only a fool would act in any other way. Would it not be disturbing if this were true?

My central claim will be that adherence to the precepts of justice is a rational strategy for a self-interested actor. I intend to demonstrate that con men and covenant-breakers do not act rationally when violating an agreement. I will trace the concept of justice as it evolves through philosophical history and show that, while the concept of justice changes as the underlying concept of human nature and psychology changes, the argument in favor of the rationality of just behavior remains coherent throughout. Each historical interpretation will advance some form of the claim that the consistent observance of cooperative agreements is a rational strategy, and at each point an interlocutor will object. I will show that these interlocutors are mistaken.

My motivating goal is to show that justice, understood as the consistent observance of cooperative agreements, is rational. I want to respond to the clandestine cheaters and other skeptics who believe that just behavior is for suckers, because, if the skeptics are right, and justice is indeed irrational, then those among us who are acting in a just manner are paying an unnecessary cost.

INTRODUCTION

On April 29, 1938, a baby boy was born to young working-class parents in Queens, New York. The boy grew into a bright young man and graduated from Hofstra University. He briefly attended law school, but dropped out in 1960 to found his own Wall Street investment firm, which he financed with money he had saved from working as a lifeguard and lawn sprinkler installer. His firm was a pioneer in the use of information technology in the trading of securities, and it quickly grew to become one of the largest market makers on Wall Street.

In the 1970s, his firm opened a wealth management division, which also enjoyed tremendous success. He was described as a "master marketer," and he generated most of the firm's clients himself through relationships he cultivated at exclusive country clubs in New York and Palm Beach. His investment returns were remarkably consistent in both up and down markets, and by 2008 his wealth management division had grown to $17 billion in assets under management. His clients adored him and they routinely referred their friends and family members to his firm, hoping they would also be granted the privilege of having him manage their money.

Then, on December 11, 2008, the man from Queens was arrested for securities fraud. In the ensuing days it became clear that this man, Bernard Madoff, had been operating the largest financial scam in the history of the United States. The consistent returns he had been reporting to his clients were completely fabricated, and the complaint filed by U.S. Attorney for the Southern District of New York on March 10, 2009 alleged that Madoff had defrauded his clients of nearly $65 billion in assets. As of the time of this writing, Madoff is serving a 150-year

prison sentence, and an ongoing effort to recover his clients' assets has so far yielded less than $12 billion.

We share our world with many people who ignore the principles of justice and who regularly take advantage of others by breaching trust or breaking agreements. Bernie Madoff is the most famous case, mostly because he got caught. But we all see less dramatic examples every day: A politician reneges on promises she made to her constituents, confident in her ability to survive the backlash and win re-election. A businessman makes a comfortable profit by financing a project with borrowed money and defaulting on the loan, secure in the protection of bankruptcy laws. A married man has a one-night affair with another woman while on a business trip, certain that his wife will never know. A patron at a restaurant notices that the waiter forgot to charge her for the second round of drinks, yet she pays the bill as if it were accurate. "That is their problem," she thinks to herself.

This dissertation is about the irrationality of the actions of these covenant-breakers. It addresses con men, cheaters, dishonest restaurant patrons, and anyone who has ever taken advantage of others by breaking an agreement. To varying degrees, it addresses all of us. When an individual elects to behave in an unjust manner by violating an agreement, she typically does so with the belief that she is acting rationally and in her own best interest. What if the covenant-breakers are right? Is justice a farce? Are those among us who consistently behave in a just manner merely suckers to be taken advantage of by more clever individuals who recognize that justice is not in one's own best interest? A covenant-breaker, whether he is Bernie Madoff or some less-harmful agent, believes that unjust behavior is to his advantage and that only a fool would act in any other way. Would it not be disturbing if this were true?

The central claim of this dissertation will be that adherence to the precepts of justice is a rational strategy for a self-interested actor. That is, I intend to demonstrate that con men and covenant-breakers do not act rationally when violating an agreement. I will trace the concept of justice as it evolves through philosophical history and show that, while the concept of justice changes as the underlying concept of human nature and psychology changes, the argument in favor of the rationality of just behavior remains coherent throughout. Each historical interpretation will advance some form of the claim that the consistent observance of cooperative agreements is a rational strategy, and at each point an interlocutor will object. I will show that these interlocutors are mistaken.

My motivating goal is to show that justice, understood as the consistent observance of cooperative agreements, is rational despite its costs. I want to respond to the clandestine cheaters and other skeptics who believe that just behavior is for suckers, because, if the skeptics are right, and justice is indeed irrational, then those among us who are acting in a just manner are paying an unnecessary cost. Those of us who advocate justice have agreed to constrain our behavior in the belief that to do so is in our own self-interest; we believe the cost of this constraint is outweighed by the benefits of just cooperation. If it turns out that clandestine cheating is a superior strategy, then justice is not in our self-interest, justice has negative instrumental value, and those among us who are observing the precepts of justice are acting irrationally from an instrumental standpoint. In some respects, this is a battle between our moral intuitions and the claims of a moral skeptic.[1] I want to lend credence to our intuitions by showing that following

[1] Haidt might claim that all I am doing here is attempting to find a rational justification for an ethical intuition that I, and most other individuals, already have. I address Haidt's ideas briefly in Ch2 and Ch 5. See Haidt, Jonathan. "The Intuitive Dog and Its Rational Tail." in *The Righteous Mind,* 27-51. New York: Pantheon, 2012.

these intuitions is indeed in our best interests, and that the Madoffs of the world are acting irrationally.

In Chapter 1, I introduce the philosophical debate concerning justice as it appears in Plato's *The Republic*. The original moral skeptic, Thrasymachus, denounces justice and Socrates offers a spirited defense. While Socrates' argument is ultimately unsatisfying, he introduces three themes related to justice that will reappear throughout the dissertation: He claims that justice is a non-zero-sum game, he illustrates the distinction between the instrumental and intrinsic value of justice, and he shows that any concept of justice must be based upon an underlying concept of human nature.

Chapter 2 addresses justice as it is interpreted in the philosophy of Hobbes and Hume. They advance many of the same themes as Socrates, and they each provide a more satisfying account of human nature than the one given in *The Republic*. Since my goal is to show that justice, understood as the consistent observance of cooperative agreements, is rational despite its costs, I conclude that Hobbes's contractarian account of justice provides the better framework for the advancement of this thesis.

In Chapters 3 and 4, I present the core of my argument with assistance from Gauthier and several other contemporary contractarian philosophers. Gauthier's brilliant insight is that, when an individual adopts a strategy of just behavior, she is operating on the level of metachoice. That is, *she is making a choice about how to make choices*. Additionally, Gauthier claims that we adopt certain "deliberative procedures" that define how we have chosen to make these choices.

Yet the moral skeptic remains unconvinced. The skeptic, most notably in the form of Hobbes's Fool, claims that it is in his best interest not to adopt a deliberative procedure in accordance with justice, but instead to opt for the opportunistic violation of covenants. It is at

this point that I introduce my most significant contribution to the debate, which I refer to as the imperfect reason argument. I will demonstrate that the Fool is indeed foolish, as he is almost certainly overestimating his own ability to determine in advance which violations of which covenants will be to his advantage. I will use behavioral economics to expose some noteworthy aspects of human nature and their corresponding effects on the rationality of justice. I will show that we humans tend to reason in a flawed manner, and that this flaw in our reasoning ability ultimately leads to the conclusion that a self-interested individual is best-served by adopting a policy of just behavior.

Finally, Chapter 5 attempts to finish the project that Socrates started. I consider whether or not justice, in addition to having instrumental value, has intrinsic value as well. Our moral intuitions encourage us to believe that there are "loftier" reasons for just behavior above and beyond the avoidance of the downside of cheating, yet it remains to be seen whether these benefits are intrinsic to justice or if they are just another form of instrumental benefit.

Before moving on, a brief aside regarding the use of terms is in order. The term "justice" is used in many different contexts throughout the historical philosophical discourse, and it carries with it a wide variety of connotations. Hobbes defines justice as the performance of covenants, and for Gauthier justice is the rational disposition to agree to forego the opportunity for free ridership or parasitism in return for others foregoing the same. In keeping with this contractarian tradition, I will be using the term "justice" throughout this dissertation in the sense of "the consistent observance of cooperative agreements." I will assume non-coercion and approximate equality (in the Hobbesean sense) between the cooperating parties.

At various points along the way, the reader may have a spontaneous and negative reaction to what I have *not* said about justice. I ask the reader to keep in mind that the topic

under consideration involves a very narrow definition of justice. I am not using the term "justice" to cover the whole of morality. I am not making claims about justice in any political context, nor am I addressing the Rawlsian justice of social institutions. I have omitted these aspects of the wider definition of justice not because they are unimportant, but because each of them could occupy a separate dissertation in its own right. In the Conclusion section of the dissertation, I will briefly address the justice of social institutions and justice in situations of unequal power, but until then I ask the reader to keep in mind that "justice" will refer specifically to the consistent observance of cooperative agreements.

CHAPTER 1: JUSTICE AND RATIONALITY IN ANCIENT GREECE

The enquiry into the rationality of justice begins in Athens in the opening pages of Plato's *The Republic*. Socrates and various interlocutors attempt to define justice, and Socrates is challenged to defend the position that a just life is superior to an unjust life. He claims that justice is desirable not only because of the instrumentally valuable consequences of just behavior, but also because possessing a just soul provides us with intrinsic benefits as well. Simply put, Socrates proposes that we have a motivation to be just.

I will not attempt to definitively resolve the dispute between Socrates and his interlocutors, as this would occupy a separate book in itself. I will, however, use arguments suggested by *The Republic* to frame the ongoing debate regarding the rationality and intrinsic value of justice. Before embarking on a study of the rationality of just behavior, it is necessary to understand how the particular idea of justice that is being considered relates to our underlying human nature. I will therefore begin this chapter with a description of human nature and a corresponding definition of justice drawn from *The Republic*. I will then turn to an analysis of Socrates' argument in favor of the rationality of justice, first as a refutation of the egoistic argument of Thrasymachus and second as a response to the challenge of Glaucon and Adeimantus regarding the instrumental versus intrinsic value of justice. Next, I will investigate the relationship between Socrates' definition of the just soul and the rationality of just acts. I will show that Socrates' failure to emphasize the importance of just acts leads to an inadequate account of human psychology and represents a major weakness in his argument in favor of the rationality of justice. In the concluding section, I hope to show that, while Socrates'

characterization of justice in *The Republic* leaves much to be desired, it does introduce several concepts which are helpful in supporting the argument that just behavior is rationally required.

It is important to note at the outset that I will not assume that the Platonic character Socrates speaks for Plato or that the doctrines that are often attributed to Plato by scholars actually represent Plato's own views. Plato explicitly chooses to remain anonymous in his dialogues and we must therefore consider the very real possibility that he chose to write dialogues for philosophical reasons and that his own personal views may have been quite different from those espoused by the character Socrates. I will therefore attribute the claims made by Socrates to Socrates and not to Plato, and when addressing scholarly commentary on Plato, I will be referring to a specific scholar's interpretation of Plato's dialogues, always bearing in mind that the doctrines being addressed cannot be definitively attributed to Plato himself.

Socrates' Concept of Human Nature and the Definition of Justice

The ongoing debate over the rationality of justice in the Western philosophical tradition is rooted in Socrates' description of human nature and psychology in *The Republic*.[2] For Socrates, the primary driving force behind human motivation in a properly adjusted soul is reason. Reason is what separates us from lower animals; it is the dominant trait in the best individuals among us and it allows us access to knowledge of the Forms.

In addition to being driven by reason, Socrates claims that human beings are also driven by certain needs. We come together to form cities because we are not self-sufficient; we need help from each other in the form of protection, economic sustenance and mutual aid, and we can

[2] Plato. "The Republic." in *Plato, Complete Works,* edited by John M. Cooper, translated by G.M.A. Grube and C.D.C. Reeve. Indianapolis: Hackett Publishing Company, 1997.

only realize our full nature in the context of a community of other humans. That is, humans have an emotional and physical need for contact with other humans and we are, by nature, social animals.

Given our nature as rational, needy creatures, how should we understand the workings of human psychology as it relates to justice? For Socrates, human psychology is best described through an analysis of the proper structure of the human soul. Rather than address the structure of the soul directly, he famously begins with the use of an analogy in which he will claim that the defining characteristics of a just city are also the defining characteristics of a just soul.

He begins with the "principle of specialization."[3] As individuals come together to form societies and cities, reason informs them that each individual possesses unique natural talents and abilities. Some individuals are talented farmers, some are good at commerce, others are built for combat and still others have intellectual gifts that will make them talented philosophers. It soon becomes obvious that the needs of everyone in the community are best-served when each individual performs the function for which he is best-suited by nature.[4] This principle of specialization will dictate that each individual in the city will focus on certain tasks, which will in turn benefit the overall flourishing of the city as a whole and the individuals within it.

Specifically, the populace will be divided into craftsmen, who tend to be dominated by desire and who will conduct the business of the city, auxiliaries, who possess a high degree of courage and spirit and who will be the defenders of the city, and guardians, who, being educated in reason and philosophy, will be the rulers of the city. A just city is one in which the four cardinal virtues (wisdom, courage, temperance and justice) are fostered via the proper interaction of these three types of citizens. The rulers exercise wisdom, the guardians exercise courage,

[3] "The Republic," 370a-c, p. 1009
[4] It is important to note here that Socrates finds this unproblematically true, yet it is highly problematic.

temperance is fostered via the craftsmen and auxiliaries acquiescing to the rule of the guardians, and justice is the result of a harmony in the proper functioning between the three parts.

Having described his just city, Socrates completes the analogy by likening the just city to the just soul of an individual. As with the city, the soul has three components; the reason, the spirited part (what we might refer to as a sense of honor or pride) and the desires or passions. In an individual, wisdom arises from the exercise of reason and courage arises from the exercise of pride and spirit. When the individual knows when to obey each element of the soul he is displaying temperance, and justice results when the rational part of the soul rules and all of the aspects work together in a harmonious fashion to foster the soul's love of knowledge. In other words, justice is merely the harmony of a soul driven by reason.

Through this attempt at a definition of justice, Socrates has established the idea that the concept of justice is preceded by, and inextricably tied to, an understanding of human nature and human psychology. But is his depiction of human nature, and his subsequent definition of justice, convincing? While it is very tempting to simply answer this question in the negative and move on, it will be helpful to the central themes of this essay to understand exactly where Socrates' deficiencies lie.

The most obvious problem with Socrates' description of human psychology is that he does not place enough emphasis on the value of experience; he relies almost entirely on a priori rational enquiry. Rather than observing the psychology of those around him and generating ideas based upon those observations, Socrates seems to want to make the facts of human psychology fit his claims about the workings of the just city. He neglects many important aspects of psychology such as dispositions, motivation and psychological dissonance, which renders his version of psychology incomplete. Various commentators have attempted to save Socrates'

notion of human psychology, but as we shall see, they cannot overcome his failure to simply observe human behavior as we find it. I will address Cooper's defense of human psychology as presented in *The Republic* here, and I will return to the topic again in the final section of this chapter.

Cooper focuses his defense of "Platonic psychology" on Socrates' account of the spirited part of the soul. Cooper wants to argue that, contrary to popular belief, Socrates is not conveniently and arbitrarily dividing human psychology into a tripartite soul just to correspond to his conception of the city; he instead argues that Socrates' tripartite soul is prior to his tripartite city and that it has a basis in facts about individual human motivation.[5]

For Cooper, Socrates' tripartite theory is merely the idea that there are three psychological determinants of choice and voluntary action, each of which has its own distinct type of pleasure and motivation. Reason is clearly present in the human psyche, and reason is motivated to rule. Appetites are clearly present as well, and they are obviously distinct from reason. Thus, the only element of Socrates' argument that may seem arbitrary is spirit. Socrates includes many different things under the heading of spirit, such as honor, pride, anger, shame, outrage and a desire to assert oneself, and it is this vagueness and lack of a unifying factor that makes spirit appear to be an arbitrary addition to the picture.

Cooper claims that there is, in fact, a unifying factor in the Socratic idea of spirit; something we would call "competitiveness" or the desire for self-esteem and the esteem of others. He argues that the psychological importance of competitiveness is evidenced by the fact that reason and competitiveness are often in conflict, causing us to feel differently that we think,[6]

[5] Cooper, John M. "Plato's Theory of Human Motivation." *History of Philosophy Quarterly* 1, no. 1 (January, 1984): p. 4
[6] Cooper, p. 15

and that competitiveness is therefore a third source of human motivation of the same importance as reason and appetite. He concludes that spirit (understood as competitiveness and the desire for esteem and self-esteem) is "an innate form of human motivation, distinct from the appetites and reason itself and equally as basic as they are to human nature."[7]

Cooper's attempt at salvaging Plato's version of human psychology falls short on two levels.[8] First, if he wants to argue that the Socratic concept of the tripartite soul is based in facts about human motivation, Cooper needs to show that there are three *and only three* elements of the soul, as Socrates claims. While Cooper goes to considerable effort to show that spirit is as much an essential element of human motivation as reason and appetite are, he ignores the possibility that there are other sources of motivation as well. That is, Cooper makes no attempt to argue that these three elements constitute a *comprehensive* list of the elements of the soul, and without demonstrating this he cannot plausibly claim that Socrates is giving an accurate account of the facts of human psychology.

Second, even if Cooper's claims about the composition of human motivation turn out to be a reasonable account of the facts, his claims about the nature of spirit simply do not correspond to Socrates' arguments in *The Republic.* That is, Socrates is not making the argument that Cooper attributes to him. The only reasonable conclusion is that Socrates produces a theory of human psychology that is incomplete, arbitrary and inconsistent with everyday human experience. However, while Socrates' description of human psychology certainly has its flaws, it is still useful in that it sets the stage for the more important question at hand, namely, whether or not just behavior is rational.

[7] Cooper, p. 17

[8] Recall that I am not attributing a specific account of human psychology to Plato here; I am addressing the account that Cooper attributes to Plato, while recognizing that this may not have been Plato's actual view.

Thrasymachus on the Rationality of Justice

Socrates' main ambition in the later books of *The Republic* will be to convince us that it is rational to be just. To this end, he will offer various arguments and proofs in response to the challenges of Glaucon and Adeimantus. However, before embarking on his more intricate defense of justice, Socrates must address a more primitive argument raised by Thrasymachus in Book I.

Thrasymachus is the original protagonist of the "justice is irrational" argument in *The Republic*. Socrates begins Book I by asking Cephalus and some others for a definition of justice. Soon after, Thrasymachus interrupts the conversation, and Socrates' description of him provides a hint regarding Thrasymachus' views on justice: "He coiled himself up like a wild beast about to spring, and he hurled himself at us as if to tear us to pieces."[9]

In keeping with his demeanor, the definition of justice that Thrasymachus offers is far different from justice as it is commonly understood. He says that justice means being concerned with one's own good, while injustice means being concerned with the good of another. He characterizes justice as "the advantage of the stronger,"[10] by which he means to argue that those in power establish laws to serve their own interest, and justice is nothing more than the acquiescence of weaker individuals to the oppressive laws of the stronger. While this is certainly a controversial definition, the question of its coherence is not relevant; in fact, Thrasymachus is not genuinely attempting to provide a definition of justice at all. His intent is to indicate his contempt for justice as it is commonly defined, and with this in mind he turns to a presentation of his own arguments on the rationality of justice after Socrates confounds his attempt at a clear definition of the term.

[9] "The Republic," 336b 3-5, p. 981
[10] "The Republic," 338c 1-2, p. 983

Thrasymachus is arguing that adhering to the principles of justice as they are commonly understood is foolish. For him, any concept of justice is temporal; it derives its meaning and validity from the prevailing public opinion of the time and it is not based upon any enduring truth. Under this characterization, the only reason people obey the rules of justice is that they are afraid of the legal consequences or social stigma of getting caught behaving in what is currently considered to be an unjust way, so, to the extent that one can get away with unjust behavior, one should do so. Those fools who obey the rules of justice out of respect for justice itself are placing an unnecessary and unreasonable restraint on themselves, and they will inevitably be taken advantage of by more clever, enterprising individuals who ignore the rules, behave in an egoistic fashion and get away with it. The sophist Antiphon[11] argued that a wise person treats law as important in the presence of witnesses and nature as important when there are no witnesses, and Thrasymachus would certainly agree with him. While Thrasymachus may concede that having a system of justice in place is beneficial for society as a whole, he believes that an individual who always adheres to that system of justice is simply foolish. For Thrasymachus, in many situations being unjust is preferable to being just, and an individual who has an inclination to observe justice on every occasion is irrational.

Socrates responds to Thrasymachus' challenge with three arguments in favor of justice. While these arguments as a whole are far from satisfying, Socrates does raise one important point here that we can build upon. He recognizes that all cooperative human behavior, even among unjust persons, requires some element of justice:

> …for when we speak of a powerful achievement of unjust men acting together, what we say isn't altogether true. They would never have been able to keep their hands off each other if they were completely unjust. But clearly there must have been some sort

[11] Antiphon, Diels-Krans 87 B44

> of justice in them that at least prevented them from doing injustice among themselves at the same time they were doing it to others.[12]

Socrates recognizes that all human group activity is, at its most basic level, dependent upon just behavior. If any of us are to accomplish anything above the level of basic physical survival, we will need the assistance of other individuals and we will need some assurance that we will not be exploited by those individuals. Even if the individual participants in a given venture are unjust persons who are striving for an unjust purpose, they still require some element of justice among themselves in order to successfully accomplish their goal. The fact that Socrates does not pursue this line of argument further is unfortunate. The sarcastic disengagement of Thrasymachus at this point of the discussion leads Socrates to abandon the entire argument thus far and start afresh with Glaucon and Adeimantus in Book II.

If we wish to engage Thrasymachus at the point where Socrates disengages, our own response to Thrasymachus must be made on two levels. As Reeve indicates,[13] Thrasymachus is arguing not only that just *actions* are irrational, but that the acquisition of a just *character* is irrational as well. That is, Thrasymachus believes that just actions are irrational because they directly prohibit one from acquiring the things that one desires, and a just character is irrational because it subjects one to the whim of the rulers in power. Thrasymachus must be addressed on both accounts.

The refutation of Thrasymachus' argument against just actions is not particularly difficult. Thrasymachus seems to view the natural state of human relations as a perennial free-for-all in which each individual is fighting with other individuals in an effort to acquire for himself the maximum amount of wealth and power. However, as Socrates indicates in his own

[12] "The Republic," 352 c, p. 996

[13] Reeve, C.D.C. "Glaucon's Challenge and Thrasymacheanism." *Oxford Studies in Ancient Philosophy* 34 (May 29, 2008): p. 100

description of human nature, humans are social animals. We naturally come together to form societies because we need help from each other and because we can only realize our full potential in the context of a community of other humans. If we did not share a common understanding of just behavior, we would not be able to thrive as a species and it is likely that we would devolve into a Hobbesean state of war with each other. Socrates recognizes this (albeit in a cavalier fashion) when he shows that some element of justice even exists within groups of unjust individuals. While this "honor among thieves" may not be true justice in the robust sense that Socrates is seeking, it does illustrate one very important aspect of justice: Despite the fact that each individual in such a group has an unjust character, they all benefit from the mutual observance of rules or norms in their relations with one another.

The driving force behind this phenomenon is the fact that human interaction is not a zero-sum game.[14] That is, the practice of justice is not analogous to a poker game where one individual's gain is dependent upon and is equal to another individual's loss. Justice is, instead, a value-producing practice in its own right. While Thrasymachus appears to believe that humans are involved in a struggle to obtain the largest possible share of a fixed amount of "goods" such as wealth and security, a cursory examination of actual human interaction will demonstrate that this is absolutely not a zero-sum game. The most essential reason that humans choose to interact and cooperate is their mutual desire to increase the total amount of wealth and security available to the human species as a whole. The practice of just actions does not benefit some individuals at the expense of others; it allows *every* individual to have more than she otherwise would.[15]

[14] Barney, Rachel. "Callicles and Thrasymachus." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Winter 2011 Edition), URL = <http://plato.stanford.edu/archives/win2011/entries/callicles-thrasymachus/>.

[15] It should be noted here that there is some disagreement on this point. Some characterizations of justice will claim that justice generates tradeoffs in which some individuals or groups benefit at the expense of others. In Chapter 2 of this paper and beyond, I will argue that, while justice does not benefit all individuals equally, it does make every

Thrasymachus also wants us to believe that humans are driven by an insatiable *pleonexia*, or the drive to have as much wealth and power as possible, always at the expense of others in a zero-sum world.[16] Yet, as our own experience tells us, our motivation to acquire more wealth and power is reduced by the law of diminishing marginal utility. It is a fact of human psychology that, as we acquire more and more of a particular asset, our desire for another unit of that asset diminishes.[17] Although some individuals are content with very little of a given asset while others will require much more of the same asset before becoming satiated, nearly all individuals reach a point where more of the same brings very little additional utility. Yet, while we may reach a level of *material* wealth and security where our motivation to acquire more of the same diminishes, we will still likely be interested in acquiring goods such as friendship, clear conscience and the esteem of oneself and of others; goods that cannot be had by unjust means.

Where an unjust individual seeks to obtain a larger piece of a fixed basket of goods, a just individual realizes that, by cooperating, she can increase the overall size of the basket of goods for everyone, increase her own odds of obtaining a larger absolute amount (although not necessarily a larger relative amount) of those goods than she otherwise would have had, and increase her ability to secure possession of her own goods, all without having to deprive others of theirs. When individuals deal with each other in a just fashion, fear is reduced, trust is fostered, commerce is made more efficient and the overall opportunity set of each individual is improved. Thrasymachus' claim that just behavior is irrational simply does not survive this basic observation.

---

individual better-off than she otherwise would be. See Gauthier, David. *Morals by Agreement*. New York: Oxford University Press, 1986, p. 320-321.

[16] See Barney, Rachel. "Socrates' Refutation of Thrasymachus." in *The Blackwell Guide to Plato's Republic*, edited by Gerasimos Santos, 44-62. Hoboken: Wiley-Blackwell, 2008, p. 46

[17] See Gauthier (1986), p. 318

It should also be noted that the "non-zero sum game" argument could be used to illustrate why a group of unjust individuals (as described by Socrates in 352c) would be better-served if they were to behave in a just fashion towards non-members as they have towards members of the group. That is, if this unjust group chooses to behave in a just fashion to outside individuals or other groups, it will be able to achieve more, both for the group and for the individuals within the group, than it did by acting in an unjust manner.[18]

As mentioned above, in Book I Thrasymachus is arguing against the rationality of having a just character as well as the rationality of just actions. We have overcome Thrasymachus' claims about just actions, but the refutation of his argument against the acquisition of a just character will prove more difficult. Socrates' attempt at this refutation is the central point in his conversation with Glaucon and Adeimantus in the later Books of *The Republic*. I will therefore postpone a detailed analysis of this argument until the next section, but a few general points against the argument of Thrasymachus in Book I will be a useful prelude.

The main flaw in Thrasymachus' rejection of a just character is that the benefits that may accrue to an individual as a result of performing acts of injustice are not necessarily the proper yardsticks with which to judge whether a course of action is rational or even desirable. When Thrasymachus claims that justice is "the advantage of the stronger,"[19] by "advantage," he means "that which leads to the accrual of wealth, power and influence over others." But, it is not necessarily rational to pursue these things at all costs. That is, while it may be true that these advantages do sometimes accrue to individuals who practice injustice, it does not necessarily follow that it is rational to pursue these advantages.[20]

[18] This idea will be made more explicit in the discussion of Robert Axelrod's research in Chapter 4.
[19] "The Republic," 338c 1-2, p. 983
[20] See Barney (2011), p. 6

In his account of injustice, Thrasymachus is pretending that wealth and power are the only rewards available to us. By doing so, he is failing to recognize the psychological opportunity cost of unjust behavior. Injustice may reap great material rewards for its practitioners, but benefits such as friendship, family relationships and clear conscience will evade the unjust individual. While such psychological benefits lack the tangible value of the goods Thrasymachus is after, it is still rational to pursue these benefits, and, as we will see later, the intrinsic value of the psychological benefits of justice make these goods preferable to Thrasymachus' material goods. Thrasymachus is simply ignoring the value of many of the best things in life, and he is unintentionally making the false claim that happiness is a zero-sum game.

In the remainder of *The Republic*, Socrates' main project will be to argue against Thrasymachus' character-based version of "egoistic eudaimonism"[21] and to demonstrate just how much human psychology and human happiness Thrasymachus has left out of his own account. In order to make his case, Socrates is forced to abandon the conventional understanding of justice and to start afresh. He needs to show that justice is in our own best interest, not only for the material benefits that just actions provide, but for the intrinsic value of a just character as well. That is, he needs to demonstrate that it is rational for us to prefer a life of *actual* justice to a life of *apparent* justice. Socrates' subsequent conversation with Glaucon and Adeimantus will serve this purpose.

Glaucon: The Intrinsic Value of Justice

Book II of *The Republic* begins with Glaucon's taxonomy of the different types of goods. He characterizes goods as either being good in themselves only (goods such as simple pleasures),

[21]See Reeve, p. 100

good for themselves and for the sake of other things (goods such as health and the senses), or good only for the sake of the related benefits they bring (goods such as money and exercise). Glaucon and Adeimantus are willing to concede that justice is one of these three types of goods, but the debate with Socrates will center upon which of the three types it is. Glaucon will argue that justice is a good of the third type; it is beneficial to us only because of the instrumental benefits it provides. Socrates, however, wants to argue that justice is a good of the second type; that it is good in itself as well as being good for its outward benefits. Socrates' efforts in the reminder of *The Republic* will be focused on his attempt to prove to Glaucon and Adeimantus that, while behaving in a just manner and giving an outward appearance of justice is in an individual's rational self-interest because of the instrumental benefits it conveys, it is also in one's self interest to develop a just character and to actually be a just person because justice brings intrinsic benefits as well. That is, Socrates is claiming that there are both moral and non-moral reasons for being just, and that justice is its own reward.

Glaucon is arguing against justice from two separate but related angles. First, he claims that it is better to be unjust than just. The unjust individual, he claims, is able to satisfy all of her physical desires and to secure all of the wealth she wants, while the just individual will be deficient in these areas. Glaucon therefore concludes that it is better to appear just and to act unjust than to actually behave in a just manner and to develop a just character. This is a similar but more sophisticated version of the second part of Thrasymachus' argument addressed above.

Glaucon's second argument is that justice is valuable only instrumentally. He acknowledges that, because justice is instrumentally valuable, it is in one's own self-interest to behave in a just manner. However, because justice is valuable *only* instrumentally (and not intrinsically), we behave in a just manner only for non-moral reasons. That is, he is calling our

motives into question and arguing that our motivation for observing the conventions of justice is not a moral motivation, but a selfish one. When we behave in a just manner, we only do so because of the social benefits that such behavior brings, and because we fear the consequences of getting caught if we behave unjustly. He uses the example of The Ring of Gyges to make his point: If I were able to avoid detection by others, I would have no reason behave in a just manner because such behavior would no longer bring me any instrumental benefits.

Socrates attempts to refute Glaucon's two-pronged argument via a series of proofs in Book IX. In the first proof, Socrates argues that justice is superior to injustice because the just person is the happier person. He contrasts the tyrannical person (and eventually the tyrannical ruler) with the aristocratic person, arguing that the injustice of the tyrant leads her to be fearful, unable to satisfy her desires and, consequently, unhappy, whereas the justice of the aristocratic person allows her to fulfill her more lofty desires and to attain happiness. This proof is loosely related to a proof from Book IV in which Socrates compares justice to health by arguing that the healthy body, like the just soul, derives its virtue from the fact that its component parts are properly arranged and in correct relation with each other. He is arguing, in response to Glaucon, that justice, like health, is good in itself and for the sake of something else. Socrates believes both of these proofs demonstrate that justice leads to psychological health and is therefore a rational course of action for a self-interested individual.

In the second proof, Socrates separates people into three categories: those who love pleasure and money, those who love honor, and those who love truth. He views these three objects of love as a hierarchy with pleasure and money at the bottom and truth at the top. The philosopher (the lover of truth) is the only type of person who has experienced all three of these pleasures, and is therefore the only one qualified to judge which type of life is best. The fact that

the philosopher has chosen the life of truth serves as proof that the life of truth (and justice) brings the best kind of pleasure. It follows from this that it is more rational to be just than unjust.

The third proof makes use of pleasure as well, claiming that the pleasure of the philosopher is the only real, permanent pleasure which does not have a corresponding pain. The lower, bodily pleasures can never be completely satisfied and are merely a temporary removal of pain. Clearly the real pleasure of the philosopher is superior to the other types of pleasures, and it is shown once again that it is rational to be just.

The argument that Socrates is making is similar to the Epicurean notion that fear is the enemy of happiness.[22] He is trying to convince us that justice, as the chief legislator of a balanced psychology, allows us to lead a life of fulfillment, self-confidence, social acceptance, serenity, clear conscience and freedom from fear. Injustice, on the other hand, is a state of constant fear, strife and inability to fulfill one's desires. In other words, the just life is the pleasant life, and we act in our own rational best interest when we behave in a just way.

It should be noted that each of these proofs attempts to demonstrate (with various levels of success) that justice is better than injustice because it results in happiness or a lofty type of pleasure. The problem for Socrates is that the challenge posed by Glaucon still has not been completely overcome. Recall that Glaucon's argument is two-fold: First, he claims that it is better to appear just and to act unjustly because such a strategy will allow one to reap all of the material benefits of just behavior and to exploit other individuals for one's own gain. Socrates has, to an extent, shown that Glaucon is wrong on this point because Glaucon is using an incomplete notion of human good. That is, by emphasizing only the material benefits of unjust behavior, Glaucon is misrepresenting human nature and ignoring the greatest benefits of having a

[22] Fear will reappear in the 17th century as a central theme of Hobbesean psychology and moral theory.

just character. Justice is worthwhile not only for the material benefits that can also be had by pretending to be just, but also for loftier benefits such as philosophical happiness which cannot be attained by someone who is merely faking it. Socrates has shown that, because of the psychological benefits of justice, it is in one's rational self-interest to behave in a just manner.

However, Glaucon also claims that justice is a type-three good, that is, justice is only beneficial because it leads to other valuable goods. Unfortunately for Socrates, none of his proofs convincingly overcomes this claim by demonstrating that justice is good in itself. Happiness is certainly a more worthwhile benefit of justice than the material rewards cited by Glaucon (and Thrasymachus before him), however, happiness is still a "by-product" of justice, not an inherent part of it. While this does not diminish the point that it is rational to behave in a just manner and to have a just character, it leaves Socrates open to Glaucon's contention that justice is good only because it leads to other, greater goods.

Recent scholars have attempted to save Socrates' argument in various ways. Shields claims that the difficulty Socrates faces in satisfying the challenges of Glaucon and Adeimantus is partially due to the fact that the challenges themselves are internally problematic.[23] Glaucon begins his challenge by describing three types of goods: those that are good in themselves (type-one), those that are good in themselves and for the sake of something else (type-two) and those that are good only for the sake of something else (type-three). However, as noted by Shields, Glaucon's taxonomy of goods does not seem to be exhaustive. Glaucon argues that type-two goods are the best of the three because they are good in themselves and good for the sake of something else. But why does he favor these type-two goods over type-one goods that are simply good in themselves? That is, what reason should we have for preferring an intrinsic good that

[23] Shields, Christopher. "Plato's Challenge: the Case against Justice in Republic II." in *The Blackwell Guide to Plato's Republic*, edited by Gerasimos Santos, 63-81. Hoboken: Wiley-Blackwell, 2008, pp. 67-68

leads us to another intrinsic good over an intrinsic good alone? Should we not prefer the direct route? Glaucon and Adeimantus themselves even seem to be confused about what they are asking: "I want to know what justice and injustice are and what power each itself has when it's by itself in the soul. I want to leave out of account their rewards and what comes from each of them."[24] Yet, they also want Socrates to "…praise justice as a good of that kind, explain how – because of its very self – it benefits its possessors and how injustice harms them."[25]

Shields' solution to this dilemma is that there is a higher type of good that lies behind this whole picture. The "simple pleasures" described by Glaucon as comprising the type-one goods are not of the same kind as the intrinsic goods that type-two goods lead us to. That is, there is a higher "type-four" good that stands behind the three types enumerated by Glaucon, and the type-two goods are of interest because they combine the simple pleasures of type-one goods with the added utility that they lead us to more important type-four goods, whatever those may be. Our initial impression that the type-one goods are of the same importance as the goods that are the object of the type-two goods is mistaken.

While the claim made by Shields does help Socrates somewhat by highlighting the inconsistency of Glaucon's line of questioning, it does not add any positive force to Socrates' own argument. A cleverer attempt at responding to Glaucon's claim that justice is only superior because of the benefits it provides is proposed by Reeve.[26] He begins with the observation that most scholars tend to argue that Socrates' reply to Glaucon's challenge needs to be a deontological one. That is, scholars believe that if we are to claim that justice is good for reasons other than the reputation it brings, we should not attempt to make this claim via a

[24] "The Republic," 358b 4-6, p. 999
[25] "The Republic," 367d 3-5, p.1007
[26] Reeve, p. 77-78

consequentialist argument. It therefore seems odd that both the challenge and Socrates' response are leveled on consequentialist grounds (612d 3-8).

Reeve claims that we do not need to make a deontological argument after all. A consequentialist defense of Socrates' position will suffice, but it needs to be a consequentialist argument of a distinct type. According to Reeve, when Glaucon claims that justice is beneficial only for its consequences, he is not making a claim about justice at all; he is making a claim about *reputed* justice. That is, all of the benefits of justice that Glaucon is interested in can just as easily be attained via a simulation of just behavior. If I merely pretend to be just, I can gain all of the reputational benefits that Glaucon claims are the driving force behind the motivation to justice. However, the same is not true of the loftier benefits of justice that Socrates is concerned with; these are not attainable via a mere simulation of justice. In Reeve's terms, the class of consequences of actual justice is larger than the class of "simulator accessible" consequences, and since Glaucon is only interested in the simulator accessible consequences, he is making a claim about simulated justice instead of the real thing.

Reeve concludes that a consequentialist defense of Socrates' argument is perfectly acceptable, as long as the defense is made using non-simulator accessible consequences. The consequences that Socrates mentions when he offers his three proofs in Book IX are of the non-simulator accessible type, so we can accept his consequentialist argument that actual justice is preferable to apparent justice. However, we have still come up short. While Reeve's formulation of Socrates' argument has lent some additional credence to the claim that justice is superior to injustice, it has not definitively shown that justice is good in itself. That is, Reeve has shown that justice is good independent of the rewards of having a reputation for justice, but this is different from pure intrinsic goodness.

All of this leaves us in a somewhat unsatisfying position. Some of the questions posed by Glaucon have been shown to be internally problematic, but Socrates' response leaves unproven his claim that justice is intrinsically valuable. Rather than attempt to resolve this shortcoming in the context of Socrates' argument, it is preferable to take the insights we have gained from *The Republic* and move on. Socrates has introduced the concept of the intrinsic versus instrumental benefits of justice and he has shown how difficult it is to define and distinguish between them. The recognition of this distinction will be of central importance to the conversation about the rationality of justice in the early modern period, as will the recognition of the importance of actions to an evaluation of justice.

The Importance of Actions and Their Influence on Psychology

Before moving away from ancient Greece, it will be helpful to pause and recognize one more shortcoming of Socrates' account of justice in *The Republic* that will play a major role in the account of justice given in seventeenth and eighteenth century Europe, namely, his apparent belief that actions are of little importance to an account of justice. For Socrates, justice is an internal psychological state of the soul that results from the proper education of desire, and the benefits of justice are based almost entirely on this internal psychological state. While he does believe that just actions will necessarily result from the development of a just soul, he does not believe that actions are the primary source of the benefits of justice, that actions play a significant role in the formation of a just character, or that actions should be the focal point for the evaluation of justice.

The linking of justice to a psychological state of the soul instead of actions creates two difficulties for Socrates. First, it prevents him from recognizing one of the most fundamental

aspects of our commonsense notion of justice, namely, outward-directed behavior that regards the welfare of others. Second, it influences him to give a questionable account of human psychology that ignores the impact of actions and dispositions on the development of a just character.

The claim that Socrates makes regarding virtuous action is this: If you have true knowledge of what virtue is, you will behave in a virtuous way because your reason will compel you to do so. When he speaks of knowledge in this context, Socrates is not referring to knowledge in the sense of "knowing that Tallahassee is the capital of Florida" or "knowing how to swim." He is referring to knowledge in the sense of knowing what to do in a given situation or knowing why A is more important than B. Knowledge of virtue in this sense is a state of the soul, and any individual who possesses this knowledge will necessarily be driven by reason to live in accordance with virtue through a consistent outward application of this virtuous inner state.

The problem here is that Socrates fails to account for justice in what most of us would consider the everyday sense of the term, namely, actions that take into account the interests of others. That is, under Socrates' characterization of justice, knowing what is important and how to act is a sufficient condition for virtue. But how can we be sure that an individual with a proper state of the soul will necessarily refrain from performing unjust actions towards others?

While Socrates does not explicitly address this concern, we can infer from his characterization of human nature that Socrates' just individual will behave in a just way towards others because of her desire to be connected with other people. Socrates recognizes that human beings are needy. We come together to form cities because we are not self-sufficient and we can only realize our full nature in the context of a community of other humans. Social connections

and unity with others are prerequisites of our own happiness and survival, and in order to make these connections we need to behave in a just manner towards others.

Singpurwalla argues along these lines[27] and invokes the unhappy tyrant as an example of the irrational nature of unjust behavior. As Socrates indicates, "someone with a tyrannical nature lives his whole life without being friends with anyone, always a master to one man or a slave to another and never getting a taste of either freedom or true friendship."[28] The tyrant is wealthy and powerful, but his disregard for just behavior prevents him from attaining happiness. Socrates wants to contrast this individual with the just individual, whose knowledge of the good allows her to recognize that her own good is intertwined with the good of others, and who will therefore necessarily behave in a just way towards others in order to attain this good for everyone.

While Singpurwalla is able to assist Socrates by inferring an account of other-regarding actions from his account of justice, Socrates' failure to give an adequate account of human psychology and his failure to emphasize the importance of actions in the development of a just character are far more damaging to his project. Aristotle recognizes these two shortcomings in Socrates' account of justice, and we can use Aristotle's own account of justice to see just how significant these shortcomings are.

These shortcomings in Socrates' account of justice are primarily due to the method that he uses. Where he employs rational enquiry alone, Aristotle's use of a posteriori practical experience allows him to give an account that more closely resembles the facts as we find them in everyday human life. Cooper suggests that there is a gradual progression from the Socratic argument presented in *Protagoras,* to the Platonic argument in *The Republic,* and finally to

[27] Singpurwalla, Rachel. "Plato's Defense of Justice in *The Republic*." in *The Blackwell Guide to Plato's Republic*, edited by Gerasimos Santos, 263-279. Hoboken: Wiley-Blackwell, 2008, p. 277
[28] "The Republic," 576a 3-5, p. 1184

Aristotle's argument in *Nichomachean Ethics*. The movement is from an account of virtue as pure rationalism to virtue as an "interfusion of reason and desire."[29] This is a plausible account of the arguments as presented, and it provides support to the observation that the movement from the Platonic Socrates to Aristotle is a movement from a rational approach which emphasizes mental states to an observational approach which emphasizes virtuous action.

Aristotle's emphasis of the importance of dispositions is critical. He explicitly emphasizes that a person's progress to a state of virtue involves not just a change in his knowledge, but a change in his psychology as well. Not only does he recognize that having the proper disposition is a prerequisite for an action to be virtuous, but more importantly, he understands that there are instances where having the proper disposition is not enough to lead a person to perform virtuous action. In other words, Aristotle recognizes that *psychological dissonance can be present in a just person*. This is a fact that is entirely obvious from observation, and one that Socrates unfortunately fails to emphasize.

It should be noted that Socrates does recognize some notion of dispositions, but he understates their importance. According to Joseph,[30] Socrates is referring to dispositions when he discusses the harmony that results when a person "puts himself in order" (*Republic*, 443c8 – 444a1). Joseph also attempts to draw an analogy between Aristotle's dispositions and Socrates' emphasis of courage in the virtuous soul.[31] While there may be some similarity here, this analogy is far from perfect. Aristotle is making an explicit claim about the psychological state of a

[29] Cooper, p. 3

[30] Joseph, Joseph, H.W.B. "Aristotle's Definition of Moral Virtue, and Plato's Account of Justice in the Soul." *Philosophy* 9, no. 34 (April, 1934): p. 173

[31] Joseph, p. 177

person when that person is involved in action of an ethical nature, whereas Socrates is not directly concerned with action or observational evidence.

The importance of Aristotle's methodology is even more evident in his emphasis of the importance of action in the development of a virtuous character. Like Socrates, Aristotle does believe that knowledge of virtue is a necessary condition for the development of a virtuous character, but for Aristotle it is not a sufficient condition: "intellect itself moves nothing."[32] Aristotle believes that for an individual to be virtuous, knowledge of virtue must be combined with a long-standing habit of performing virtuous actions which results in a disposition to behave in the correct way in a variety of situations. He realizes that there are no hard, fast rules that can be universally applied to the ethical decision-making process. Instead, He emphasizes the importance of action and the use of correct deliberative judgment in particular ethical instances. In real life no two ethical situations are exactly alike and it is therefore necessary for one to have a good deal of practical experience in order to ensure that one is trained to properly analyze the situation within an ethical framework and to take the correct action based upon that analysis. Aristotle's characterization of virtue explicitly accounts for this fact of life, whereas Socrates' does not.

London is quite helpful in this regard when he stresses the point that "actions are in the particulars."[33] That is, knowledge of universals is a helpful tool in making correct ethical judgments, but since ethical situations are varied and unique, knowledge of the particulars is more important. Only experience in the actions of daily life can lead to knowledge of the particulars involved in virtuous choice, and this experience is therefore essential to the

[32] Aristotle. "Nichomachean Ethics." in *The Basic Works of Aristotle*, edited by Richard McKeon, 927-1112. New York: Random House, 1941, 1139a35-6, p. 1024

[33] London, Alex J. "Moral Knowledge and the Acquisition of Virtue in Aristotle's 'Nichomachean' and 'Eudemian Ethics'." *The Review of Metaphysics* 54, no. 3 (March 2001): p. 568

development of the ability to deliberate correctly: "Nor is practical wisdom concerned with universals only – it must also recognize the particulars; for it is practical, and practice is concerned with particulars."[34] Aristotle recognizes that experience in real ethical situations is a necessary condition of the development of a virtuous character. That fact that Socrates underemphasizes this point diminishes the force of his argument.

Aristotle's experience-based approach, his superior account of human psychology, and his emphasis on deliberation and action expose serious flaws in the account of justice offered by Socrates. However, the exposure of these flaws raises three points that will be quite helpful in the development of the central argument of this essay. First, a proper account of justice needs to be based upon an observational investigation of the facts as we find them in everyday life, and not upon an other-worldy rational enquiry. Second, in order to give a coherent account of justice we must first have an understanding of human nature and human psychology; dispositions and psychological dissonance are important in this regard. Third, when constructing an argument in favor of the rationality of justice, the emphasis should be placed on just actions rather than justice as a virtue or state of the soul.

Conclusion

The account of justice given by Socrates in *The Republic* may be incomplete and unsatisfying, but it does succeed in several ways. Socrates is successful in refuting the primitive argument against justice posed by Thrasymachus. By recognizing that the benefits of justice are not limited to material goods, he allows us to focus on the loftier benefits of justice and to

[34] "Nichomachean Ethics," 1141b14-6, p. 1028. Note that the word "practice" here is sometimes translated as "action."

recognize that these benefits are not gained in the context of a zero-sum game. Also, the mere presence of Thrasymachus introduces us to the literary method of the "amoral interlocutor," which will recur in the 17th century in the person of Hobbes's Fool.

Also, he raises the critical distinction between the intrinsic and instrumental benefits of justice. More will be said on this topic as the essay progresses, but two ideas should be briefly recognized here. First, Socrates has shown us that demonstrating the intrinsic value of justice is a far more difficult task than demonstrating its instrumental value. He is able to quickly dispense of the instrumental-based argument made by Thrasymachus, but his attempt to satisfy Glaucon's request for a poof of intrinsic value is ultimately unsuccessful. Second, Socrates has demonstrated that justice can be a rational strategy even if it is only instrumentally valuable. That is, the intrinsic benefits of justice, if they can be proven, may provide a nice bonus to the benefits of just behavior, but the instrumental benefits alone are sufficient for the advancement of an argument in favor of the rationality of justice.

Finally, and most importantly, Socrates has shown us that any concept of justice that we advance must be based upon an underlying concept of human nature and psychology. He assists us further by demonstrating that if the underlying concept of human nature and psychology is flawed (as his is), the resulting formulation of justice will be flawed as well. As we will see when examining the 17th century debate over justice, as the underlying concept of human nature evolves over time, the resulting concept of justice will evolve with it. The various views of the natural state of humans advanced by Hobbes, Hume and others of their era, as well as Hobbes's emphasis on actions rather than virtues or states of the soul, will have major ramifications for the direction of the debate over the rationality of justice. To this we will now turn.

CHAPTER 2: HOBBES, HUME AND THE EARLY MODERN CONCEPT OF JUSTICE

The preceding chapter was meant to demonstrate that many of the recurring themes relating to justice can trace their origins all the way back to the very foundation of western philosophy. In this chapter, I will address these same themes as they reappear in the early modern period. Many philosophers of the early modern period address the topic of justice from various perspectives, but the two accounts that have had the most impact and generated the most commentary are those of Thomas Hobbes and David Hume. These two figures will be the primary emphasis of this chapter.

One of the distinguishing features of the moral philosophy of this period is the explicit recognition of the importance of human nature and psychology to the formulation of a moral system. Where Socrates proposes an unsatisfying rationalistic account of human nature in an attempt to support the theory of justice he is promoting, Hobbes and Hume each offer a detailed account of human nature and psychology upon which they subsequently build their theories of morality in general and justice in particular. Hobbes views the establishment of the rules of justice as a rational response to the fear of the predations of others. Hume, in contrast, characterizes justice as a feeling of sympathy with public utility that we experience when we observe the performance of an act of justice. While these two accounts of justice differ dramatically from one-another, they are both distinguished from the earlier account offered by Socrates in that they are based on plausible accounts of human nature and the human psychological constitution.

Although Socrates' moral psychology is abandoned in the early modern period, his implicit argument that justice is not a zero-sum game is embraced. Both Hobbes and Hume (and other philosophers of this period) recognize that those who adhere to the rules of justice will receive benefits that they could not attain if they operated outside these rules. This concept is the driving force behind Socrates' instrumental refutation of Thrasymachus and, as we will see, it will serve as the motivation to exit Hobbes's state of nature as well as the reason for an individual's willingness to adopt the conventions of justice as proposed by Hume.

One aspect of Socrates' account of justice that is neglected in the early modern period is the distinction between the instrumental and intrinsic value of justice. Hobbes, in particular, views justice solely instrumentally.[35] For him, justice is merely the self-interested establishment of a covenant as a form of protection from others; he is not concerned with the "loftier" benefits of justice espoused by Socrates. For the time being, we can move forward without making significant reference to the intrinsic value of justice because the primary goal of this essay is to demonstrate that adherence to the rules of justice is a rational strategy for a self-interested actor, and this can be demonstrated by referencing the instrumental benefits of justice alone. However, I will return to the question of intrinsic value in later chapters.

The chapter will proceed as follows: I will interpret the theories of justice proposed by Hobbes and Hume, with particular emphasis on their contrasting descriptions of human nature and psychology. Next, I will argue that Hobbes's account of justice is more satisfying because his act-based instrumental approach is better-suited to a defense of the rationality of justice than Hume's virtue-ethical approach is. Finally, I will briefly address Adam Smith's moral

[35] The notion that justice is valuable solely for instrumental reasons was viewed as an Epicurean idea, and Hobbes and Hume were both considered to be Epicureans because they emphasized the instrumental value of justice and largely ignored the possibility of intrinsic benefits. See Epicurus, *Principal Doctrines*, #33, 36-38.

psychology, both as an improvement to Hume's account and as a prelude to some important concepts to be presented in later chapters.

Hobbes: Justice as a Rational Response to Fear

The philosophy of Thomas Hobbes and the contractarian philosophers who followed him will hold a place of central importance in the remainder of this essay. I will therefore dedicate a significant amount of effort to an exegesis of his philosophy in order to point out several relevant themes that have their origins in his work. Hobbes's ideas on justice, and morality in general, are found mostly in *Human Nature* and *Leviathan*, so my commentary will focus on these two sources.

Hobbes does believe that there is such a thing as human nature, that certain traits are common to all humans, and that we are capable of understanding what these traits and this nature are. However, his account of our human nature departs from the Greeks in several respects. First, Hobbes's account is not teleological. That is, unlike Aristotle before him, Hobbes does not believe that humankind is forever striving towards some ultimate good or purpose. In fact, Hobbes is a moral relativist. He denies that there are any moral facts in nature at all and instead claims that terms such as "good" and "evil" are agent-relative. "Good" is merely what an agent desires and "evil" is merely what an agent dislikes. Despite his relativism, however, Hobbes will claim that as a matter of observational fact there are some goods upon which we will all agree, such as the preservation of our own lives. This general agreement is what eventually leads to the formation of a covenant.

Second, Hobbes does not believe that humans are political animals. Where Socrates and Aristotle believe that humans always find themselves in some familial or societal situation and that humankind cannot be conceived of without making reference to society or politics, Hobbes believes that society is merely a contingent construct created among individuals through the use of reason in order to advance their own interests. Hobbes believes that humans, in their natural state, are both anti-social and anti-political. Third, and most importantly, Hobbes, unlike Socrates and Aristotle, does not believe that humans are driven primarily by reason. As we will see, Hobbes views humans as being moved by appetites (such as avidity and a lust for power) and aversions (such as fear) rather than by reason. His concept of human nature can therefore be fairly summarized as a mechanistic string of appetites and aversions. Human action takes place when this string of appetites and aversions (known as deliberation) comes to an end; the last appetite or aversion is what we call the will, and the will is directed towards the promotion of one's own self-interest or preservation.[36] Thus, Hobbes sees human nature as a selfish pursuit of appetites that we find desirable and a selfish avoidance of aversions that we find undesirable. This does not mean, however, that reason will play no part in Hobbes's account; in fact, Hobbes will argue that we use the faculty of reason when forming covenants and adhering to the rules of justice. The point is merely that Hobbes does not believe that reason is the primary driver behind the actions of humans in their natural state.

What exactly are these appetites that we find so desirable and these aversions that we strive to hard to avoid? According to Hobbes, the primary object of human appetite is power. Because we are self-interested creatures, every action we perform is done with a view to our own preservation. We strive to attain and retain as much power as possible in order to secure our

[36] See Gauthier, David. *The Logic of Leviathan.* Oxford: Clarendon Press, 1969, p. 8

comfort and continue our own existence. This quest for power never stops because, due to the potential rapaciousness of our fellow humans, we can never attain enough power to completely ensure our own security. As Hobbes puts it:

> So that in the first place, I put for a general inclination of all mankind, a perpetual and restless desire of power after power, that ceaseth only in death. And the cause of this is not always that a man hopes for a more intensive delight than he has already attained to, or that he cannot be content with a moderate power, but because he cannot assure the power and means to live well, which he hath present, without the acquisition of more.[37]

It should be noted that this quest for power is not necessarily the result of a lust for competitiveness or competition in themselves. That is, humans pursue power as a means to secure their own interest, and competition is the inevitable result of such pursuit in a world of scarce resources. The conflicts that result from this competition are not the object of this appetite, but they are nevertheless an unavoidable side-effect.[38]

As the lust for power drives an individual into competition with his fellows, fear, an even more compelling aversion, restrains him. Fear played a major role in Hobbes's moral philosophy and his life,[39] and fear is the primary force behind humankind's eventual move out of the state of nature and into a society governed by the rules of justice. It may seem ironic that a creature driven by self-interest and a lust for power is even more forcefully driven by fear, but these two forces are actually self-reinforcing. As we observe the potential for rapacious power-seeking behavior in our neighbors, the fear of violent death at their hands compels us to engage in rapacious power-seeking behavior of our own. It is not even necessary for each of us to postulate that all other people are aggressive and violent; the fact that *some* individuals *may* be aggressive and violent is enough to stoke the fear of violent death in one's heart and cause even a

[37] Hobbes, Thomas. *Leviathan,* edited by Edwin Curley. Indianapolis: Hackett Publishing Company, 1994, ch. 11, p. 58

[38] See Gauthier (1969), p. 17

[39] In his Latin verse autobiography of 1672 Hobbes writes, "fear and I were born twins together."

moderate person to behave in an aggressive, uncooperative and violent way in order to defend her own interests. It is because we find ourselves in this constant state of fear that Hobbes claims that humankind cannot be driven by reason as Aristotle and Socrates would have us believe. Fear encourages us to ignore the advice of reason and to instead act according to our fickle emotions, which keeps us mired in an endless cycle of movement towards power and away from dangerous interactions with other individuals.

Hobbes refers to life in a world characterized by this endless cycle of power juxtaposed with fear as humankind's "state of nature." He will eventually claim that entering into a covenant with an all-powerful sovereign is the only way to exit the state of nature, but he first wants to describe the state of nature in detail in order to show us just how awful life would be in the absence of a sovereign power. It is important to note that the state of nature Hobbes describes is not necessarily a state that humanity experienced in the past; it is sufficient for his purposes that the state of nature is merely a potential state that we could inhabit in the absence of a sovereign power or in a condition of civil war. Hobbes wants to show the state of nature as a potentiality that we are afraid of; it is what drives our fear.

Because Hobbes's human nature is a mechanistic succession of appetites and fear, when conflicts inevitably arise during the competition for scarce resources, the only possible result is all-out war. Hobbes famously describes his state of nature as a war "of every man against every man," and human life as "solitary, poor, nasty, brutish and short."[40] This unfortunate situation is characterized by two concepts that will be critical to the theory of justice that Hobbes develops later in Leviathan. First, the state of nature is a situation of equality:

> if we consider how little odds there is of strength or knowledge between men of mature age, and with how great facility he that is the weaker in strength or wit, or

[40] *Leviathan*, ch. 13, p. 76

> in both, may utterly destroy the power of the stronger, since there needeth but little force to the taking away of a man's life; we may conclude that men considered in mere nature, ought to admit amongst themselves equality; and that he that claimeth no more, may be esteemed moderate.[41]

The equality Hobbes speaks of here is not equality in the contemporary sense of equal rights; as we will see shortly, there are no rights (in the liberal sense) in the state of nature. The equality that arises in the state of nature is more akin to an equal ability to kill one-another. Despite the fact that some individuals will be more powerful, clever and ambitious than others, the range of power among individuals is relatively small. Human life is somewhat frail, and the ease with which one individual can kill another puts us in a situation where even the weakest among us can destroy the strongest.[42] Because of this, we are wise to assume equality of ability when we find ourselves in a state of war.

Another key element of Hobbes's state of nature is that it is a "pre-moral" state in which there is no law and might makes right. That is, when we are competing for scarce resources in a violent environment where we each have a roughly equal ability to kill one-another, there are no rules of conduct. There is nothing objectively wrong with killing another individual who is competing for scarce resources that one wants, and stealing is a perfectly acceptable act because the concept of property cannot possibly exist. The very idea of theft would be incomprehensible. Hobbes is trying to show us that, in a situation where there is no sovereign, morality can have no place, nor can any practice of justice. The implication is that we owe all of our moral conduct and even the very idea of justice to our covenant with the sovereign, and if the covenant disappears, morality and justice will disappear with it.

---

[41] Hobbes, Hobbes, Thomas. *Human Nature,* edited by J.C.A. Gaskin. New York: Oxford University Press, 1994, ch. 14, p. 78

[42] See Hoekstra, Kinch. "Hobbes on the Natural Condition of Mankind." in *The Cambridge Companion to Hobbes's Leviathan,* edited by Patricia Springborg, 109-127. New York: Cambridge University Press, 2007, p. 110

The situation described by the state of nature seems hopeless. If we are creatures driven by appetites and fear, using approximately equal abilities to compete for scarce resources in an environment of all-out war with no rules of conduct, how can we possibly escape this unenviable situation? Fortunately, Hobbes has an answer: reason. While we humans are primarily driven by our passions, reason does play an important part in our psychology. Hobbes claims that the application of reason to the misery of the state of nature will inevitably lead us to recognize that it is in our own best interest not only to form a covenant in which we exchange liberty for safety, but also to actually keep the covenants we make. Individuals will covenant with each other to establish a sovereign in order to ensure the compliance of all parties to the covenant. Our fear of the state of nature and our selfish desire for our own preservation makes us want to escape, and our reason will show us the way.

Once he has introduced reason as our ticket out of the state of nature, Hobbes introduces us to several "laws of nature", or, general rules of conduct discovered through reason, which prohibit an individual from doing that which is destructive to his life. These laws will guide us from our undesirable pre-moral condition to the formation of a state. Before he delves into the details of the laws of nature, Hobbes posits a single "right of nature" that we all possess, which is simply to defend oneself by any means necessary:

> The Right of Nature…is the liberty each man hath to use his own power, as he will himself, for the preservation of his own nature, that is to say, of his own life, and consequently of doing anything which, in his own judgment and reason, he shall conceive to be the aptest means thereunto.[43]

This right of nature grants us the freedom to engage in acts that foster our own survival and to condemn acts that hamper that ability to survive.

[43] *Leviathan,* ch. 14, p. 79

According to Hobbes, the right of nature combined with human reason will guide us to no fewer than nineteen laws of nature, but the bulk of his theory regarding justice can be found in the first three laws. The first of these laws of nature is simply stated: seek peace, but if peace cannot be obtained, seek war. Clearly the state of peace is rationally preferable to the state of war, and for this reason Hobbes argues that humans will, through fear and rational self-interest, seek and attain a state of peace. However, reason also dictates that the state of peace is desirable only under certain conditions, and Hobbes describes these conditions in the second law of nature.

The second law of nature is similar to the contemporary conception of the Golden Rule. Since an individual forfeits a certain degree of liberty when she agrees to make peace with others, this state of peace is rationally acceptable only under conditions of reciprocity. As he states in his definition of the second law of nature:

> that a man be willing, when others are so too, as far-forth as peace and defence of himself he shall think it necessary, to lay down this right to all things, and be contented with so much liberty against other men, as he would allow other men against himself.[44]

In other words, Hobbes believes that our rational self-interest will lead us to enter into an agreement in which we exchange some of our liberty for protection. This agreement is in the form of a *contract*, and it is the basis for Hobbes's notion of moral obligation. However, we will only enter this contract if it is in keeping with the right of nature. If we lay down our own rights but others refuse to do so as well, we are violating the right of nature by placing our very lives in danger. Thus, reason dictates that we will only relinquish our liberty under conditions where we can reasonably expect *reciprocity* from others.

[44] *Leviathan*, ch. 14, p. 80

The third law of nature is where Hobbes's concept of justice first appears. This law states that it is not enough simply to make contracts; we are also obligated to keep the contracts we make. The third law states:

> that men perform their covenants made, without which covenants are in vain, and but empty words, and the right of all men to all things remaining, we are still in the condition of war…And in this law of nature consisteth the fountain and original of JUSTICE.[45]

While these laws of nature and their corresponding definition of justice are quite simple on the surface, in the context of Hobbes's larger theory of human nature and commonwealth, they give rise to several implications which merit further consideration. First, justice is a direct result of the first two laws of nature. In order to seek peace, we must transfer rights via the use of contracts and we must be able to reasonably expect that others will transfer rights as well and comply with the contracts they make. Justice is therefore an inevitable product of our natural desire to exit the state of nature; without justice we would be unable to attain peace and, according to the first law of nature, we would then be obliged to seek war. Justice is simply a matter of observing contracts after they are made,[46] and the expectation of justice is a prerequisite to an escape from the rapacious state of nature.

Second, if the formation of a commonwealth via covenant is a rational and self-interested act, as Hobbes argues it is, then we must conclude that acting in a just manner is in our rational self-interest as well. That is, Hobbes is explicitly demonstrating how and why justice is a rational strategy for a self-interested individual. He has shown that, without contracts, we will find ourselves in an undesirable state of war, and that without justice (some assurance that others will observe these contracts), we have no reason to enter into the contracts in the first place.

[45] *Leviathan*, ch. 15, p. 89
[46] See Sorell, Tom. "Hobbes's Moral Philosophy." in *The Cambridge Companion to Hobbes's Leviathan*, edited by Patricia Springborg, 128-153. New York: Cambridge University Press, 2007, p. 140

Since we fear the horrors of the state of nature, it is in our rational self-interest to make these contracts, we have an obligation to observe these contracts, and (for Hobbes at least) we need a sovereign to assure that the contracts will be kept.

It is important to note here that the fact that Hobbes makes reference to appetites, aversions, passions and fear should not discourage us from analyzing his argument from the standpoint of rational choice. As Hoekstra indicates, rational choice can be driven by purely instrumental considerations based upon our desire to satisfy our passions. Our goals may be formed by a quest for power or an aversion to fear, but as long as our actions succeed in allowing us to achieve those goals, those actions can be considered rational.[47] In other words, justice places a rational constraint on some of our passions in order to allow us to satisfy more important passions.

A third important implication of Hobbes's laws of nature is the emphasis that they place on the importance of reciprocity. As indicated above, the idea of reciprocity first appears briefly in the second law of nature, but it plays a role of central importance in the remainder of his theory of justice. The need for reciprocity is obvious. Since we find ourselves in an initial situation of unbridled aggression, when we decide to lay down our right to violence we need to have some assurance that other individuals will reciprocate by also laying down their own right to violence, lest we become easy prey to their violent whims. To lay down one's own arms unilaterally without assurance of the same from the other side is simply irrational.

Many later commentators[48] on Hobbes have characterized the problem of reciprocity within the game-theoretic framework of the Prisoner's Dilemma (hereafter known as the PD).[49]

[47] See Hoekstra (2007), p. 115-116
[48] See David Gauthier, Kinch Hoekstra, Gregory Kavka, and Tom Sorell, among others.

The application of the PD to the reciprocity problem in Hobbes's second law of nature is straightforward. Suppose we have two individuals, Ashley and Brittany, who find themselves in Hobbes's state of nature. They are each driven by fear of violent aggression from the other, so they would each prefer to enter into a contract of mutual non-aggression rather than remain in the current state of war. The problem is that each of them stands to gain if they themselves fail to uphold the contract while the other party does uphold it. That is, if Ashley defects on the contract and Brittany upholds it, Ashley is better-off than if there were no contract at all. On the other hand, if Brittany defects on the contract and Ashley upholds it, Brittany is better-off than if there were no contract at all. Regardless of what the other party does, both Ashley and Brittany are better-off if they defect. If we operate under the assumption that Ashley and Brittany are rational actors, we can only conclude that, without some assurance of mutual compliance to contracts, both parties will defect and a state of peace will never emerge from the state of nature.

This is where the need for a sovereign enters Hobbes's moral theory. In order to assure reciprocity through mutual compliance with contracts made, a sovereign power must be instituted. The transferring of individual rights to this sovereign is in the best interests of all covenanters because, stated in the framework of the PD, the possibility of punishment at the hand of the sovereign significantly reduces the potential rewards of defecting on one's contracts. With the sovereign in place, Ashley and Brittany are no longer tempted to defect and they can each have a high level of assurance that the other party will perform, so they are very likely to be willing to exit the state of nature and enter a state of peaceful cooperation.

[49] Because the PD is perhaps the best-known "game" in all of game theory, I will not spend time here describing the mathematics behind it. I merely indicate that it is a valid framework from which to begin an analysis of the reciprocity problem.

However, Hobbes has to recognize that, even in a situation where the threat of punishment by the sovereign or ostracism by one's fellows should deter any individual from defecting on her contracts, there will still be some individuals who will attempt to get away with defection. Thus, where Socrates has his Thrasymachus, Hobbes has his Fool:

> The fool hath said in his heart: 'There is no such thing as justice'; and sometimes also with his tongue, seriously alleging that: 'every man's conservation and contentment being committed to his own care, there could be no reason why every man might not do what he thought conduced thereunto, and therefore also to make or not make, keep or not keep, covenants was not against reason, when it conduced to one's benefit.'[50]

The Fool argues that if humans by nature act in their own rational self interest, there is no such thing as the justice that Hobbes describes. According to the Fool, an individual's decision whether to keep or break a specific covenant should depend upon whether or not the breaking of the covenant will provide a benefit to that individual. Essentially, the Fool says that if it is to the benefit of an individual to break a covenant and he can get away with it, it is in his rational self-interest to do so, and therefore justice as defined by Hobbes can often be contrary to rational self-interest.

Hobbes, however, clearly does believe that justice is rational and beneficial to the individual, as he indicates in his response to the Fool's argument. The crux of his response is based not upon the idea that breaking covenants is "wrong" or "bad," but on the fact that a person who regularly breaks covenants cannot avoid the repercussions forever.[51] Hobbes argues that the very reason for forming a commonwealth is to enjoy the benefits of the safety provided by the commonwealth. If an individual seeks to break covenants whenever it serves him well to

[50] *Leviathan*, ch. 15, p. 90

[51] Again, we can see that this aspect of Hobbes' concept of justice is Epicurean: "It is impossible for the person who secretly violates any article of the social compact to feel confident that he will remain undiscovered, even if he has already escaped ten thousand times; for right on to the end of his life he is never sure he will not be detected." See Epicurus, *Principal Doctrines*, #35.

do so, that individual will almost certainly be discovered eventually, since getting away with such transgressions can only be achieved via the ignorance and errors of others, which are random and unpredictable. In other words, a defection strategy will not work in the long-term, and even if it does work in the short-term, it is due to luck rather than ability, and the defection strategy was therefore not reasonably followed. Once (inevitably) discovered, the individual who has broken the covenant will almost certainly either be banished from the commonwealth and returned to the state of nature, or he will be punished within the commonwealth for the injustice that he has practiced. These are both undesirable outcomes which result in a loss of the safety and comfort provided by the commonwealth, and this course of action is therefore irrational and against the individual's own self-interest.

Hobbes's initial reply to the Fool is brief and incomplete, and it fails to address many potential rebuttals. The core of his argument against the Fool is simply that the downside of breaking covenants is huge and success is uncertain. While he is certainly on the right track, Hobbes fails to consider situations in which the perceived utility of cheating is high and the perceived probability of being detected is low. I will therefore present a thorough analysis of the Fool's argument and offer a neo-Hobbsean contractarian reply to the argument in Chapter 4 of this essay. Before turning to that, however, I will examine a competing theory of the rationality of justice proposed by David Hume. Hume approaches justice from a very different perspective than Hobbes, so his ideas will serve as an effective contrast to the Hobbesean tradition. Although Hobbes's contractarian account will provide a more firm foundation for my argument in favor of the instrumental rationality of justice, Hume does have an influence on many contemporary contractarian philosophers and his philosophy will be useful when we return to questions regarding the intrinsic value of justice in Chapter 5.

Hume: Justice as Convention

Hume's theory of justice is written in response to other early modern British moral philosophers who want to find the nature of justice in actions or who, like Hobbes, want to characterize justice as a part of natural law. Hume is also rejecting ethical rationalism; he wants to deny that moral facts are discovered by reason and that reason is in accord with moral goodness.[52] Hume instead gives an experience-based account of human nature and morality in which he links morality to a feeling of sympathy that we have with the motive behind a particular act. He is attempting a virtue-ethical rebuttal to the law-based morality of his recent predecessors.

Like Hobbes before him, Hume is not a proponent of reason as the driving force behind human psychology. He believes that moral distinctions arise from our passions and not from the faculty of reason and that reason cannot motivate the will. When we see another individual performing an act that we consider to be virtuous or vicious, our moral evaluation of that act stems not from rational contemplation of the act, but instead from a moral sentiment of approval or disapproval that we feel.

This Humean idea of reason as subordinate to the passions is skillfully explained in contemporary terms by Haidt.[53] He uses evolutionary psychology to describe how reason evolved *after* intuition (intuition being Haidt's substitute for Hume's passions), not because reason was a replacement for intuition, but because reason served a useful purpose for intuition. In Haidt's analogy, reason is to intuition as the rider is to the elephant; the rider came into being in order to help the elephant reach its goals and explain its actions, but the elephant still holds the

[52] See Cohon, Rachel. "Hume's Moral Philosophy." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Fall 2010 Edition), URL = http://plato.stanford.edu/archives/fall2010/entries/hume-moral/, p. 9-10
[53] See Haidt (2012)

power. Haidt's experiments support Hume's claim that people make moral judgments based on intuition and emotion, and that "moral reasoning was mostly just a post-hoc search for reasons to justify the judgments people had already made.[54] [55]

In addition to rejecting the supremacy of reason, Hume also rejects the notion that any action can be intrinsically right or wrong in itself. Instead, he claims that the morality of an action is determined by the motive behind that action, and the motive itself is often driven by a particular character trait. These motives and the character traits that drive them do have an intrinsic morality, but the actions they produce do not:

> 'Tis evident, that when we praise any actions, we regard only the motives that produc'd them, and consider the actions as signs or indications of certain principles in the mind and temper. The external performance has no merit.[56]

Virtues and vices are character traits that are possessed by an individual who performs an action, and our moral evaluation of that action is driven by sympathetic feelings that we have concerning the character traits that gave rise to the motive behind the action.

It is important to note that Hume believes that these feelings of sympathy[57] are a social phenomenon and the moral judgments derived from them are made without regard to self-interest. That is, our feelings of sympathy are sympathetic with the public utility and general societal welfare that is created when virtuous acts are performed. We consider actions moral or

[54] Haidt (2012), p. 40
[55] Haidt's claims regarding reason and intuition are similar to those made in Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Strauss and Giroux, 2011. Kahneman's work will be addressed Chapter 4.
[56] Hume, David. *A Treatise of Human Nature*, edited by David F and Mary J Norton. New York: Oxford University Press, 2000, 3:2:1.2, p. 307
[57] Hume's use of the term "sympathy" would be better represented in contemporary American English as "empathy." He wants to convey the idea that when we see the effects of a passion or the causes of an emotion in another person, we ourselves actually feel what the other individual feels. See *Treatise*, 3:3:1.7, p. 368

immoral only with respect to how they affect others, not with respect to how they affect ourselves, and correct moral judgments can only be made from "some common point of view."[58]

Hume's account of human psychology has a profound influence on his description of humankind's natural state. He claims to reject Hobbes's state of nature as "a mere fiction,"[59] but his own account of humankind prior to the recognition of justice probably has more in common with Hobbes than he would like to admit.[60] Hume's state of nature is characterized by a mean between the extremes of superabundance and superscarcity. That is, Hume claims that in order for a moral system to have any relevance, it must be assumed that there is not a superabundance of goods (because in such a situation there would be no conflicts among possessions) nor is there a superscarcity of goods (because in such a situation the rules of morality would be suspended and we would likely descend into a Hobbesean state of war). Hume also finds it necessary to assume that humans are self-interested in their natural state. He bears some similarity to Hobbes in this respect; Hume's human is selfish and aggressive by nature and in his initial situation of limited scarcity he is incented to act in his own interest by taking what he needs by the use of force. It should be noted that Hume is somewhat more charitable than Hobbes in his claim that some moral virtues such as benevolence do exist in humankind's natural state, but he is explicit in his assertion that justice will not be found there.

While Hobbes spends considerable effort in giving a detailed description of life in the state of nature, Hume is less concerned with describing humankind's natural state and far more concerned with providing details on how we managed to remove ourselves from it. For Hume, this transition from the state of nature to moral society is where reason and self-interest make

[58] *Treatise*, 3:3:1.30, p. 377
[59] *Treatise*, 3:2:2.15, p. 317
[60] See Pack, Spenser J. and Eric Schliesser. "Smith's Humean Criticism of Hume's Account of the Origin of Justice." *Journal of the History of Philosophy* 44, no. 1 (2006): p. 49

their most important contribution to humanity's well-being. In the state of nature, an individual is motivated to act in her own self-interest and to use force to take what she needs. We move out of this state by a sequential process: First, via the use of reason we discover practices that foster economic cooperation and coordination. Second, we become aware that these cooperative activities are value-creating and that our own self-interest is served by observing these practices because they are not part of a zero-sum game (whereas using force to take what we want *is* a zero-sum game). Third, adherence to these rules becomes regarded as moralized conduct via our sympathy with the public good that is served when people follow the rules. Rather than using an explicit purposeful *contract* with a sovereign power to escape the state of nature in an instant, Hume's natural individual, through a sort of "spontaneous order" gradually adopts certain *conventions* that allow him to escape the state of nature over time.

Hume's notion of the conventional origins of justice can be understood as analogous to the conventional origins of language. Justice, like language, develops over time without design or intention. In the case of language, we have a need to communicate, so our linguistic practices emerge over time to fulfill that need. In the case of justice, we have a need for social interaction and commerce, so the practice and principles of justice emerge over time in order to facilitate this interaction. In both cases, a useful and eventually indispensible practice evolves, but no explicit agreement among individuals is required.[61]

Thus, Hume characterizes many of our moral virtues, and the virtue of justice in particular, as "artificial" virtues.[62] Rather than being a naturally-occurring part of the natural

[61] This interpretation owes much to a personal conversation with Professor Colin Heydt

[62] Note that, In Hume's time, the term "artificial" meant "a work of reason." The term did not have the negative connotation that it has today and it was used to describe actions that are performed by design and with intention. See Rawls, John. *Lectures on the History of Moral Philosophy*. Cambridge: Harvard University Press, 2000, p. 52-3

state of humankind, justice is a matter of "honesty with respect to property";[63] it is a product of reason and self-interest which develops as a convention as humanity moves from its primitive state into a state of moral society. The rules of justice are initially adopted only in the context of a mutual self-interest to maintain property rights, but as they develop into conventional norms, these rules become vested with a sense of morality due to the sympathy that we mutually feel upon contemplating their benefits for public utility:

> we are to consider this distinction betwixt justice and injustice, as having two different foundations, *viz*. that of *self-interest*, when men observe, that 'tis impossible to live in society without restraining themselves by certain rules; and that of *morality*, when this interest is once observ'd to be common to all mankind, and men receive a pleasure from the view of such actions as tend to the peace of society, and an uneasiness from such as are contrary to it. 'Tis the voluntary convention and artifice of men, which makes the first interest take place; and therefore those laws of justice are so far to be consider'd as *artificial*. After that interest is once establish'd and acknowledg'd, the sense of morality in the observance of these rules follows *naturally*, and of itself…[64]

As alluded to earlier, Hume distinguishes between artificial virtues, such as justice, and natural virtues such as benevolence. He claims that we can distinguish between the natural and artificial virtues by examining the good that arises from each of them. Whereas the good that arises from natural virtues such as benevolence is evident in every act that is motivated by them, the good that arises from artificial virtues such as justice arises only from their continued practice and is not necessarily immediately obvious:

> The only difference betwixt the natural virtues and justice lies in this, that the good, which results from the former, arises from every single act, and is the object of some natural passion: Whereas a single act of justice, consider'd in itself, may often be contrary to the public good; and 'tis only the concurrence of mankind, in a general scheme or system of action, which is advantageous.[65]

[63] Cohon, p. 18
[64] *Treatise*, 3:2:7.11, p. 342
[65] *Treatise*, 3:3:1.12, p.370

Acts which are motivated by natural virtues such as benevolence give rise to unambiguous sentiments of approval every time we observe them. Acts motivated by artificial virtues such as justice can at times give rise to sentiments that are not so clear. For example, if I see someone help a stranger change a flat tire, this act of benevolence causes me to feel a sentiment of approval that is immediate, direct and obvious because this act is unambiguously beneficial to overall public utility. However, if I see someone being issued a speeding ticket, my sentiments regarding the justice of this act may be delayed, indirect and obscure. Having been on the receiving end of traffic tickets in the past, I will likely sympathize with the financial loss incurred by the offender. I may inwardly protest that the offender's act was benign, that she is probably a decent person and she was not placing any other individuals in danger, and therefore the act of issuing her a ticket is not virtuous. It is only when I contemplate the fact that the enforcement of safe driving rules are beneficial to the long term utility of society that I can sympathize with the justice inherent in the act of writing a traffic ticket and recognize that goodness does arise from it. Hume would likely agree that the recognition of the good arising from this particular act of justice requires effort and the use of reason because justice is not a natural virtue.

In summary, Hume provides us with a characterization of justice that emphasizes sympathy with public utility while still allowing a place for reason and self-interest. In Hume's framework, moral judgments are made without reason and self-interest, but reason and self-interest do come into play when they lead us to adopt conventions in order to escape the state of nature and enter moral society. Self-interest is necessary for the practice of justice to originate, but justice matures from there, eventually becoming a conventional sympathy with public utility. While Hume does think justice is artificial in that it is a product of reason and not a part of

humankind in its primitive state, he does not think that the fact that we arrive at the convention of justice is at all arbitrary. Through a mechanism of spontaneous order, we will necessarily arrive at a convention of justice without the need for an explicit social contract. Justice is therefore a necessary result of rational self-interested behavior and it is inseparable from the human species.

Hume's account of justice draws markedly different conclusions than that of Hobbes, but their starting point is quite similar. Like Hobbes, Hume bases his account of justice on a corresponding account of humankind's state of nature. Hume is in general agreement with Hobbes in that they both envision humankind's natural state as one of rapaciousness and self-interest where justice exists neither in concept nor in practice (although Hobbes is more graphic and radical in his portrayal of this beastly nature). In both accounts, the world in which we humans initially find ourselves is unpleasant and inefficient, we have a self-interested incentive to escape this situation, and we will allow our reason to show us the way.

While Hobbes and Hume mostly agree on the state of nature and the need to escape it, they differ significantly in their description of the means that we use to make our escape. For Hobbes, we make our escape via an explicit and purposeful agreement with a sovereign power. Human society (and therefore justice) is the direct result of this covenant with the sovereign, and without this explicit purposeful agreement, humankind is forever doomed to remain in the state of nature and no concept or practice of any moral virtue (including justice) can arise. Hume needs no such purposeful agreement; for him, our escape from the state of nature is the inevitable result of a mutually beneficial accident. While the natural virtues, such as benevolence, are already present in Hume's state of nature, the artificial virtues, such as justice, arise via spontaneous convention. Hume recognizes no explicit agreement among individuals or action on

the part of a central authority that saves us from the state of nature and drives us to the recognition of property rights and justice. Instead, justice is the inevitable result of a spontaneous order that develops over time via an iterative process of reason. We find that the observance of justice is useful to us, and we therefore agree amongst ourselves to observe its tenets.[66]

While this idea of unplanned convention or spontaneous order is in sharp contrast to Hobbes's explicit social covenant, it is quite similar to Adam Smith's famous idea of the "Invisible Hand" from *Wealth of Nations*. Smith describes how hundreds or thousands of independent actors motivated by their own economic self-interest unintentionally work together to bring a product or service to the end user.[67] For example, in order for me to drive to work today, I had to buy a car at some point, I had to put fuel in that car and I have to have faith that if my car breaks down, I can depend upon a reliable repair service to get my car running again. We take for granted that all of these things will be available to us whenever we need them, but it is no small miracle that all of these necessities come together through the independent work of complete strangers without explicit planning on the part of a central authority or sovereign. Hume believes that the convention of justice arrives via a similar mechanism. We find that other individuals, whether they are family members or complete strangers to us, will agree that it is in everyone's best interest to respect the property of others provided others agree to respect theirs.

[66] It should be noted that Hobbes' account of justice can also be characterized as conventional in the sense that justice is chosen by us rather than imposed upon us by human nature or the natural world around us. The key difference between Hobbes and Hume on this point is that in Hobbes' account the convention is the result of an explicit agreement among individuals, whereas Hume sees the convention arising without any such agreement. See Rescorla, Michael. "Convention." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Spring 2011 Edition), URL = http://plato.stanford.edu/archives/spr2011/entries/convention/.

[67] Smith, Adam. *The Wealth of Nations,* edited by Edwin Cannan. New York: Modern Library, 2000

That is, we have all come to a common understanding through a mutually beneficial accident without the assistance of an explicit agreement.

Despite the differences in their respective accounts of humankind's escape from the amoral state of nature, Hume and Hobbes do concur on one critical point: The escape is facilitated by the recognition of the rules of justice (whether by explicit covenant or spontaneous convention) on the part of rational self-interested actors. Hume does believe that human action is driven by passions that rule over our reason, however, it is reason that allows us to re-channel our passions and to recognize the benefits of coordinated behavior. Postema characterizes this as a "paradox of avidity." Individuals seek to survive and to acquire status and security through the competitive acquisition of possessions, but the more passionate this contest becomes the more dangerous and destructive it is to the participants. Reason allows us to establish the conventions of justice in order to redirect these potentially dangerous passions from socially destructive to socially beneficial uses, while continuing to act in our own self-interest.[68] Reason will never be completely in charge of the passions, but reason does guide humanity out of its undesirable natural state and into a state of peace and cooperation with others where justice can be found. Thus for Hume, as for Hobbes, justice is a rational strategy for exiting the state of nature.

A Fork in the Road

In their respective accounts of justice, Hobbes and Hume each succeed in advancing some of the most important themes introduced by Socrates. They each offer an account of human nature and psychology that is more scientifically sound than Socrates', which allows

[68] Postema, Gerald. "Whence Avidity? Hume's Psychology and the Origins of Justice." *Synthese* 152, no. 3 (Oct. 2006): p. 390

them to base their accounts of justice on a more firm foundation. They also advance Socrates concept of the value-added nature of justice by explicitly recognizing that justice is not a zero-sum game and that a general adherence to the rules of justice makes all participating individuals better-off than they would be in the absence of justice. And, while Hume underemphasizes and Hobbes completely ignores the possibility that justice has intrinsic value, they are both able to make a coherent argument in favor of justice as a rational strategy by referencing the instrumental benefits alone.

However, we have now reached a fork in the road. Despite some similarities, the two accounts of justice proposed by Hobbes and Hume are irreconcilable. In order to move forward it is necessary to decide which of them will provide a better framework within which to assess the rationality of just behavior. Although Hume will be of assistance at several points along the way, I believe that Hobbes gives us a better launching pad for an analysis of justice for several reasons. First, while Hume's account of human psychology is far superior to that of Socrates, and is probably more accurate than Hobbes's account in light of our contemporary notions, it still poses several problems for his account of justice. For example, Hume claims that our sympathy with justice arises because we recognize the utility that justice has for the public good. While this claim fits nicely within his larger theory of morals, it is not consistent with the facts as we find them in everyday life. Most people simply do not consider the greater good of society as they go about their daily activities, and when we applaud acts of justice or condemn acts of injustice, we seldom have utility in mind. Adam Smith was among the first to comment on this shortcoming of Hume's: "But few men have reflected upon the necessity of justice to the existence of society,

how obvious soever that necessity may appear to be."[69] Another contemporary of Hume's, Thomas Reid, rebuts Hume's utility argument in a very simple way; he merely uses a formulation of the Golden Rule to remind us that utility or "public good" is not a necessary consideration in our understanding of the virtue of justice:

> The simple rule, *Don't do to your neighbor what you would think wrong to be done to yourself* would lead him to the knowledge of every branch of justice, without any thoughts about public good or laws and statutes made to promote it. So it isn't true that public usefulness is the only standard of justice, and that the rules of justice can be derived only from their public usefulness.[70]

In addition to these obvious flaws raised by Smith and Reid, Hume's claims regarding sympathy for public utility create a more subtle problem for him. If sympathy with the public good is the driving force behind an individual's motive for justice, it is difficult to argue that the individual is viewing the situation from a self-interested perspective; such a motive is more aptly characterized as disinterested, and possibly even altruistic. Also, if the benefit of justice is to be found at the societal level, it remains to be proven that justice is a rational strategy on an individual level as Hume claims it is.[71] That is, while the value of justice for public utility makes justice a rational strategy for society as a whole, this does nothing to discourage a rational

---

[69] Smith, Adam. *The Theory of Moral Sentiments*, edited by D.D Raphael and A.L. Macfie. Oxford: Oxford University Press, 1976, II.ii.3.9, p. 89

[70] Reid, Thomas. *Essays on the Active Powers of Man, Essay V, Chapter V*. Early Modern Texts. URL<http://www.earlymoderntexts.com/rea5.html>, pp. 35-6.

[71] See Woozley, A.D. "Hume on Justice." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 33, no. 1 (Jan 1978): p. 90. Gauthier attempts to defend Hume in this respect by arguing that when Hume insists that sympathy with public utility is the origin of justice, he is not appealing to a utilitarian concept of total utility, but to a contractarian concept of mutual advantage. Each individual expects benefits for himself from justice, and these benefits do not enter into the moral approbation accorded to justice by Hume. According to Gauthier, what Hume really means is, "That public utility (i.e., mutually expected advantage) is the *sole* origin of justice, and that reflections on the beneficial consequences of this virtue are the foundation of its merit (i.e., moral approbation)." See Gauthier, David. "David Hume, Contractarian." *The Philosophical Review* 89, no. 1 (Jan 1979): p. 18

individual from behaving in an unjust way if doing so can be seen as beneficial to her without damaging the societal institution of justice.[72]

Hobbes's account is far more satisfying in this regard. For Hobbes, the motivation for justice can be described in purely self-interested terms, his account explicitly claims that justice is beneficial on an individual as well as on a group level, and his conversation with the Fool denounces the perceived advantages of unjust behavior by self-interested individuals. Hobbes has no need to make reference to sympathy with the public good because in his view individuals acting in their own self-interest in a coordinated way leads to an unintended, but welcome, increase in the overall public good as a by-product. Any mention of sympathy would be superfluous.

Hume's account of human nature is also flawed in that he insists on maintaining the distinction between natural virtues such as benevolence and artificial virtues such as justice. He views justice and the other artificial virtues as less reliable and somehow subordinate to the natural virtues. Smith argues, in contrast to Hume, that justice is the core foundational virtue and that it is subordinate to none:

> Though Nature, therefore, exhorts mankind to acts of beneficence, by the pleasing consciousness of deserved reward, she has not thought it necessary to guard and enforce the practice of it by the terrors of merited punishment in case it should go neglected. It is the ornament which embellishes, not the foundation which supports the building…Justice, on the contrary, is the main pillar that upholds the whole edifice.[73]

Hume's peculiar insistence on addressing justice only as it applies to property is particularly damaging. Reid and other contemporaries of Hume recognized six types of injustices

[72] It should be noted that this critique is based on flaws in Hume's argument for the *instrumental* value of justice only. Elsewhere Hume does argue in favor of the *intrinsic* value of justice on an individual level. See Hume, David. *An Enquiry Concerning the Principles of Morals*, edited by J.B. Schneewind. Indianapolis: Hackett Publishing Company, 1983, IX, part II, pp. 79-82

[73] Smith, *The Theory of Moral Sentiments*, II.ii.3.3, p. 86

that an individual can suffer: injuries in our persons, our families, our liberty, our reputation, our goods and our contracts. The fact that Hume's characterization of justice applies only to the last two of these categories of justice is a serious problem for his argument and a source of disbelief for Reid.[74]

As Woozley[75] and many others have recognized, Hume's preoccupation with property prevents him from seeing that there are obvious natural motives for justice. He emphasizes property rights because he believes that external goods are the only goods that can be taken from us.[76] However, this is clearly a misrepresentation of the various ways in which one individual can harm another. When we witness others being injured in their persons, in their reputations or in other naturally occurring goods, we clearly have a natural reason to feel sympathy with them and to feel sympathy with the utility of any system of justice that prevents these injuries from occurring.

Not only does Hobbes offer a more internally consistent account of human nature than Hume, Hobbes's reliance on laws of nature and moral acts as opposed to virtues provides a better context in which to analyze the game-theoretic aspects of justice. The idea that justice is a non-zero-sum game is merely the most basic concept in the argument for the rationality of justice. As the challenges to the argument become more sophisticated, we will need to respond in a more sophisticated manner by analyzing the rationality of justice in the context of game theory and prospect theory. Hobbes's reply to the Fool and the characterization of this reply as a version of the Prisoner's Dilemma give us the best starting point from which to launch a more sophisticated defense. While Hume also demonstrates some understanding of reciprocity and justice as non-

[74] See Reid, *Essays on the Active Powers of Man, Essay V, Chapter V*
[75] See Woozley, p. 94-99
[76] See *Treatise,* 3:2:2.7, p.313

zero-sum, his virtue-ethical approach does not permit analysis of the problem from a game-theoretic angle. Hobbes's account allows us to see in more objective and supportable terms why justice is instrumentally rational for a self-interested actor.

In addition, Hobbes's act-based version of the justice story is more suitable for our purposes because it is purely instrumental, whereas Hume employs both instrumental and intrinsic aspects of justice in his virtue-ethical argument. When confronted by the Fool (or with another skeptic in the tradition of Thrasymachus), Hobbes can remain consistent and make his reply on purely instrumental grounds, but Hume, when replying to the sensible knave, is forced to change tactics and attempt a reply from the intrinsic angle. The game-theoretic argument that I will make in favor of the rationality of justice must remain consistent throughout and draw only on the instrumental value that justice provides. I will initially show that the instrumental benefits alone are sufficient to entice a rational actor to behave in a just fashion, but at the end of this paper, Hume's virtue-ethical approach will reappear as I consider the possibility that there is an intrinsic value to justice that serves as an added benefit to an already sufficiently rational choice. In the meantime, our argument is best levied in a purely instrumental context, and Hobbes's instrumental approach provides a better vehicle for this project.

A Brief Note on Adam Smith

Thus far, I have mentioned Adam Smith only as a critic of Hume's account of justice and as the source of the concept of the invisible hand. However, Smith was an accomplished moral theorist in his own right, and, while his own theory of justice is reminiscent of Hume's, the argument that Smith makes in The Theory of Moral Sentiments is an improvement on Hume in

several respects. For example, Smith believes that justice is part of the basic human constitution and not a result of artifice and convention. He recognizes the innate sense of *fairness* as a motivating force for human behavior and he firmly believes that this sense leads to the desire for justice and is part of our primal nature.[77] He disagrees with Hume's contention that sympathy with the public good is behind our motive for justice partly because he recognizes that sympathy is erratic and unpredictable, while the concern for justice and fairness is constant.

It is also worth noting that Adam Smith can be considered a proto-behavioral economist.[78] Behavioral economics is a field of study that combines economic theory with behavioral psychology. It has grown in popularity over the past two decades due to its ability to explain decision making under uncertainty, and its recent contributions to our understanding of financial markets and investor behavior are immense. However, as I will explain in more detail in Chapter 4, behavioral economics will also be useful in analyzing decision making as it relates to the rationality of justice and the keeping of covenants. For now, it will suffice to acknowledge that some of the most important concepts in current behavioral economics were alluded to by Adam Smith in the eighteenth century. For example, Smith was aware of the problem of intertemporal choice, or the irrational tendency of individuals to have a preference for utility today over utility at some future date:

> The pleasure which we are to enjoy ten years hence, interests us so little in comparison with that which we may enjoy today, the passion which the first excites, is naturally so weak in comparison with that violent emotion which the second is apt to give occasion to, that the one could never be any balance to the other, unless it was supported by the sense of propriety…[79]

[77] Ashraf, Nava, Colin F. Camerer, and George Loewenstein. "Adam Smith, Behavioral Economist." *Journal of Economic Perspectives* 19, no. 3 (Summer 2005): p. 136
[78] See Ashraf, et al.
[79] Smith, *The Theory of Moral Sentiments*, IV.2.ii, p. 190

Another central tenet of behavioral economics is the human tendency to be overconfident. As I will explain in detail in Chapter 4, most individuals believe their abilities and their luck to be better than average. Smith was aware of this fact centuries before it was the subject of formal academic study: "The chance of gain is by every man more or less over-valued, and the chance of loss is by most men under-valued…"[80] This tendency leads to poor decisions in cases of uncertainty, and, as we will see, it leads Hobbes's Fool to believe his chances of evasion are better than they actually are.

For reasons given earlier, from this point forward it is preferable to pursue Hobbes's line of argument rather than Hume's, and since Smith's theory of morality and justice is far more reminiscent of Hume than of Hobbes, I will be leaving Smith aside for now. However, I will return to him in later chapters where his early contributions to behavioral finance theory will be of assistance in making the case against free riders and Hobbsean Fools.

Conclusion

So far I have conducted a rather broad historical enquiry into the topic of justice from which several important common themes have emerged. In particular, Socrates, Hobbes and Hume are in agreement that concepts of justice are dependent upon human nature, that justice is not a zero-sum game and that the instrumental benefits alone are sufficient to demonstrate that justice is a rational strategy for a self-interested actor.

From here the focus will become much narrower as I emphasize the nuances of the contractarian version of the argument. Hobbes's account of human nature and psychology may not be entirely accurate from a scientific standpoint, but it provides us with an adequate

[80] Smith, *The Wealth of Nations*, I, X, 1, p. 124

launching pad for a more sophisticated examination of the rationality of justice. In the next chapter, using Hobbes's description of human nature and justice as a starting point, I will examine how the neo-Hobbsean contractarians use concepts from the fields of game theory and financial economics to offer a more sophisticated argument in favor of the rationality of justice. Then, in Chapter Four I will use a contractarian argument to refute the claims made by Hobbes's Fool and other free-riders.

CHAPTER 3: GAME THEORY, DISPOSITIONS AND THE INSTRUMENTAL VALUE OF JUSTICE

Chapter 2 examined the Hobbesean and Humean accounts of justice and concluded that Hobbes's contractarian version provides a better framework for a more detailed enquiry into the topic. This chapter will present the contractarian philosophy of David Gauthier as a more refined version of Hobbes's theory. I will begin the chapter with a detailed discussion of Gauthier's most influential work, *Morals by Agreement,*[81] placing special emphasis on his use of game theory and competitive market theory to advance a novel argument in favor of the rationality of justice. I will then present an analysis and critique of Gauthier's claims, drawing on several of the contractarian commentators who have responded to Gauthier's compelling argument, as well as modifications to the original argument made by Gauthier himself in his more recent work. I conclude that Gauthier makes a valuable contribution to the advancement of contractarianism, most notably with his introduction of the concepts of competitive markets and game theory to the issue of justice and his recognition of decision-making at the level of metachoice. Gauthier will show us that, when choosing a deliberative procedure, an individual is actually making a rational choice *about how to make choices.*

[81] Gauthier (1986)

Gauthier's *Morals by Agreement*

Overview

The remainder of this essay will lean heavily on several of the ideas introduced in *Morals by Agreement* (hereafter *MbA*), so I will dedicate considerable effort to a careful description and analysis of Gauthier's argument in *MbA* and his modification of these ideas in subsequent works. In the preface to *MbA*, Gauthier proposes to address three core problems related to morality, advantage and justice. The first of these is the principle of rational cooperation. Stated in Hobbesean terms, he wants to demonstrate that it is rational for an individual to agree to principles of morality by accepting constraints on her own liberty. Gauthier wants to make a formal break with Hume in this regard, and he wants to argue that agreeing to accept moral constraints is not based on sympathy, but on reason: "If moral appeals are entitled to some practical effect, some influence on our behavior, it is not because they whisper invitingly to our desires, but because they convince our intellect."[82]

Second, Gauthier will argue that it is not only rational to agree to constrain one's behavior, but that it is also rational for an individual to comply with the agreement, or, as Hobbes would say, to "perform their covenants made."[83] It is important to note that Gauthier does not deny Glaucon's claim that "Someone who has the power to do (injustice), however, and is a true man, wouldn't make an agreement with anyone…"[84] Gauthier recognizes that any individual would prefer to avoid the constraints that justice requires, but he is arguing (again, in the Hobbsean tradition) that our own limitations, that is, our own inability to consistently and reliably get away with cheating on our agreement to constraint, makes compliance a rational

[82] Gauthier (1986), p. 1
[83] *Leviathan*, ch. 15, p. 89
[84] *The Republic*, 359b 1-3, p. 1000

strategy. His rational choice approach "allows us to state…why rational persons would agree *ex ante* to constraining principles…and why rational persons would comply *ex post* with the agreed constraints."[85]

Third, as a response to Rawls, he wants to introduce a basic concept of rights from which to propose an initial bargaining position for the formation of a social contract. This chapter and the next will focus on the first two of these core problems.[86] Gauthier will claim, and I will agree, that, "To choose rationally, one must choose morally."[87] Justice will initially be described by Gauthier as a *disposition* not to take advantage of others as long as others can reasonably be expected to be similarly disposed.

As noted in the prior two chapters, ancient and early modern theories of justice involved several consistent themes, including the idea that justice is not a zero-sum game, the need to understand underlying human nature and psychology, and the distinction between the intrinsic and instrumental value of justice. Gauthier employs each of these themes in a new light in *MbA*, and he uses each of them to advance his own argument in favor of the rationality of just behavior.

He employs the concept of free markets to illustrate that cooperation among individuals is not a zero-sum game. He invokes Smith's invisible hand (as well as a different and *visible* hand, to be explained below) to demonstrate that societal interaction offers individuals more than mere protection from physical harm. Interactions among humans provide tremendous benefits for all of the players involved, and justice is a necessary element of this interaction. Because

[85] Gauthier (1986), p. 10
[86] Chapters 7 -9 of *Morals by Agreement* deal with fairness and coercion in the initial bargaining position and justice with respect to the distribution of the economic benefits of cooperation. This section provides valuable commentary on the work of John Rawls and would be instrumental in facilitating a comparison of the work of Rawls, Nozik, and Sen on these topics, but it is beyond the scope of the current project.
[87] Gauthier (1986), p. 4

participation in human interaction is in our rational self-interest, Gauthier will argue that a disposition to observe the rules of justice, as a prerequisite to this interaction, is in our self-interest as well.

Gauthier makes extensive use of game theory and the prisoner's dilemma (PD) to advance his argument. He shows that, in the context of an imperfect market, a situation in which individuals agree to constrain their behavior delivers better results for everyone than does a situation in which each individual seeks to maximize his own individual utility. This results in a seemingly paradoxical outcome: constraint dominates individual advantage, yet the acceptance of this constraint is advantageous to each individual because those who are disposed to constraint (what Gauthier will call "constrained maximization") will enjoy opportunities for mutually beneficial cooperation that an uncooperative "straightforward maximizer" will fail to attain.

The concept of human nature and psychology that Gauthier will advance in *MbA* is closely intertwined with the distinction between the instrumental and intrinsic value of justice. As mentioned above, Gauthier's argument for the rationality of justice is based on market interaction among individuals. He therefore initially characterizes human nature in terms of a caricature that he calls "economic man." Economic man is willing to place constraints on his behavior as long as others do so as well, because his reason allows him to realize that it is in his best interest to do so. The psychology of economic man is such that he recognizes that it is rational for him to dispose himself to constrained maximization in order to increase his opportunities to benefit from cooperation.

Gauthier's economic man contrasts sharply with the Humean concept of humanity because economic man operates under conditions of mutual unconcern; he takes no interest in the interests of other individuals and he is not motivated by sympathy. Economic man therefore

requires individuals to mutually agree to constrain their behavior and to refrain from force and fraud against one-another. The beauty of Gauthier's constrained maximization is that it can convert this mutual unconcern into a mutual benefit through cooperation. The one unfortunate consequence is that economic man observes justice for instrumental reasons only; the intrinsic benefits of justice for its own sake are meaningless to him.

Obviously, Gauthier has to acknowledge that economic man is not an accurate representation of what humankind really is. Actual humans do take interest in one another's interests, and this is where the intrinsic value of justice arises. Gauthier claims that humankind's natural state is a mean between economic man (who values things only instrumentally) and utopian man (who values things only intrinsically). The "liberal individual" who is found in the mean between these extremes values things for both instrumental and intrinsic reasons, and it is in this context where the intrinsic value of justice must be sought.

Rational Choice

Before he begins the exposition of the central themes in *MbA*, Gauthier needs to establish a theory of rational choice. This aspect of Gauthier's argument will not be of central importance to this paper, but it is necessary to define it as a foundational concept before moving on to more pertinent topics. He will claim that reason is not concerned with the content of the preferences that an individual has; it is only concerned with the interrelations and measurement of particular preferences. In keeping with Hobbes's idea of value, Gauthier believes that value is a subjective and relative measure of individual preference, over which reason holds no sway. Good and evil are merely a matter of personal preference and no individual can rationally claim that the preferences of another are objectively wrong. Furthermore, Gauthier explicitly sides with Hume

with respect to Hume's claim that it is "not contrary to reason to prefer the destruction of the world to the scratching of my finger."[88] While such a preference may be considered insane due to a defect with the subject's affections, Gauthier wants us to recognize that to make such a choice is not irrational (if the subject in question actually has this insane ordering of preferences), nor is it arbitrary.

While reason is not concerned with the content of preferences, it is concerned with the measurement of those preferences. Gauthier's rational choice theory designates utility as the measure of preference and it states that reason leads a self-interested actor to attempt to maximize his own personal utility. He wants to be clear that he is distinguishing between a normative *evaluation* of preference and a *measure* of preference: "The theory of rational choice sets its course between the dogmatism of assuring a standard for preference and the scepticism of denying a measure of preference."[89] As we shall see, Gauthier makes frequent use of the concepts of rational choice and utility when constructing the central arguments of *MbA*. While his claims are certainly open to debate, I will not attempt to resolve any potential critique concerning the soundness of Gauthier's theory of rational choice here; in order to maintain the focus of the essay, I will accept this aspect of his argument and move on.

Game Theory and the Prisoner's Dilemma

Gauthier begins the core of his argument with an analysis of strategic rationality, or rationality in the context of interaction with other actors. His initial goal is to demonstrate that it is rational for a self-interested actor to agree to moral constraints provided that others agree to

[88] *Treatise*, 2:2.3.4, p. 267
[89] Gauthier (1986), p. 26

those constraints as well, and he uses a game-theoretic approach to accomplish this goal. He creates a theoretical model in which he assumes that certain ideal conditions apply. Specifically, he assumes that each person's choice is a rational response to the choices she expects others to make, that each person expects others' choices to be rational as well, and that each person expects her choices and expectations to be reflected in the expectations of others.[90] Armed with these (admittedly very strong) assumptions, he can move on to the heart of his game-theoretic analysis of the problem.

Gauthier's game-theoretic argument is based upon the distinction between equilibrium outcomes and Pareto-optimal outcomes. A (Nash) equilibrium outcome occurs when each of the players involved in a particular interaction is making the best move that he can for himself given the actions of the other players. That is, an equilibrium is a set of strategies in which each player maximizes his own utility given fixed expectations about the strategies of the other players; no player can do better for himself by unilaterally changing his own strategy. The equilibrium outcome is contrasted with the (Pareto) optimal outcome. A Pareto-optimal outcome occurs if and only if there is no other potential outcome affording some person a greater utility and no person a lesser utility.[91]

At first glance, optimality and equilibrium appear to be quite similar since they are both utility-maximizing. However, there are subtle differences between equilibrium and optimal outcomes that have profound implications for Gauthier's moral theory. In an equilibrium outcome, the emphasis is on inputs; the input strategy of each player gives her the maximum utility given the strategies of the other players. In an optimal outcome, the emphasis is on output payoffs; each player receives his maximum payoff given that no other payoff is decreased.

[90] See Gauthier (1986), p. 61
[91] See Gauthier (1986), p. 76

Equilibrium and optimal outcomes are both utility-maximizing, but an equilibrium outcome maximizes outcomes for each individual whereas an optimal outcome maximizes the group outcome while ensuring that no individual can be made better-off without making another individual worse-off.

The distinction between equilibrium and optimality is a focal point for Gauthier's moral theory because an optimization strategy is often not utility-maximizing on an individual level. That is, it may be argued that choosing an optimizing strategy is not strategically rational for a self-interested actor. Gauthier not only recognizes this distinction, he embraces it as the very basis for morality: "Moral theory is essentially the theory of optimizing constraints on utility-maximization."[92] While some theorists will reject the strategic rationality of optimization, Gauthier will argue that the choice of an optimizing strategy is rational because the actor who chooses such a strategy is doing so on the level of metachoice. That is, *she is making a rational choice about how to make choices.*

The classic example known as the Prisoner's Dilemma (PD) will help to clarify the link that Gauthier wants to draw between rational choice and optimality. In the situation postulated by the PD, the equilibrium outcome is mutual confession; regardless of what action his partner takes, each prisoner maximizes his own utility by confessing. The problem with this situation is that the equilibrium outcome is not optimal; each prisoner would have an outcome better than the equilibrium outcome if they both remained silent. Thus, the PD illustrates the conflict that can arise between optimizing strategies and utility-maximizing strategies. In the PD, two strategically rational utility maximizers acting independently will both confess, but this will result in a worse outcome for both of them than if they had adopted the supposedly irrational

[92] Gauthier (1986), p. 78

optimization strategy of mutual silence. That is, the players should rationally prefer the optimal outcome, but they will attain the equilibrium outcome unless they *cooperate*.

Gauthier uses international relations to illustrate the point. Sovereign nations are essentially in a Hobbsean state of nature with one-another. In this state of nature, individual maximizing behavior (such as violating the terms of a disarmament agreement), which is, by definition, rational, results in a sub-optimal outcome which is potentially disastrous to all players.[93] In this context, equilibrium strategies can be viewed as short-term oriented and based on maximization at the individual level, whereas optimal solutions are long-term oriented and based on maximization at the group level. The main thrust of Gauthier's project will be to demonstrate the utility-maximizing rationale for choosing an optimal strategy instead of an equilibrium strategy. In an ideal world with a perfectly competitive market, equilibrium and optimization will coincide without any action on the part of the individuals involved. However, in the real world, where market imperfections exist, cooperation among individuals is necessary for the promotion of optimization strategies. I will now address the ideal conditions prevailing in a perfect market before turning to the real world example of market imperfection and cooperation.

Competitive Markets

The perfectly competitive market is the antithesis of the PD and of Hobbes's state of nature. Under a perfectly competitive market model, private goods, free market activity, mutual unconcern, the absence of externalities and conditions of certainty in production and exchange

[93] See Gauthier, David. "Thomas Hobbes: Moral Theorist." *The Journal of Philosophy* 76, no. 10 (Oct. 1979): p. 551

are all assumed.[94] The resulting strategic situation is a non-zero-sum game in which value is created from mutually beneficial exchange without the need for an explicit agreement. Where the PD places the players in a situation in which equilibrium and optimality are mutually exclusive, the perfect market guarantees the coincidence of equilibrium and optimality: "Adam Smith's invisible hand is thus made visible by the economist's analysis…each individual, intending only her own gain, promotes the interest of society, in bringing about a mutually beneficial optimal outcome, even though this is no part of her intention."[95] That is, when self-interested individuals operate within the confines of a perfect market, the invisible hand serves, without the purposeful intent of any of the individual actors, to overcome Hobbes's state of nature and to bring equilibrium in-line with optimality. The need for a rational actor to make a meta-decision between equilibrium strategies and optimization strategies does not arise. It should be noted that Gauthier does not invoke the concept of cooperation at this point, because the presence of the invisible hand makes cooperation unnecessary in a perfect market. Cooperation only arises when some market imperfection makes it necessary for individuals to agree to refrain from force and fraud.

Gauthier's model of the perfect market implies that the market is a "morally free zone."[96] Where most advocates of laissez-faire capitalism claim that a perfect market in a state of equilibrium and optimality is morally right, Gauthier will instead claim that in such a market morality has no place at all. In a perfect market, choice is neither morally right nor wrong because the coincidence of equilibrium and optimality removes the need for the constraints that morality provides. Moral constraints appear only when market imperfections are present and a

[94] See Gauthier (1986), p. 89. These are all quite strong assumptions, as Gauthier readily acknowledges.
[95] Gauthier (1986), p. 89
[96] Gauthier (1986), p. 90

gap emerges between optimal mutual benefit and the rational pursuit of gain on an individual level. Under an imperfect market scenario morality is needed to in order to keep the optimal outcome from deteriorating into a less-desirable (on a group level) equilibrium outcome.

Not only is morality not required in Gauthier's perfect market, the individual actors in the market are non-tuistic, that is, they mutually exhibit no concern for one-another. It is important to note here that he is not claiming that actual people have no concern for their fellow humans; most certainly do. What he wants to show is that this mutual concern need not be present for optimality and equilibrium to coincide under the conditions of a perfect market. While an individual does have concern for family and friends, this same level of concern is not extended to acquaintances or other individuals with whom she may have infrequent contact. The beauty of the perfect market lies in the fact that the invisible hand renders this mutual unconcern ineffective in disrupting the coincidence of equilibrium and optimality.

Unfortunately for Gauthier and the rest of us, the real world is not a perfectly competitive market. As indicated above, one of the assumptions of Gauthier's perfect market model is the absence of externalities. A cursory examination of the facts as we find them will reveal that market externalities certainly do exist, most notably in the form of free-riders and parasites. A free-rider is an actor who enjoys a benefit without sharing in its cost, such as the guy in your office who drinks four cups of coffee each day but does not contribute to the coffee fund. A parasite is an actor who enjoys a benefit while passing-on all or part of the cost of that benefit to another party who does not enjoy the benefit. Goldman Sachs and other large investment banks are examples of parasitism. When they make risky wagers with their clients' capital, they have the potential to reap significant economic benefits for themselves and only themselves, but some of the costs they incur, in the form of systemic risk and periodic bailouts, are born by their clients

and the public. In the presence of free riders, parasites and other market externalities, the market will fail to align equilibrium and optimality, and this market failure gives rise to the need for morality as a constraint on action. This is where cooperation and justice emerge.

Cooperation and the Circumstances of Justice

When equilibrium and optimality fail to coincide, rational actors will respond by cooperating. It is at this point that Gauthier makes the most explicit declaration of his concept of justice:

> Where market interaction, with its pre-established harmony between equilibrium and optimum, is beyond good and evil, and natural interaction, in the presence of free riders and parasites, degenerates into force and fraud, cooperative interaction is the domain of justice. Justice is the disposition not to take advantage of one's fellows, not to seek free goods or to impose uncompensated costs, provided that one supposes others similarly disposed.[97]

Cooperation is not needed in a perfect market because free riders and parasites do not exist there. These externalities do exist in the Hobbesean state of nature, but rational cooperative action eliminates them. The result is justice: the rational disposition to agree to forego the opportunity for free ridership or parasitism in return for others foregoing the same.

Gauthier is in agreement with Hume and Rawls, who both recognize scarcity and individual bias as the circumstances of justice. He argues, however, that the presence of market externalities must be added to this list. A perfectly competitive market will eliminate most of the need for cooperation, because the invisible hand does most of the work itself. It is the presence of externalities (an imperfect market) that makes the *visible* hand of cooperation necessary: "the fundamental circumstances of justice, those features of the human situation that gave rise to co-

[97] Gauthier (1986), p. 113

operation, are awareness of externalities in our environment, and awareness of self bias in our character."[98] Hume and Rawls failed to appreciate the role of the perfect market in bringing about an optimal outcome. It is only when the perfect market becomes imperfect that the need for cooperation, and in turn, the need for justice, arises.

It is important to note here the distinction between cooperation and bargaining. Bargaining is an individual utility-maximizing strategy in which each person's behavior is a response to her expectations of the behavior of others in a zero-sum game. This is not what Gauthier's rational actor is engaged in. Cooperation, on the other hand, is a joint strategy in which each person's behavior is an attempt to optimize the collective outcome in a non-zero-sum game.[99] This is what the rational agent is striving for. When cooperating, individuals are acting in agreement with each-other with the joint goal of optimization, where the outcome will be determined by what Gauthier calls the principle of minimax relative concession.

The principle of minimax relative concession states that given a range of outcomes, each of which requires concessions by some or all persons, an outcome will be selected only if the maximum relative concession it requires from a single person is as small as possible.[100] This principle serves as a rational basis for an impartial accord in which each person agrees to restrain his self-interested maximizing behavior, provided other individuals agree to do likewise. Justice is merely the self-imposed disposition to abide by this self-imposed restraint once the restraint has been adopted. In Gauthier's words: "co-operation is the visible hand restraining persons from taking advantage of their fellows, but restraining them impartially and in a way beneficial to all. Such restraint commands rational acceptance; this is the idea underlying morals by

[98] Gauthier (1986), p. 116
[99] See Gauthier (1986), p. 129
[100] Gauthier's detailed derivation of minimax relative concession can be found in Gauthier (1986), p. 129-146.

agreement."[101] As we will now see, the conscious adoption of a *disposition* to restrain one's own behavior is the foundation upon which Gauthier's theory of justice rests.

The Disposition to Constrained Maximization

In arguing for the minimax principle, Gauthier claims that a rational utility-maximizing actor is better-off if she agrees to constraints on her behavior (provided that others agree to the same constraints) than she would be if she refused to accept any constraints at all. She rationally recognizes the superiority of joint optimal outcomes over individual equilibrium outcomes, and she chooses, on joint utility-maximizing grounds, not to make future decisions based on individual utility-maximizing grounds. This obscure concept is clarified by Gauthier's distinction between straightforward maximization (SM) and constrained maximization (CM).

A person practicing SM bases his decisions on the input strategies of others, that is, he wants to maximize his utility based on the strategies of the other players. He is akin to an individual in a Hobbsean state of nature. In contrast, an individual practicing CM bases her decisions on the output payoffs of others; she wants to maximize her utility given the utilities of the other players.[102] Unlike the SM, the CM is willing to take the benefits of the other players into account in an attempt to increase the utility of everyone, including herself.[103] Gauthier uses this distinction between SMs and CMs to make one of the most important claims in all of *MbA*. He states:

> In defending constrained maximization we have implicitly re-interpreted the utility-maximizing conception of practical rationality. The received interpretation, commonly accepted by economists…identifies rationality with utility-

[101] Gauthier (1986), p. 150 - 151
[102] See Gauthier (1986), p. 167
[103] It is important to note here that, while an individual who is practicing a SM strategy is in a situation similar to a Hobbesean state of nature, it does not follow that adopting a CM strategy must necessarily place an individual under the rule of an absolute sovereign, as Hobbes' account does.

> maximization at the level of particular choices. A choice is rational if and only if it maximizes the actor's expected utility. We identify rationality with utility-maximization at the level of dispositions to choose. A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition.[104]

This is a strong claim with profound implications. Gauthier is arguing that the relevant difference between SMs and CMs is in their dispositions. He wants to show that these differences in dispositions can be observed by others, and that the disposition of a CM is superior to that of a SM.[105]

Gauthier's claim that the disposition of a CM is superior to that of a SM is based simply on the now-familiar idea that cooperation is not a zero-sum-game. It is true that SMs can, on occasion, take advantage of CMs. However, as long as there is a sufficient number of CMs, the CMs will obtain numerous benefits from cooperative interaction with each other that are unavailable to SMs. That is, the benefits of being a CM are found not in the specific individual choices that a CM makes, but in the CM's wider range of opportunities to choose.[106]

The observability of dispositions is relevant because it will impact the ability of SMs to take advantage of CMs. That is, if a SM wants to take advantage of the group-regarding disposition of a CM, her likelihood of success will depend upon her ability to conceal her true disposition. It is rather obvious that dispositions are at least somewhat observable, but how exactly does the degree of observability impact the force of Gauthier's argument? If the dispositions of individuals were entirely transparent, deception would be impossible and CMs would avoid interactions with SMs. If dispositions were entirely opaque, government force

[104] Gauthier (1986), p. 182-183

[105] It is important to note that Gauthier has subsequently abandoned the notion of constrained maximization. In his most recent work, Gauthier acknowledges that his advocacy of CM was a mistake, and he instead invokes the concept of Pareto-optimal cooperation. Under this characterization, the reason that agents reach an optimal agreement is not because they want to maximize, it is because they want to cooperate; they see value in cooperation itself. See Gauthier, David. "Twenty-Five On." *Ethics* 123, No. 4 (July, 2013): 601-624

[106] See Gauthier (1986), p. 183

would be necessary to keep CMs from being taken advantage of by SMs. Gauthier observes that human dispositions are more aptly described as translucent. While we may not be able to always ascertain the intentions and dispositions of other actors, we can often approximate intentions and dispositions with more accuracy than random guessing. In addition, reason will require that a CM cultivate a talent for detecting the dispositions of others (although it is not obvious how a CM will accomplish this). The stronger this ability for detection becomes, the more valuable is the choice to adopt the disposition of a CM.

Gauthier is making three claims here: First, that the dispositions of individuals are detectable. Second, that most individuals are able to detect the dispositions of others with some level of accuracy, and third, that if most individuals can accurately detect the dispositions of others, then everyone has a reason to be a CM. The first and third claims are mostly uncontroversial, but the second claim, that most individuals are able to detect the dispositions of others with some accuracy, is the subject of much debate.[107] Various retorts to this aspect of Gauthier's argument have been offered for many years, most famously by Hobbes's Fool, who is essentially the epitome of Gauthier's SM. The Fool claims that an individual can maximize his utility by making a covenant, but that he does not necessarily always maximize his utility in keeping the covenant because the limited ability of other individuals to detect his true disposition will allow him to get away with periodic violations.[108] That is, the Fool wants to argue that it is rational to appear just, but to behave unjustly in select situations.

According to Gauthier, Hobbes, in his reply to the Fool, is missing the critical point that the rationality of keeping one's agreements is distinct from the rationality of being disposed to keep one's agreements. The Fool is condemning reason for a lack of benefit in performance, but

[107] I owe this characterization of Gauthier's claims to a personal conversation with Professor Hugh Lafollette.
[108] *Leviathan*, ch. 15, p. 91

he needs to be able to condemn reason for a lack of benefit in *disposition to perform.* Since the disposition of a CM provides advantages that the disposition of a SM cannot, and since dispositions are translucent, Gauthier argues that the Fool's argument is unsuccessful. A detailed analysis of The Fool's argument will occupy most of Chapter 4 of this essay, so I will leave the finer points of this argument until that time.

Economic Man, Utopian Man and the Liberal Individual

In his discussion of imperfect markets, CM, and dispositions, Gauthier uses an idealized concept of humankind to make his case. This ideal, which he refers to as "economic man," is non-tuistic (he takes no interest in the interests of others), he is constantly seeking to appropriate more utility for himself and he values justice only for instrumental reasons. For economic man, Gauthier's perfectly competitive market is the ideal societal situation.[109] Economic man has no concern for his fellows and he derives no pleasure and no value from interacting with others; for him, interaction is valuable only as a more efficient means to obtain more of the things he desires.

The problem with economic man is that he is merely a caricature; his nature is not like human nature as we actually find it. Real humans do take an interest in each-others' interests and they find interacting with one another a pleasurable activity in itself. Real persons have interests other than appropriating more goods for themselves and they experience diminishing utility as they appropriate more. Gauthier recognizes these obvious differences between economic man and actual humans, yet he maintains that morals by agreement are applicable to both of them.

[109] See Gauthier, David. "Rational Cooperation." *Nous* 8, no. 1 (March 1974): p. 62

Economic man is contrasted with utopian man. Where economic man values things only instrumentally, utopian man values them only intrinsically. Utopian man lives in a world without scarcity, and, as Hume argues[110] and Gauthier agrees, utopian man therefore has no need for justice. However, utopian man is no more human than economic man is, because utopian man has no need to seek or strive. Nietzsche recognized that an integral part of human life is the need to struggle towards higher goals, and Gauthier understands this:

> it is scarcity that gives rise to activities with instrumental value. If they are necessary to human fulfillment, then scarcity is necessary too. The idea of a human society based not on scarcity but on plentitude is chimerical; to overcome scarcity would be to overcome the conditions that give human life its point.[111]

Precisely because he lacks the experience of scarcity and the need to struggle, utopian man is no more the human ideal than economic man. His plentitude and lack of worry is no more applicable to the human condition than economic man's non-tuism and indifference to human affection. Gauthier believes that humans as we actually find them are located between these two extremes, and this caricature he calls the liberal individual. The liberal individual values things both instrumentally and intrinsically; he can engage his intellect in a market context as well as engage his affections in personal interaction with others. He finds justice in both contexts.

The liberal individual will play a significant role in the argument regarding the intrinsic value of justice to be addressed in Chapter 5, so I will not dwell on him here. The relevant point at this juncture is that the fact that economic man is not like humankind as we actually find it does not invalidate Gauthier's argument for morals by agreement. By using economic man as a caricature, Gauthier is trying to demonstrate that a disposition to justice is instrumentally

[110] *An Enquiry Concerning the Principles of Morals,* III, Part I, p. 21-22
[111] Gauthier (1986), p. 333

rational, and, as I will argue, he is mostly successful. The remaining question is, when we relax our assumptions to include a creature such as the liberal individual who is more in-line with humankind as we actually find it, how does this impact the argument in favor of just dispositions? That is, can it be demonstrated, in terms familiar to Gauthier, that justice is intrinsically valuable as well? In Chapter 5 I will address this question in detail.

Gauthier's Critics

Overview

The contractarian theory of morals and justice presented by Gauthier in *MbA* is both provocative and profound, and it demands a reply. The majority of the contractarian philosophers who have responded to Gauthier's claims have, for the most part, done so with respect for the brilliance of Gauthier's work, and at times even with assistance from Gauthier himself. In this section I will address some of the relevant points of contention over Gauthier's claims in *MbA*, beginning with some relatively minor quibbles over minimax concession and the assumption of mutual unconcern, then moving on to a more substantial critique of constrained maximization and Gauthier's emphasis of dispositions. I will claim that Gauthier's recognition of the importance of making decisions on the level of metachoice represents a major contribution to the contractarian argument for the rationality of justice, but that his failure to include an element of reciprocity is a material shortcoming. As mentioned earlier, I will defer my discussion of Gauthier's treatment of Hobbes's Fool until Chapter 4, and Gauthier's liberal individual will appear again in Chapter 5.

Mutual Unconcern and the Minimax

The assumption of mutual unconcern is probably the easiest aspect of Gauthier's argument to challenge, simply because it is so obviously and so patently false. As Gauthier readily acknowledges, real people do take an interest in one-another's interests, and economic man is a poor estimate of what humans are actually like. Clearly, the assumption of non-tuism does not hold empirically, but to what extent does this diminish Gauthier's larger project?

Vallentyne[112] argues that, since people's actual preferences are not mutually unconcerned, an agreement that is rational under an assumption of non-tuism is not necessarily rational when this strong assumption is lifted and the actual preferences of other-regarding individuals are considered. When postulating a system of rational constraints, he wants to avoid making any assumptions at all about preferences and simply consider the actual preferences that individuals would have. Essentially, he believes Gauthier needs to drop the assumption of non-tuism if he wants to have a coherent theory of rational interaction.

While the central point of Vallentyne's critique is true, it does not pose a significant threat to the strength of Gauthier's argument. The fact that Gauthier's rational actors are non-tuistic when agreeing to CM does not mean that relaxing the non-tuistic assumption would result in a rational agreement other than CM. Gauthier uses the non-tuistic assumption, not because it helps him argue that rational actors would adopt CM, but because it allows him to include real-life situations that are non-tuistic within the confines of CM. Real humans *are* non-tuistic in many situations involving justice; outside of our relatively small group of friends and family, we often take little or no interest in the interests of others when interacting with them, especially when these interactions are impersonal in nature. When we pay for a transaction with an

[112] Vallentyne, Peter. "Contractarianism and the assumption of mutual unconcern." in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 71-75. New York: Cambridge University Press, 1991

unknown individual over the internet, faithfully pay our taxes, or keep current on the mortgage that we owe to a parasitic banking institution, we usually do not take an interest in the other party's interest. Gauthier would say that we are merely acting on a prior disposition to CM, without regard for the other party. The assumption of non-tuism merely allows Gauthier to incorporate these infrequent and less personal interactions under the umbrella of CM. Furthermore, his characterization of the liberal individual in the final chapter of *MbA* allows him to acknowledge and account for the fellow-feelings that we have for others. Gauthier certainly does recognize that humans are tuistic, but in the context of CM, they do not need to be.

Another element of Gauthier's project that invites criticism is the principle of minimax relative concession. As outlined above, the principle of minimax relative concession states that given a range of outcomes, each of which requires concessions by some or all persons, an outcome will be selected only if the maximum relative concession it requires from a single person is as small as possible. The principle serves as a rational basis for an impartial accord in which each person agrees to constrain his behavior. One might reasonably question whether this outcome would actually be selected given Gauthier's assumptions.

Kavka, in his review of *MbA*,[113] claims that the minimax principle would not be chosen because it rewards tendencies that our common sense intuitions find blameworthy and punishes tendencies that we would commend. The minimax principle states that each actor will minimize the *relative* concession that she has to make. Since some gluttonous individuals have a psychological tendency to derive a significant amount of utility simply from having more of the cooperative surplus than others have, in order to reach agreement under the minimax principle, the relative concession of these individuals will have to be smaller than the relative concession of

[113] Kavka, Gregory. "Morals by Agreement, by David Gauthier." *Mind*, New Series 96, no. 381 (Jan, 1987): 117-121.

individuals who have a psychological tendency to a more equal distribution. In other words, greed is rewarded and parity is punished.

Kavka's critique of the minimax principle is insightful and correct. It is also relevant to Gauthier's project as it relates to social justice and the justice of distributive shares, and it is a critique that Gauthier needs to address.[114] However, in this paper I am presenting an argument in favor of the rationality of justice in the Hobbsean sense of keeping covenants made, and this particular observation of Kavka's is not relevant to Gauthier's argument that forming a disposition to CM is a rational strategy. The minimax principle could be replaced with any number of other principles for dividing the cooperative surplus without diminishing the force of Gauthier's core argument in this regard.

Dispositions

Although cogent objections can be raised regarding the minimax principle and the assumption of non-tuism, these objections do little to undermine the force of Gauthier's theory. However, his claims regarding the role of dispositions are of central importance to the coherence of his entire argument in *MbA*, and any flaws in this portion of his theory must be addressed. Recall the passage cited above:

> We identify rationality with utility-maximization at the level of dispositions to choose. A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition…She benefits from her disposition, not in the choices she makes, but in her opportunities to choose.[115]

[114] For a contractarian alternative to Gauthier's minimax principle, see Hampton, Jean. "Equalizing concessions in the pursuit of justice: A discussion of Gauthier's bargaining solution." in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 149-161. New York: Cambridge University Press, 1991..
[115] Gauthier (1986), p. 182-183

Gauthier is making three claims regarding dispositions. First, he claims that humans have the ability to choose dispositions in general. Second, assuming that the first claim holds, it is rational to choose a disposition to CM. Third, once we choose the disposition to CM, it is rational to comply with that disposition. I will address the first and third claims in this section and the second claim will be addressed in the section on CM below.

The first claim is really a claim about human psychology. Gauthier seems to simply assume that we have complete control over our choices and that we can choose to be automatically disposed to act in a particular way across a wide variety of situations. This is especially peculiar in light of the fact that he claims to agree with Hume's contention that reason is subservient to passion: "Desire, not thought, and volition, not cognition, are the springs of good and evil."[116] In her commentary on Gauthier,[117] Hampton recognizes that the ability to will ourselves to be in accord with a particular disposition is at odds with Hobbesean psychology, and possibly at odds with contemporary moral psychology as well.[118] Given the importance of the choice of dispositions to his argument, this challenge needs to be addressed if Gauthier's theory is to have any force at all.

In Chapter 1 I emphasized the importance of a dispositional account of justice as characterized by Aristotle. Aristotle believes that virtue is found, not in particular acts of virtue, but in a permanent state of character that disposes an individual to act in a virtuous way

[116] Gauthier (1986), p. 21

[117] Hampton, Jean. "Two faces of contractarian thought." in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 31-55. New York: Cambridge University Press, 1991, p. 41

[118] The situationist account offered by Doris highlights the potential problem with Gauthier's version of dispositions. Through the use of empirical studies in the field of moral psychology, Doris argues that the variability that we observe between the ethical behavior of individuals is more a function of differing situations than of differing dispositions between individuals. He claims that any consistency in ethical behavior that we do observe is likely attributable to the fact that the situations in which an individual finds herself are consistent from day-to-day, rather than to any consistent and enduring ethical disposition. In short, Doris would doubt that we can choose a particular disposition as Gauthier claims, because he does not believe that individuals possess enduring ethical dispositions at all. See Doris, John. "Persons, Situations, and Virtue Ethics." *Nous* 32, no. 4 (Dec, 1998): 504-530.

regardless of the benefits it brings in a particular situation. That is, Aristotle and Gauthier both recognize that decisions regarding justice should be made, not on the level of particular individual choices, but on a meta-level; a disposition is a choice about how to make choices.

Given this apparent similarity between Aristotle and Gauthier, it is tempting to invoke Aristotle's account of dispositions in support of Gauthier's account, but this attempt will fail. Gauthier's account of dispositions is quite different from that of Aristotle in that Gauthier's account implies the automatic application of a principle whereas Aristotle's is based entirely on sound judgment. Aristotle believes that the ethical situations in which we find ourselves will vary so greatly that we will never be able to come up with global rules that will apply in every situation. Instead of rules, Aristotle wants to rely on sound judgment to guide our actions in particular situations, and he believes this judgment is the result of strong moral education and a self-reinforcing combination of dispositions and actions over many years of practice. In contrast, Gauthier's account is based on the explicit choice of a disposition which will be automatically applied without further thought in a consistent manner across a wide variety of situations. Aristotle will be of no help to him in this regard.

Fortunately, in subsequent writings Gauthier has recognized the shortcomings of the dispositional account given in *MbA*, and he has replaced the dispositional account with an account based upon "deliberative procedures." Deliberative procedures are more like rules for decision making than dispositions, and they "are rational if and only if the effect of employing them is maximally conducive to one's life going as well as possible, where this effect includes, not only the actions they determine, but also the actions they make possible."[119] Gauthier originally invoked the concept of dispositions because it allowed him to emphasize the

[119] Gauthier, David. "Assure and Threaten." *Ethics*, no. 104 (July 1994): 690-721.

importance of making decisions based not upon immediate single outcomes, but upon the larger universe of outcomes that such decisions make possible. He wants to make a choice about how to make choices. The notion of deliberative procedures allows him to attain the same goal without having to make a controversial claim about human psychology and our ability to program ourselves to act in a certain way. We may not be able to intentionally choose a permanent disposition to automatically behave in a specific, consistent manner, but we certainly can choose to adopt a particular deliberative procedure to be applied across a variety of situations.

As mentioned at the beginning of this section, the third claim Gauthier makes regarding dispositions in *MbA* is that once we choose a disposition to CM, it is rational to comply with that disposition. Having abandoned the dispositional account in favor of an account based on deliberative procedures, he must now show that once we choose a particular deliberative procedure, it is rational to comply with it. It has been observed by several critics that adopting a policy of CM (whether via a disposition or a deliberative procedure) may be rational at the same time that the individual actions the policy proposes are not.[120] Again, due to the key role that dispositions and deliberative procedures play in Gauthier's philosophy, this challenge must be addressed.

In "Assure and Threaten," Gauthier makes an attempt to bridge the gap between the rationality of choosing a deliberative procedure and the rationality of observing it: "Deliberative procedures are rational if and only if the effect of employing them is maximally conducive to one's life going as well as possible, where this effect includes, not only the actions they

[120] See Kavka (1987), Yi, Byeong-Uk. "Rationality and the Prisoner's Dilemma in David Gauthier's Morals by Agreement." *The Journal of Philosophy* 89, no. 9 (Sept. 1992): 484-495 and Finkelstein, Claire. "Pragmatic Rationality and Risk." *Ethics* 123, No. 4 (July, 2013): 673-699

determine, but also the actions they make possible."[121] He is attempting to show that, if the general application of a deliberative procedure is rational, then each instance in which that procedure is employed is also rational, even if, in a particular instance, the employment of the principle is not conducive to one's life going as well as possible.

There is definitely a strange connection between the choice to adopt a deliberative procedure and the choice to comply with that procedure in any one particular instance, but Gauthier has still failed to demonstrate that the rationality of the former necessarily implies the rationality of the latter. The problem that the critics cite is the lack of a necessary connection between the deliberative procedure and the individual acts it demands.

Aristotle recognized this gap and attempted to bridge it via his concept of dispositions. Aristotle claimed that merely having a disposition implies that you have complied and will continue to comply with it, because the disposition owes its existence to the sum of the individual acts. That is, if you decide that the disposition is rational and you make the choice to adopt the disposition, the particular acts that are in accord with the disposition are necessarily rational because they are helping to form and reinforce the disposition in a virtuous cycle. For Aristotle, you cannot have the disposition without a consistent habit of acting in accordance with it, and, furthermore, if you view compliance with the disposition as being subject to violation on a case-by-case basis, the lack of performance will lead to the eventual loss of the disposition.

While it may once again be tempting to invoke Aristotle in order to assist Gauthier, as mentioned above, Aristotle's concept of dispositions is quite different from Gauthier's deliberative procedures. The continued application of a rule across a variety of situations will not help to reinforce the choice of that rule in the way that Aristotle's dispositions are self-

[121] Gauthier (1994), p. 701

reinforcing. In addition, even in the context of Aristotle's own account, it is not evident that every individual act of justice is a necessary element of the formation of a disposition to justice. It is possible that some individual acts that would be recommended by a disposition to (or a deliberative procedure of) justice could be omitted without necessarily damaging the disposition (or weakening the deliberative procedure). What Gauthier needs to show is that there is a necessary benefit emanating from the performance of each individual act done in accordance with a given deliberative procedure.

Essentially, Gauthier is faced with a problem of cross-temporal consistency in behavior. Although the general application of a deliberative procedure may clearly benefit an individual via access to a wider and superior array of opportunities, the application of the principle in a specific situation may not, in-and-of itself, lead to the individual's life going as well as possible. What Gauthier has overlooked is the possibility that cross-temporal consistency itself is an element of one's life going well.[122] To engage in actions that are inconsistent with an important deliberative procedure that one has consciously adopted is "a serious psychological fracture"[123] that can have a significant negative impact on one's self-image and sense of stability in identity. That is, the agent derives value from the consistency of her behavior, and any deviation from the adopted procedure can lead to undesirable psychological strife.[124]

Furthermore, it is important to keep in mind the main point behind Gauthier's account of dispositions and deliberative procedures: When an individual adopts a deliberative procedure, she is making a choice about how to make choices. The act of adopting the procedure itself

[122] See Bratman, Michael. "The Interplay of Intention and Reason." *Ethics* 123, no. 4 (July 2013): 657-672
[123] Bratman, p. 667
[124] Doris (1998) cites empirical studies as evidence of a lack of consistency in ethical behavior. While it may be true that individuals often behave in an ethically inconsistent manner, this does not imply that they have not chosen a deliberative procedure or that the choice of and adherence to the procedure is in any way irrational. It is important to recognize that many of the individuals in these studies likely felt remorse after they failed to perform, because the way they acted is inconsistent with the deliberative procedure they had previously chosen.

implies that the procedure will be applied in all cases; that is, compliance is included in the rule.[125] Thus, any inconsistency in applying the procedure is either a hindrance to a well-ordered life or evidence that the procedure was not really adopted in the first place. Gauthier's third claim is a viable one: Once we adopt a rational deliberative procedure it, is rational to adhere to it in all relevant situations.

Game Theory and Constrained Maximization

As indicated above, Gauthier makes three claims regarding dispositions in *MbA*: that humans have the ability to willingly choose dispositions in general, that it is rational to choose a disposition to CM, and that once we choose the disposition to CM, it is rational to comply with that disposition. In the previous section I addressed the first and third claims and offered rebuttals to the critics. These two claims remain intact, although in a modified format. Individuals do have the ability to consciously adopt rational deliberative principles (as opposed to dispositions), and it is rational for them to adhere to these rational principles in all relevant situations. Gauthier's second claim, however, is more problematic than the other two, and the shortcomings of constrained maximization as a rational principle are far more damaging to Gauthier's project.

At this point it is important to clarify the issue at stake. I have already accepted the claim that it is rational to consistently adhere to a rationally chosen deliberative procedure, but the question that remains is, "What kind of deliberative procedure is it rational to choose?" When selecting a deliberative procedure, we are making a choice about how to make choices, but this

[125] This notion of consistency in application of a procedure is in opposition to Doris' situationist account. For a refutation of Doris' claims in this regard, see Annas, Julia. "Comments on John Doris' 'Lack of Character'." *Philosophy and Phenomenological Research* 71, no. 3 (Nov. 2005): 636-642.

leaves open the possibility of choosing a deliberative procedure that advocates violating covenants when it is advantageous to do so. That is, if an agent behaves unjustly by violating a covenant, she is not necessarily in violation of her deliberative procedure; her procedure may actually call for opportunistic violation of covenants. In *MbA*, Gauthier claims that choosing a disposition to CM is rational. In his subsequent work, he abandons dispositions in favor of deliberative procedures, and he rejects CM and replaces it with agreed Pareto-optimization (hereafter APO).[126] Despite these revisions, the problem remains; he still needs to demonstrate that adopting a deliberative procedure of APO is a rational choice, and this is where his most significant problem lies.

The controversy surrounding APO is most clearly illustrated by contrasting APO with the ideas posited by evolutionary game theory (EGT). I will begin with a brief outline of EGT and I will attempt to reconcile the insights of EGT with the critiques of some of Gauthier's fellow contractarians. The conclusion will be that APO is an untenable principle.

The treatment of morality in EGT differs from contractarianism in a number of ways, but the most obvious is that EGT does not need to assume an explicit or implicit contract. In EGT, morality is not the result of a bargaining process, but is instead merely a side-effect of repeated interaction between agents who habitually act in a certain way.[127] The foundational example of EGT in the field of ethics is given by Axelrod.[128] His work is based upon a contest in which he challenged scientists, philosophers and hobbyists to come up with an optimal strategy in a multi-

[126] See Gauthier (2013)
[127] Verbeek, Bruno and Christopher Morris. "Game Theory and Ethics." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Summer 2010 Edition),
URL = <http://plato.stanford.edu/archives/sum2010/entries/game-ethics/>.
[128] Axelrod, Robert, and William Hamilton. "The Evolution of Cooperation." *Science*, New Series 211, no. 4489 (March 27, 1981): 1390-1396, and Axelrod, Robert. "The Emergence of Cooperation among Egoists." *The American Political Science Review* 75, no. 2 (June, 1981): 306-318

round (iterated) PD game. Axelrod insists on using an iterated PD game because he believes that repeated interaction is necessary for any form of cooperation to take hold; if the individual players know they will never interact again or if the number of future interactions is known, mutual defection is an evolutionarily stable strategy and the process will never move past this point.

The winner of Axelrod's contest was a simple strategy of cooperation based on reciprocity known as tit-for-tat (TFT), in which the player begins by cooperating on the first move, then continuing to cooperate on the second move if the other player cooperates in response, or defecting on the second move if the other player defects, and so on in subsequent moves. The story of morality that emerges from this game begins with Axelrod's observance that Hobbes's state of nature is equivalent to a PD game in which mutual defection (ALL D) is stable. The introduction of cooperation based on reciprocity (due to kinship or clustering) into this stable state allows the TFT strategy to take hold and become dominant if the probabilities of repeated interaction and potential retaliation are sufficiently high. Axelrod's conclusion is that cooperation can emerge even from an initial position akin to Hobbes's state of nature as long as small clusters of cooperators have even a small chance of interacting with one another. The collective stability of a cooperative strategy is stronger than that of mutual defection strategy, however, "for a (cooperative) strategy to be stable in the collective sense, it must be *provocable*. So, mutual cooperation can emerge in a world of egoists without central control, by starting with a cluster of individuals who rely on *reciprocity*." (emphasis added)[129]

On the surface, Axelrod's conclusions appear to mirror those of Gauthier: both present a scenario in which mutually beneficial cooperative interaction between individuals allows them to

[129] Axelrod (1981B), p. 317

escape the Hobbsean state of nature. However, two important differences are present. First, the link between rationality and justice is far weaker in EGT than in CM or APO. The EGT approach does not invoke a contract and it makes no assumption of full rationality or transparency in the way that CM and APO do. That is, in EGT rationality is not providing us with a reason to be moral or just; morality in this context is just an accidental outcome of interaction. If we understand justice as the keeping of covenants made, we cannot make an argument in favor of the rationality of justice under EGT because there is no contract to be kept. EGT must remain silent on the issue of justice in contractarian terms.

Second, there is a profound difference between CM / APO and EGT regarding the importance of reciprocity. A CM or APO considers it rational to cooperate even in situations where defection will not hinder future opportunities to interact and where there is no expectation of future interaction with the other party. This is in direct opposition to the TFT strategy, which has reciprocity as one of its main tenets. In *MbA* Gauthier specifically addresses this difference[130] and goes to significant effort to distinguish CM from TFT and to deny that CM is in any way dependent upon reciprocity. He wants to argue that if I dispose myself to cooperating, others will be aware of my disposition and this awareness by others will provide me with greater opportunities for future cooperative benefits, regardless of whether the other party cooperates: "It is rational to act in a mutually advantageous way in PD situations, if one gains more from one's own disposition to constraint, than one loses from one's actual exercise of constraint….It has nothing to do with mutuality."[131] In his works subsequent to *MbA*, although he has abandoned dispositions and CM, he continues to deny the necessity of reciprocity in advocating the choice

[130] See Gauthier (1986), pp. 169-170, footnote 19
[131] Gauthier, David. "Rational constraint: Some last words," in *Contractarianism and Rational Choice,* edited by Peter Vallentyne, 323-330. New York: Cambridge University Press, 1991, p. 327

of APO as a deliberative procedure.[132] Gauthier insists that reciprocity does not matter for an individual who chooses CM /APO because her permanent disposition frees her from having to consider the actions of the other party in any single interaction.

Because the aim of this dissertation is to link rationality and justice, it is tempting to jettison EGT for its inability to connect the two, and instead support Gauthier's APO. Unfortunately, Gauthier's peculiar aversion to reciprocity makes this an untenable course. Gauthier is simply (and inexplicably) trying too hard to deny the importance of reciprocity and mutuality in constructing a strategy of rational interaction, and this denial significantly diminishes the force of his argument.[133] Fortunately, Danielson[134] offers a solution that introduces reciprocity into the mix without significantly altering Gauthier's theory.

Gauthier claims that reciprocity does not matter because if I dispose myself to CM instead of SM I do not need to consider the actions of the other party.[135] Danielson observes that Gauthier is oversimplifying when he limits his PD strategies to just CM and SM, and Danielson rejects the disposition to CM in favor of a disposition that allows an actor to increase her utility by taking the need for reciprocity into account. He proposes an alternative strategy, reciprocal cooperation (RC), in which party 1 cooperates only when cooperation is necessary and sufficient to elicit the cooperation of party 2. RC differs from CM in that RC recognizes that if party 2 is an unconditional cooperator (he will cooperate no matter what), it is rational for party 1 to defect

[132] See Gauthier (1994)
[133] It should be noted that Gauthier's avoidance of reciprocity is a significant departure from Hobbes' second law of nature as well. See *Leviathan*, ch. 14, p. 80.
[134] See Danielson, Peter. "Closing the compliance dilemma: How it's rational to be moral in a Lamarckian world." in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 291-322. New York: Cambridge University Press, 1991.
[135] Although Danielson's essay was written prior to Gauthier's rejection of dispositions and CM, Danielson's critique applies equally well to APO and deliberative procedures.

because her cooperation is not necessary to elicit the cooperation of party 2. With this small change, Danielson is able to include reciprocity within a system of morals by agreement.

To summarize, the indispensable point that Gauthier is missing is this: Reciprocity matters. He is trying too hard to avoid making his argument dependent upon reciprocity and repeated interaction, even though inclusion of these elements will not diminish the moral force of his argument. As I will show in Chapter 4, the fusion of justice and rationality does need to be based on repeated interactions and the threat of retaliation. Fortunately, Danielson's contribution will allow us to continue in this regard. Danielson agrees with Gauthier's claim that it is both rational and possible to adopt a deliberative procedure of cooperation. He only disagrees with Gauthier's narrow definition of the conditions that make such a procedure rational. And, unlike Axelrod, Danielson does posit a contract and he does aim to give us a reason to be just. It is therefore possible to use Danielson's modification of Gauthier's philosophy to advance the argument that justice is instrumentally valuable to a rational actor.

Conclusion

Despite its flaws, Gauthier's work makes a considerable contribution to the argument in favor of the instrumental rationality of justice. In *Morals by Agreement* and his other works, Gauthier advances the contractarian theories of Hobbes by including competitive markets and game theory to explicitly demonstrate that rational cooperation is not a zero-sum game, and by including the idea of deliberative procedures in his account of justice. His emphasis on the distinction between optimality and equilibrium leads to his most important contribution to the argument; that an individual who chooses to adopt an optimizing strategy is making a decision

on the level of metachoice. That is, when choosing a deliberative strategy, an individual is making a rational choice *about how to make choices*.

Recall that Gauthier is now making three claims regarding deliberative procedures. First, humans have the ability to willingly choose deliberative procedures in general. Second, assuming that the first claim holds, it is rational to choose a deliberative procedure in accordance with APO. Third, once we choose the deliberative procedure in accordance with APO, it is rational to comply with that procedure in all instances. The first and third claims are sound, but this is insufficient to prove his point. If an individual decides that a deliberative procedure is rational (irrespective of whether it is a deliberative procedure in accordance with APO) and she makes the decision to adopt the procedure, the particular acts that are in accord with the procedure are also rational because the act of adopting the procedure itself implies that the procedure will be applied in all cases, and because any inconsistency in applying the procedure is a potential source of psychological strife and a hindrance to a well-ordered life.

The real question that contractarian philosophy needs to address is not whether it is rational to adhere to a deliberative procedure once you have chosen it; I have shown that adherence to a rational deliberative procedure is rational. The question that must be addressed is, "what kind of deliberative procedure is it rational to choose?" As Gauthier emphasizes repeatedly, this is a choice *about how to make choices*. When deciding on which deliberative procedure to choose, it must be acknowledged that reciprocity matters, as does repeated interaction. Both Danielson (explicitly) and Axelrod (implicitly) agree with Gauthier that it is rational to choose a deliberative procedure, but they disagree with him in that they both claim (correctly) that reciprocity is a necessary element of a rational choice of deliberative procedure. CM /APO will fail in this regard, and they must be replaced by a deliberative procedure that

includes reciprocity. But the question remains: Is it rational to choose a procedure that advocates adherence to covenants or is it rational to choose a procedure of violating covenants when it is advantageous to do so? Hobbes's Fool will argue in favor of opportunistic violation, and the response to the Fool's argument will be the topic of the next chapter.

CHAPTER 4: BEHAVIORAL ECONOMICS AND THE HOBBESEAN FOOL

Chapter 3 began with an exegesis of Gauthier's contractarian philosophy and ended with the conclusion that adherence to a rationally chosen deliberative procedure is rational and necessary. However, adhering to a deliberative procedure is not the same as adhering to a covenant. We still need to address the claim that it is rational to choose and to adhere to a deliberative procedure which dictates that one should violate covenants when it is advantageous to do so. That is, we need to compare the rationality of a deliberative procedure which recommends that we observe the covenants we make versus that of a deliberative procedure which recommends that we break covenants when it is convenient. Gauthier showed that choosing a deliberative procedure is a choice about how to make choices. This chapter will be dedicated to answering the question, "What type of deliberative procedure is it rational to choose?"

Hobbes's Fool will argue that it is indeed rational to choose a deliberative procedure to violate covenants when it is advantageous to do so. I will argue that the Fool is mistaken. This chapter will begin with a statement of the Fool's claims and a discussion of Hobbes's reply. Not only has there been a lack of agreement among scholars regarding whether Hobbes or the Fool advances the better argument, scholars have not even been able to agree on what exactly Hobbes and his Fool are claiming. I will therefore attempt to be quite explicit and specific regarding exactly what the Fool is claiming and what Hobbes is claiming in reply.

Next, I will address the work of several contemporary philosophers who have commented on Hobbes's reply to the Fool. Hampton interprets Hobbes as arguing that individuals reason in

a shortsighted, case-by-case manner, and this gives rise to conflict in the state of nature. Interpreted in this way, Hobbes will have a difficult time refuting the claims of the Fool. Hoekstra claims that Hobbes's argument against the Fool is levied only against the Fool who denies justice outwardly. He interprets Hobbes as agreeing with the Fool who denies justice only in his heart. Sayre-McCord posits the existence of "transopaque egoists" who enjoy an advantage similar to that of the Ring of Gyges. The Fool argues that it is rational for such individuals to violate covenants when doing so is to their advantage, and Sayre-McCord believes that Hobbes is unable to refute this claim. Gauthier argues against the Fool from two distinct angles. In *Moral Dealing*[136], he claims that Hobbes, in his argument for the second law of nature, wants to replace the natural reason of self-preservation with a conventional reason of peace. By re-characterizing the use of reason in the observance of covenants, Gauthier hopes to allow Hobbes to refute the claims of the Fool. In *Morals by Agreement*, he argues that the Fool fails to recognize that acceptance of Hobbes's second law of nature implies acceptance of the third law of nature, and the Fool is therefore mistaken.

Each of these accounts of Hobbes's reply to the Fool offers a unique and insightful perspective, but they are all ultimately unsatisfying. The most convincing argument against the Fool, by far, is levied by Kavka. He claims, and I agree, that the reply to the Fool should be based upon the presence of uncertainty and error in human reasoning. The Fool argues that we should violate covenants when we have a reasonable expectation of doing so without being detected. The core problem with this argument lies in the human inability to subjectively determine the expectation of success. When we humans attempt to calculate the probability of an outcome under uncertainty, we typically reason in a flawed manner. The world is highly

[136] Gauthier, David. *Moral Dealing: Contract, Ethics and Reason.* Ithaca: Cornell University Press, 1990

complex and unpredictable, and uncertainty and error are so widespread that an individual cannot reliably determine when breaking a covenant will be beneficial.[137]

I will apply several insights from the field of behavioral economics to show how human psychology makes us ill-suited to the task of judgment under uncertainty. We humans tend to be overconfident[138] in our ability to perform tasks and to calculate probability and we tend to be overly-optimistic in our assessment of possible outcomes. We envision ourselves as having more control over our environment than we actually have. We favor immediate benefits with limited utility over future benefits that have much higher utility, and, most important, we significantly underestimate the role that randomness plays in our lives. I will argue that a rational individual will choose a deliberative procedure of behaving in a just fashion and keeping the covenants she makes. She will recognize that due to these flaws in human psychology she is unable to properly assess the probability of successful violation on a case-by-case basis and that, given a sufficiently long time period, all covenant-breakers will eventually be discovered. In other words, the rational individual will recognize that "time wounds all heels."[139]

The Fool's Claim and Hobbes's Reply

The Fool makes his claim in Chapter XV of *Leviathan*. Hobbes states:

> The fool hath said in his heart: 'there is no such thing as justice'; and sometimes also with his tongue, seriously alleging that: 'every man's conservation and contentment being committed to his own care, there could be no reason why every man might not do what he thought conduced thereunto, and therefore also to make

[137] Kavka, Gregory. "The Rationality of Rule-Following: Hobbes' Dispute with the Fool." *Law and Philosophy* 14, no. 1 (Feb. 1995): 34.
[138] What behavioral economists refer to as "overconfidence," Hobbes would likely refer to as "pride" or "vain-glory."
[139] Kavka (1995), p. 27. This phrase is originally attributed to Groucho Marx.

> or not make, keep or not keep, covenants was not against reason, when it conduced to one's benefit.'[140]

The Fool argues that if humans by nature act in their own rational self interest, an individual's decision whether to keep or break a specific covenant should depend upon whether or not the breaking of the covenant will provide a benefit to that individual. The Fool is simply claiming that a rational individual will take advantage of situations in which the violation of a covenant will bring more reward than the keeping of the covenant.

Gauthier emphasizes the fact that the Fool is not arguing against Hobbes's second law of nature; he is arguing against the third.[141] That is, the Fool wants to show that, while it is in one's rational self-interest to maximize utility by entering into a covenant, it does not necessarily follow that one maximizes one's utility by keeping the covenant. The Fool believes that utility is maximized by observing the third law of nature only selectively. Like Thrasymachus before him, the Fool is arguing in the tradition of Antiphon that, "A person would make most advantage of justice for himself if he treated the laws as important in the presence of witnesses, and treated the decrees of nature as important when alone and with no witnesses present."[142]

Hobbes's reply to the Fool is immediate and direct:

> This specious reasoning is nevertheless false. For the question is not of promises mutual where there is no security of performance on either side (as when there is no civil power erected over the parties promising), for such promises are no covenants, but either where one of the parties has performed already, or where there is a power to make him perform, there is the question whether it be against reason, that is, against the benefit of the other to perform or not. And I say it is not against reason.[143]

Hobbes qualifies his remarks by reminding us that there are no covenants in the state of nature. He is concerned with situations in which a mechanism of punishment for the non-performance of

[140] *Leviathan*, ch. 15, p. 90
[141] Gauthier (1986), p. 161
[142] Antiphon, Diels-Krans 87 B44
[143] *Leviathan*, ch. 15, p. 91

covenants already exists, and he will argue that when such a mechanism is present it is not in one's best interest to violate an existing covenant because:

> there is no man can hope by his own strength or wit to defend himself from destruction without the help of confederates…He, therefore, that breaketh his covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any society that unite themselves for peace and defence but by the error of them that receive him; nor when he is received, be retained in it without seeing the danger of their error; which errors a man cannot reasonably reckon upon as the means of his security…[144]

Hobbes's response is rather straightforward: since the Fool is dependent upon the errors of others for his success in violating covenants, his odds of consistent success in violation are potentially low and definitely uncertain. Since the downside of being caught is often much larger than the potential upside of undetected violation, the Fool is well-advised to adhere to the covenants he makes. Hobbes's argument here is reminiscent of Pascal's Wager in its implicit use of the conjunction of probability and utility.[145] Although an individual may be successful in her attempt to violate a covenant, success is uncertain and both the probability and the magnitude of the downside are difficult to forecast. Even if there is merely a miniscule chance of being detected, if the downside of being detected is huge it is still rational to adhere to the covenant.[146]

Note that Hobbes is not making the strong claim that violation is irrational even if it is beneficial; he is merely claiming that violating is irrational because it cannot reliably be expected to be beneficial. According to Hobbes, even if violating a covenant turns out to be successful, it is probably due to good luck, which is unreliable, and the choice to violate was not a rational

---

[144] *Leviathan*, ch. 15, p. 91-92. Note again that Hobbes' view of justice would have been viewed in the early modern period as being Epicurean. Hobbes sees the benefit of justice as freedom from the concern that you will get caught in the violation of a covenant. Similarly, Epicurus sees the benefit of justice as knowing that the pleasures you have will not be taken away from you. See Epicurus, *Principle Doctrines*, #5 and #17.

[145] I owe this insight to a personal conversation with Professor Douglas Jesseph.

[146] Admittedly, the downside from violating any covenant is less severe than the pain of eternal damnation as referenced in Pascal' Wager. The point is that if the downside of being caught in violation of a covenant is extremely severe, even a miniscule chance of being caught will make compliance with the covenant a rational course of action.

strategy *despite the successful outcome*.[147] In addition, even if violation was successful on one occasion, it may give an individual a bad reputation and prohibit him from entering into future cooperative ventures, which is obviously against his own self-interest. At first glance violation may appear to be a good idea, but proper reflection will reveal that it is not rational.

Kavka observes that the interpretation of the Fool's argument hinges on how we interpret the phrase "conduced to one's benefit."[148] If the phrase refers to an expected beneficial outcome, the issue between the Fool and Hobbes is that Hobbes believes that the uncertainty of the outcome and the unreliability of luck make violating justice against our best interest. If the phrase refers to actual or certain outcomes, there is no substantive disagreement between Hobbes and the Fool. That is, if an individual is certain that she can violate a covenant without being caught, Hobbes would agree with the Fool that she should do so. Hobbes is merely arguing that, unlike the Lydian Shepherd in possession of the Ring of Gyges, humans in the real world never find themselves in a situation of such certainty.

At this point, an obvious objection to Hobbes's argument arises. While Hobbes may be correct in his assertion that the Fool can never be one hundred percent certain that he can get away with the violation of a covenant, it does not necessarily follow that the Fool should therefore adhere to his covenants in every situation. Hobbes's recommendation to adhere to covenants may be viable in many, or even most, situations, but certainly there will be situations in which the odds of detection appear to be quite low and the potential payoffs from successful violation are quite high. Surely, the skeptic will claim, in such a situation it is in the Fool's best interest to violate the covenant since the conjunction of probability and utility suggest that the Fool can expect to reap significant benefits. I will address this objection in detail in the sections

[147] See Sorell, p. 130
[148] Kavka (1995), p. 7

on Kavka and behavioral economics below, but for now it will have to suffice merely to suggest that the problem with the skeptic's argument is that the Fool, like the rest of us, is probably a poor judge of his own odds of success.

Before moving on to my analysis of contemporary commentators on Hobbes's reply to the Fool, it will be helpful to re-cast the Fool's argument in the context of the conclusions drawn at the end of Chapter 3. Recall that Gauthier makes three claims regarding deliberative procedures: He claims first that humans have the ability to willingly choose deliberative procedures in general, second, that it is rational to choose a deliberative procedure in accordance with APO, and third, that once we choose a deliberative procedure in accordance with APO, it is rational to comply with that deliberative procedure in all instances. At first glance it may appear that the dispute between Hobbes and the Fool is being waged over the third claim, but this is not the case; compliance with a covenant (the subject of the dispute with the Fool) is not the same as compliance with a deliberative procedure. The dispute is being waged over Gauthier's second claim regarding what type of deliberative procedure it is rational to choose; they are arguing over a choice about how to make choices. The Fool is proposing a strategy similar to Gauthier's SM, in which he will violate covenants when it is advantageous to do so. Hobbes, on the other hand, while certainly not endorsing Gauthier's APO, is arguing in favor of a strategy that involves keeping the covenants that one makes. Unlike Gauthier, Hobbes emphasizes the importance of reciprocity and repeated interaction, and unlike the Fool he understands that uncertainty plays a major role in deciding on a rational course of action with respect to the keeping of covenants. These particular aspects of Hobbes's reply to the Fool will figure prominently in the various accounts of contemporary Hobbes scholars, to which I will now turn.

Contemporary Commentary on Hobbes's Reply to the Fool

Kinch Hoekstra

Hoekstra reads Hobbes as believing (in agreement with the Fool) that adherence to covenants can at times conflict with self-interest. His interpretation of Hobbes is unique in that he views Hobbes as addressing two distinct kinds of Fools: Explicit Fools who deny justice outwardly through word or overt deeds, and Silent Fools "who deny justice only in their hearts."[149] According to Hoekstra, Hobbes's reply to the Fool is directed against the Explicit Fool only, as Hobbes wants to silence those who intend to publicly endorse disobedience to the law and the sovereign. Under this interpretation, to argue against the Silent Fool would be ludicrous, as the Silent Fool is making the relatively innocuous claim that violating a covenant is not always advantageous, but that it certainly is on some occasions. The Silent Fool believes, and Hobbes agrees, that it is reasonable to act unjustly if the rewards are great, punishment is light and the likelihood of detection is low.

Using the distinction between Silent and Explicit Fools, Hoekstra wants to demonstrate that in Hobbes's system the role of the sovereign is to change the probability payoff in such a way that the Silent Fool is hesitant to cheat. That is, Hobbes thinks the sovereign should establish a system of punishment that is strict enough so that when an individual calculates the pros and cons of breaking a covenant, the "payoff scale" encourages the covenant to be kept.[150] This is where Hoekstra, most likely unintentionally, makes his most important contribution to the argument *against* the Fool: "The Silent Foole may sometimes reasonably expect a net benefit from injustice (it is likely that he will profit), but generally would not be reasonable to act on or

[149] Hoekstra, Kinch. "Hobbes and the Fool." *Political Theory* 25, no. 5 (Oct. 1997): 623
[150] Hoekstra (1997), p. 627-628

rely on this expectation (though profit is probable, the risk is too great)."[151] This idea, which is once again reminiscent of Pascal's Wager, is known as "weighted average probability." An individual who violates a covenant may have a high expectation of succeeding, but if the downside of being caught is far greater than the upside of success, it may be reasonable for that person to refrain from the violation. The point is that it is necessary to calculate not only the probability of success, but the corresponding values of success and failure because the expected utility of the attempted violation can be negative even if the violation is expected to succeed.

An example will help to clarify this point. Suppose Bob has an exam today in his finance class for which he is insufficiently prepared, and he is tempted to cheat. Suppose further that he correctly assesses his odds of success at 90% and that a good result on the exam as the result of cheating will increase his utility by 50 units. Suppose further that if he is caught (an outcome with a 10% probability), he will surely be expelled from school which will result in a decrease in his utility of 1,000 units. His expected return looks like this:

Expected return if successful = .9 x 50 = 45

Expected return if caught = .1 x -1,000 = -100

Expected return from the act of cheating = 45 + -100 = -55

Under this scenario, Bob expects to be successful, but his expected utility (the product of his expectation of success versus failure and the relative payoffs of each) is negative. *So, he would be foolish to cheat even though he expects to get away with it.*

To complicate the matter further, psychological analysis has shown that humans are poor predictors of the probability of outcomes. This flaw in human reasoning adds further risk to the attempted violation due to the uncertainty of the utility calculation, and it makes the case in favor

[151] Hoekstra (1997), p. 631

of violation even more unconvincing. While Hoekstra's attribution to Hobbes of a distinction between Explicit and Silent Fools is dubious and difficult to support, his acknowledgment of the relevance of weighted average probability to the Fool's argument is an important contribution. As we will see later in the chapter, the field of behavioral economics has a great deal to say on this point.

Geoffrey Sayre-McCord

In his commentary on Gauthier, Sayre-McCord does not directly address Hobbes's reply to the Fool, but it requires no great stretch of the imagination to apply his argument to the dispute.[152] His main point of contention with Gauthier relates to Gauthier's assumption of the translucency of dispositions.[153] Recall from Chapter 3 that Gauthier assumes that human dispositions are best-described as a midpoint between transparency and opaqueness; while we may not be able to always ascertain the intentions and dispositions of other actors, we can often approximate intentions and dispositions with more accuracy than random guessing. This he refers to as translucency. Sayre-McCord will argue that, in the real world, the dispositions of some individuals are opaque enough that deception is often to the individual's advantage.

He employs probability equations to make the argument that, "The choice of a moral character is rational, then, only if one has a reason to think one is a (sufficiently) translucent member of a community of (sufficiently) translucent moral people."[154] If Jane finds that her

[152] Sayre-McCord, Geoffrey. "Deception and reasons to be moral." in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 181-195. New York: Cambridge University Press, 1991

[153] It should be noted that the Sayre-McCord article pre-dates Gauthier's rejection of his own dispositional account in "Assure and Threaten" (1994). I will therefore continue to refer to dispositions in addressing Sayre-McCord's criticism of Gauthier, but it should be noted that Sayre-McCord's critique of Gauthier's dispositional account can apply equally well to Gauthier's account of deliberative procedures in 1994 and beyond.

[154] Sayre-McCord, p. 191

dispositions are sufficiently opaque, she can reasonably decide that it is in her self-interest to adopt a disposition to violate covenants when it is advantageous to do so. While she may not possess all of the powers of the Lydian Shepherd, her opacity does afford her the opportunity to appear just while behaving in an unjust manner.

To see the problem with this argument, we need only refer back to my commentary on Hoekstra. Sayre-McCord employs probabilistic reasoning to make his case, but he fails to recognize that humans are notoriously inept at applying probability to situations of uncertainty.[155] He claims that a "transopaque" egoist (a creature similar to Hoekstra's Silent Fool) can expect to obtain the benefits of having a moral disposition without having to pay the costs of constraint, but he ignores the distinct possibility that the transopaque egoist's calculation of expected benefit will be biased by overconfidence and an inability to properly employ probability calculations. Sayre-McCord argues, as a rebuttal to a potential critique, that the victims of the transopaque egoist will not be able to sufficiently punish her because the victims are ignorant and irrational.[156] This point is a valid one, but it also serves to strengthen the argument against Sayre-McCord himself. Individuals are indeed ignorant and irrational as Sayre-McCord claims, but these facts are not reserved only for the victims of the transopaque egoist; they also apply to the transopaque egoist herself. Being human, it is likely that the transopaque egoist, when presented with an opportunity to violate a covenant, will reason in a flawed way and overestimate her odds of success. Thus the Fool, in the guise of the transopaque egoist, still lacks a convincing rational argument in favor of deception.

[155] See Kahneman (2011), pp. 256-257
[156] Sayre-McCord, pp. 193-194

Jean Hampton

Hampton also recognizes that humans often reason in a flawed manner, and she treats the issue in a more explicit way than Hoekstra or Sayre-McCord. Her interpretation of Hobbes's characterization of conflict in the state of nature provides some useful insight into how the human mind addresses cooperation and conflict, and she introduces us to the problem of inter-temporal choice.[157] Hampton describes three possible interpretations of Hobbes's account of conflict in the state of nature. In the first account, which she refers to as the "rationality account", she interprets Hobbes as arguing that the source of conflict is simply a result of prisoner's dilemma reasoning by creatures whose dominant passion is that of self-preservation. Humans are rational actors and they recognize that their best self-preservation strategy in the state of nature is to refuse to cooperate no matter what the other party does, so they will find themselves trapped in a situation of eternal conflict. The problems with the rationality account are twofold: First, it makes the very strong claim that conflict is always the best strategy, which is entirely implausible as evidenced by our prior analysis of iterated PD games. Second, if the rationality account holds, it is unclear how we could ever agree to the institution of a sovereign. That is, under a strict interpretation of the rationality account we would never be able to escape the state of nature.

The second interpretation of Hobbes proposed by Hampton is the "passions account". This account asks, if the laws of nature direct individuals so seek peace, why do self-interested individuals not follow the direction of these laws?[158] The answer is that individuals do not behave rationally; passions such as fear, greed and, most importantly, glory, drive them to behave in a manner that is inconsistent with reason. Hampton cites two problems with this

[157] Hampton, Jean. *Hobbes and the Social Contract Tradition.* New York: Cambridge University Press, 1986
[158] Hampton (1986), p. 64

account as well. First, the passions account seems to be inconsistent with total warfare in the state of nature. Individuals will reason that cooperation is in their best interest, frequent cooperation will occur in the state of nature, and the need for a sovereign will thus be called into doubt. However, if the passions account does generate sufficient conflict in the state of nature to necessitate the institution of a sovereign, it is no longer possible to accept Hobbes's own version of human psychology.

After rejecting the rationality and passions accounts, Hampton advocates a third interpretation of conflict which she refers to as the "shortsightedness account." According to this account, individuals are driven by reason, and cooperation in the state of nature is a rational choice, but cooperation does not take hold because individuals fail to reason in the proper way. Specifically, some individuals will fall victim to the desire for immediate gratification and give undue weight to the short-term benefits of conflict over the long-term benefits of cooperation. Hampton suggests that individuals may be shortsighted in this way because they consider the future benefits of cooperation to be too vague, uncertain or remote to warrant cooperative behavior in the short-term, or it may simply be that the average individual in a given society is not aware that he is involved in a multi-period iterated prisoner's dilemma game with every other individual with which he interacts (this lack of awareness may seem shocking to some game theorists, but it is entirely plausible nonetheless). Other individuals, perceiving this shortsightedness in those with whom they are interacting, will anticipate the conflict-oriented reasoning of their counterparts and will therefore decide not to cooperate themselves despite the fact that they understand the long-term benefits of cooperation.[159] Still other individuals will be influenced by a desire for glory and will tend towards conflict because they are overconfident

[159] Hampton (1986), p. 81

and have overestimated their chances of winning and underestimated the abilities of their adversaries.[160]

Hampton's shortsightedness account is a superior interpretation of Hobbes because it can explain the presence of conflict in the state of nature without sacrificing Hobbes's version of human psychology. Under the shortsightedness account, conflict is driven not by "disruptive passions" but by "fallacious reasoning", and this is entirely consistent with Hobbes's psychology because overconfidence and a preference for immediate over future gratification is a mistake that rational beings whose primary motivation is self-preservation will naturally make.[161]

By recognizing the importance of shortsightedness, Hampton gives us our first glimpse of the idea that will be the focal point of the final section of this chapter, namely, that when dealing with situations of uncertainty, *human beings reason in a flawed manner*. The only problem with Hampton's argument here is that she does not take it far enough. She cites the tendency to discount future benefits and the tendency to be (abnormally) risk averse as "congenital deformities" and she says that, "Hobbes could not plausibly argue that they are powerful or widespread in the population."[162] As demonstrated in the section on behavioral economics below, these and other tendencies to flawed reasoning are not only widespread in the population, they are actually quite normal (despite being flawed) aspects of human behavior. To find evidence of shortsightedness and individuals' inability to accurately assess future values, we need look no further than the behavior of the typical American consumer with respect to credit card debt. Agreeing to pay 17% interest in order to buy an upgraded iPhone today rather than six months from now provides strong evidence in favor of the existence of flawed reasoning in

[160] Hampton (1986), p. 86-87
[161] Hampton (1986), p. 85
[162] Hampton (1986), p. 84

situations of inter-temporal choice. These tendencies may be quite valuable in a pre-societal environment where overconfidence, immediate gratification and risk aversion promote survival of the species, but they are potentially harmful as we move towards a better life via the formation of a social contract, and they are definitely harmful in a competitive market environment, as my discussion of behavioral economics will clearly demonstrate. These flaws in our reasoning give rise to conflict in the state of nature, but we can overcome them if we recognize their existence and adjust our reasoning and our expectations accordingly.

Hampton's shortsightedness account is helpful to our analysis of Hobbes's response to the Fool because it demonstrates how flawed reasoning might lead the Fool to believe that violation of a covenant is in his self-interest when it actually is not. Kavka[163] recognizes that Hampton sees Hobbes as agreeing with the Fool's claim that we should violate a moral rule or law of nature when, on a case-by-case basis, it is to our own advantage to do so. Kavka agrees with this interpretation of Hobbes, but he contends that Hobbes also wants to claim that if we reason properly, we will see that compliance with the covenants one makes brings a higher expected benefit than violation. In this way, Kavka is hinting at the notion of deliberative procedures, and he wants to replace Hampton's case-by-case reasoning (which can often lead to a utility-destroying choice to violate) with a rule-based reasoning which will lead to the utility-maximizing choice of compliance.

David Gauthier

Gauthier addresses Hobbes's Fool in several places within his published works. He is consistent in his conclusion that the Fool is mistaken, but the particulars of his argument vary

[163] Kavka (1995), p. 10-12

widely. He uses two principal strains of argument against the Fool, which I will refer to as the "conventional reason approach" and the "deliberative procedures approach." Gauthier advances the conventional reason approach in his book *Moral Dealing*[164] and in the article "Thomas Hobbes: Moral Theorist."[165] Where the Fool says that any restriction of the pursuit of self-interest is against reason, Gauthier replies that one must not only restrict self-interest when adhering to the second law of nature, one must restrict reason as well.[166] That is, Gauthier claims that the second law of nature requires us to give up some of our natural right and some of our natural reason and, "In place of natural reason, one must accept the conventional reason of the law, which directs one to adhere to one's covenants."[167] Furthermore, Gauthier claims that the Fool is missing the point that once we enter a covenant, self-preservation is no longer the standard of reason; peace becomes the standard of reason.[168] In this context, Gauthier is viewing Hobbes's moral theory as "…a dual conventionalism, in which a conventional reason, superseding natural reason, justifies a conventional morality, constraining natural behavior."[169]

It is difficult to see why Gauthier believes this notion to be correct, and even more difficult to understand how he attributes such a notion to Hobbes. Adherence to Hobbes's second law of nature does not require that an individual restrict his own self-interest; it requires him to restrict his own liberty and to "be contented with so much liberty against other men, as he would allow other men against himself."[170] For Hobbes, this restriction of liberty is very much in keeping with one's own self-interest, as it allows an individual to expand her opportunity set and to escape the suffocating fear of the state of nature. Fortunately for Gauthier and for my

[164] Gauthier (1990)
[165] Gauthier (1979)
[166] Gauthier (1990), p. 143
[167] Gauthier (1990), p. 143
[168] Gauthier (1979), p. 557
[169] Gauthier (1979), p. 547-548
[170] *Leviathan*, ch. 14, p. 80

own argument in this dissertation, he offers a more satisfying and apparently unrelated reply to the Fool in *Morals by Agreement*, which I refer to as the deliberative procedures approach.[171] According to the deliberative procedures approach, the Fool is guilty of error on two levels, both of which relate to ideas I addressed in Chapter 3. First, the Fool fails to recognize that real acceptance of the second law of nature is possible only among those who have adopted the third law of nature as well. In other words, an individual is worthy of consideration as a partner in mutually beneficial interaction only if that individual has adopted a deliberative procedure which demands consistent compliance with all of the covenants she makes. According to Gauthier, in his reply to the Fool Hobbes does not go far enough in distinguishing the rationality of keeping one's agreements from the rationality of adopting a deliberative procedure of keeping one's agreements.[172] The Fool is relating reason directly to the benefits of performance in a single instance rather than to the longer-term benefits of a deliberative procedure of performance,[173] and he rejects the claim that the third law of nature necessarily follows from the second. That is, the Fool rejects the third of Gauthier's three claims that I addressed in Chapter 3; the claim that it is always rational to comply with a deliberative procedure if it is rational to adopt it.[174] In Chapter 3 I argued extensively that Gauthier is correct in this claim, so I will not revisit the argument here. At this point it is only necessary to recognize that the rejection of this claim can be attributed to the Fool.

---

[171] When addressing Hobbes' Fool in *Morals by Agreement*, Gauthier actually refers to dispositions rather than to deliberative procedures. However, given Gauthier's subsequent rejection of dispositions in favor of deliberative procedures (1994), it is now appropriate to interpret his argument against the Fool in the context of deliberative procedures.

[172] Note that it is not only rational to have such a deliberative procedure, it is also rational to communicate one's possession of this deliberative procedure to other individuals. As noted in the section on Sayre-McCord above, due to the ability of other individuals to detect insincerity, it is strongly preferable, and possibly even necessary, to actually possess the deliberative procedure as opposed to merely appearing to possess it.

[173] Gauthier (1986), p. 162

[174] Gauthier (1986), p. 165

The second error that the Fool makes under Gauthier's deliberative procedures approach is his failure to realize that individuals who adopt a deliberative procedure of adherence to the covenants they make will tend to have better future outcomes than those who have adopted no such procedure because those who have adopted such a procedure will enjoy the benefits of cooperative interaction in a non-zero-sum game.[175] The Fool believes that it is rational to maximize his utility on an individual level in each instance of cooperation, but by maintaining this belief he is completely missing the point of cooperative interaction. The Fool fails to recognize the benefits of substituting an optimal joint strategy for an equilibrium individual strategy. He is operating under the assumption that human interaction is a zero-sum game and he is ignoring the mutual benefits of cooperation. Because he does not adopt a deliberative procedure in accordance with justice, the Fool will not be admitted into many of the beneficial arrangements that cooperative interaction provides; that is, his future opportunity set will be limited.

While Gauthier's conventional reason approach to the Fool is highly implausible, his deliberative procedures approach is helpful in constructing an adequate reply to a portion of the Fool's claims. The Fool rejects Gauthier's claim that it is always rational to comply with a deliberative procedure that we have chosen, and Gauthier does a good job of refuting him. However, the Fool also rejects Gauthier's second claim, namely, that it is rational to choose a deliberative procedure of APO, and, as mentioned in Ch 3, Gauthier's support for this claim is unsuccessful. The question, "What type of deliberative procedure is it rational to choose?" still lacks a satisfactory answer, and the Fool remains defiant in his implicit support of SM. While Gauthier is unable to respond to the Fool on this point, Kavka is able to refine Gauthier's account

[175] Gauthier, David. "Why Contractarianism?" in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 15-30. New York: Cambridge University Press, 1991, p. 24-25

of dispositions and deliberative procedures, and by adding some insights of his own he is able to provide an adequate defense of Gauthier's third claim and a sophisticated rejection of the Fool's support of SM as a rational strategy. Kavka will argue that it is indeed rational to adopt a deliberative procedure (what he refers to as a policy or set of rules) of adherence to covenants, and as we will see in the next section, he will advance a powerful argument that shows why the Fool is mistaken in his claim that opportunistic violation is a rational strategy.

Kavka's Argument from Uncertainty

Gauthier goes part of the way towards refuting the Fool by demonstrating the importance of deliberative procedures and the value of cooperative interaction. Hampton provides some significant assistance by showing that humans often reason in a flawed manner, and Hoekstra's use of weighted average probability applies a useful statistical technique to the argument. However, none of these writers is successful alone in convincingly refuting the Fool. It is Kavka who combines the best elements of these arguments in a novel way and thus sets the Fool up for the fatal blow to his claims.

Kavka's interpretation of Hobbes rests on four core ideas, all of which are relevant to Hobbes's response to the Fool. First, practical reasoning is always forward-looking. Second, real choices regarding adherence to covenants are made under uncertainty. Third, risk aversion is rational. Fourth, rules-based reasoning (precommitment) is rational because attempting to calculate risks on a case-by-case basis subjects an individual to errors and biases.[176] I will address each of these in turn.

[176] Kavka (1995), p. 21-22

According to Kavka, Hobbes must believe that practical reasoning is always forward-looking because this is how Hobbes justifies the formation of covenants in the first place. For Hobbes (and for Kavka), honoring a covenant is rational because it gives an individual access to the benefits of future cooperative ventures. If an individual chooses not to honor the covenants he makes, he can expect retaliation from those whom he has wronged and he will be excluded from the benefits of future cooperation. This emphasis of forward-looking reasoning places Kavka's interpretation of Hobbes and his subsequent argument against the Fool in direct conflict with Gauthier's CM, to Kavka's credit. As mentioned in Chapter 3, Gauthier's refusal to include repeated interaction and reciprocity as elements of CM was a major failure, and Kavka's inclusion of them gives his reply to the Fool more force.

Kavka's claim that practical reason must always be forward-looking has a critical implication for his later claim that adherence to a covenant is rational. In order to properly understand this implication, it is necessary to re-visit some of the game theory that was introduced in Chapter 3. Recall that Axelrod insists on using an iterated PD game because repeated interaction is necessary for any form of cooperation to take hold among rational players; if the individual players know they will never interact again or if the number of future interactions is known, mutual defection is an evolutionarily stable strategy and the process will never move past this point. The reasoning behind Axelrod's contention is this: If there are two players, A and B, in an iterated PD game, and they each know that the number of interactions, *n*, is fixed, then the player who has agreed to comply in round *n* (assume it is player B) will defect in round *n* because he knows that he will not have to face reciprocal consequences in a subsequent round of the game. However, if player A is rational, he will conclude that player B will defect on round *n*, and player A will therefore defect on round *n-1*. Player B, being rational

as well, will conclude that A will defect on round *n-1*, leading player B to defect on round *n-2*. This backward induction reasoning will continue until the players conclude in round 1 that it is irrational to comply, and cooperation will never take hold. Furthermore, Kavka points out that it is not even necessary to assume that the players know exactly how many rounds there will be. All that is required is mutual knowledge that the number of interactions is finite and the way the players play the game will have no impact on the number of rounds.[177]

This argument, while logically valid and certainly interesting from a game-theoretic perspective, is unsound. As Kavka later observes,[178] the argument assumes that each individual is perfectly rational and that each individual is aware that all other individuals are perfectly rational. We need only consider the reasoning ability of our friends, fellow citizens and ourselves to see that this is a very strong and entirely implausible assumption. Many, if not most, individuals are completely befuddled by simple mathematical calculations, so it would be quite a stretch of the imagination to attribute to such individuals a necessary recognition that all of their cooperative interactions are separate rounds of an iterated PD that must account for backward induction. In other words, the backward induction argument does not correspond to the rationality that we find in real individuals.[179] In the real world, human rationality is far from perfect, and this fact will allow Kavka to maintain his claim that practical reasoning is forward-looking without having to sacrifice his contention that adherence to covenants is rational.

Recall that Kavka's interpretation of Hobbes is based on four core ideas. The first, the idea that that practical reasoning is always forward-looking, leads Kavka to an initial recognition

[177] For an explanation of this point, see Kavka, Gregory. "Hobbes's War of All against All." *Ethics* 93, no. 2 (Jan, 1983): p. 302.
[178] Kavka (1983), p. 303
[179] For a useful discussion of the backward induction argument, see Skyrms, Brian. "The Shadow of the Future." in *Rational Commitment and Social Justice*, edited by Jules Coleman and Christopher Morris, 12-21. New York: Cambridge University Press, 1998, p. 12-21

of the fact that the human ability to reason is flawed. The remaining three ideas will naturally follow from this observation about human reasoning. The second concept that Kavka attributes to Hobbes is that choices regarding adherence to covenants are made under uncertainty. This idea is related to the "Cheater Bob" example that was addressed in the discussion on Hoekstra. In the Cheater Bob example, it was demonstrated that an individual may be rationally justified in keeping a covenant even if he expects to get away with violation, due to the possibility of a very negative outcome. Kavka points out that real-world choices of this type are further complicated by the uncertainty surrounding the probabilities of the outcomes and the utilities of the various payoffs. That is, Cheater Bob calculates that his odds of being caught are 10% and the utility of this outcome is -1,000, but can he be confident of these numbers? Surely the numbers he is using for probability and utility are merely estimates; to attribute any more reliability to these numbers would be to assume a level of predictability that our world does not have.

The third idea Kavka attributes to Hobbes is the rationality of playing it safe. In keeping with the Cheater Bob example, Hobbes recognizes the need to account for the small possibility of a devastating outcome, and he contends that the rational actor will forego the potential benefit of violation in order to insure against the potential of a huge loss.[180] In order to support his reply to the Fool, Hobbes does not need to claim that individuals have an irrational aversion to risk (even though they may have such an aversion); he merely needs to show that a rational choice under uncertainty will account for the weighted average probability of the downside to any gamble.

The fourth core tenet of Kavka's interpretation of Hobbes is where all of the elements of the argument against the Fool come together. Kavka states:

[180] It is important to note here that playing it safe is not the same as risk aversion. A risk-averse person is one who avoids risk to a degree that is in excess of what a pure weighted-average utility function would recommend.

> *precommitment* (or rigid rule-following) is rational in the following sense: one is likely to do better overall by rigidly following the core moral rules than by calculating acceptable risks on particular occasions, because errors and biases in such calculations will tend toward leading you to take excessive risks in particular cases.[181]

Practical reasoning is forward-looking, so we must consider our choices in a long-term context. The choices we make are conducted in situations of uncertainty, so complex calculations will be necessary. The downside from an unfavorable outcome can be severe, so we need to be aware of the risks, and, the human ability to reason is flawed by psychological biases, so we are better-served by establishing and following a rational set of rules rather than attempting to conduct complex calculations under situations of uncertainty on a case-by-case basis.

In short, Kavka is claiming the following: Our ability to reason in complex situations of uncertainty is so limited that it is very difficult to determine ahead of time when a potential rule violation is definitely in our self-interest. We will therefore have better outcomes over time if we adopt a less-risky deliberative procedure that assumes that adherence to covenants is in our best interest at all times.[182] The validity of this strong claim rests primarily on the assumption that our ability to reason is deeply flawed. If it can be shown that this assumption is true, the Fool will be in an untenable position. Kavka believes the plausibility of his interpretation can be seen simply by considering the savings of decision costs and the avoidance of errors due to

[181] Kavka (1995), p. 22

[182] It should be noted that the claim I am making here is not in accord with Martinich's interpretation of Hobbes' reply to the Fool (see Martinich, A.P. *Hobbes.* New York: Routledge, 2005, p. 101-104). Martinich claims that because Hobbes is providing a science of politics, and since a science must consist of necessarily true propositions, the Fool cannot base his recommendation to violate a covenant on the positive consequences that he expects to result because these results are not necessarily true. For Martinich, in order for the Fool's claim to be valid, the violation of covenant would have to be beneficial with 100% certainty. The problem with Martinich's interpretation of Hobbes is that, if Hobbes did actually require 100% certainty, he could not justify the institution of a sovereign because it cannot be guaranteed with scientific certainty that the sovereign will make things better than they were in the state of nature. My claim is not that the violation of a covenant is against one's self-interest because it does not offer an objective 100% probability of success; my claim is that the subjective human inability to correctly assess what the probability of success really is makes violation of covenants risky and uncertain. I am grateful to Professor Doug Jesseph for his comments in this regard.

cognitive biases such as the short-sightedness that Hampton addresses and our own self-deception.[183] I will argue that not only is Kavka correct in this belief, it is far more important and forceful than he likely imagines. Behavioral biases cause humans to reason in a deeply flawed way, and this fact will deal a fatal blow to the argument of the Fool. We will now take Kavka's argument one step further by examining some of the most important insights from behavioral economics.

Behavioral Economics and Flawed Reasoning

Kavka's case against the Fool rests on his claim that the human ability to reason on a case-by-case basis is flawed and unreliable. This argument is not new: Recall from Chapter 2 that Adam Smith recognized overconfidence and a preference for short-term over long-term utility as pervasive tendencies of human psychology. Over 200 years later, contemporary behavioral economists have shown through extensive experimentation that overconfidence, inter-temporal choice preferences, the illusion of control, and an underestimation of the impact of randomness all have a huge impact on our ability to make rational choices in situations of uncertainty. I will argue, in keeping with Kavka, that these flaws in our capacity to reason make it impossible for us to reliably estimate the potential costs and benefits of covenant violation on a case-by-case basis, and it is therefore in our best interest to reason on a rule-oriented basis and to resolve to always keep the covenants we make. It must be noted that this account of human psychology is far different from the account given by Hobbes, so we cannot use these insights to claim that Hobbes wins his argument with the Fool. We can, however, use them to declare victory in our own argument against the claims of the Fool.

[183] Kavka (1995), p. 25

Overconfidence[184]

> *"Overconfidence is almost certainly the most important bias in behavioral finance. But most people still think I'm not talking about them."*
> - Ken French, Dartmouth College

Most of us go about our daily lives with an overall sense of confidence. We generally believe that we will survive until the end of the day and we are not incapacitated by concerns over terrorist attacks or the possibility of our home burning down. This sense of confidence helps us to cope with setbacks and plan for the future, but it often reaches too far, leading us to be overconfident in our own abilities.

Numerous studies have demonstrated evidence of "illusory superiority" among individuals; we tend to overestimate our positive qualities and underestimate our negative qualities relative to others. In one study,[185] a group of American drivers was asked whether their driving skills were better than average. If the study participants were accurately describing their own driving skills, we would expect that just about 50 percent of the respondents would claim to have better than average skills and 50 percent would claim to be worse than average. In fact, 93 percent of the participants claimed to be better than average drivers. Lichtenstein, et. al. found that overconfidence is pervasive in average tasks and severe when dealing with difficult tasks.[186] When subjects were asked to give a response to a question with a 99.9% chance of being correct (a 1-in-1000 chance of being wrong), they were actually correct only 81% to 88% of the time.

[184] See MacLean, Frederick and Tim Slattery. "The Collision of Pride and Memory." *The Light Magazine*, September, 2010

[185] See Svenson, Ola. "Are We All Less Risky and More Skillful than Our Fellow Drivers?" Acta Psychologica 47, 1981: 143-148

[186] Lichtenstein, Sarah, Baruch Fischhoff and Lawrence D. Phillips. "Calibration of probabilities: The state of the art to 1980." in *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 306-334. New York: Cambridge University Press, 1982, p. 315

When asked to give a response with a 99.9999% chance of being correct (a 1-in-1,000,000 chance of being wrong), they were actually correct only 90% to 96% of the time.

Severe overconfidence is also present in an individual's ability to predict future events. This tendency leads to optimism and a phenomenon known as the planning fallacy. When forecasting the outcomes of risky projects, people are overly-optimistic, overestimating benefits and underestimating costs. They emphasize scenarios for success and overlook the potential for mistakes and miscalculation. For example, the chance that a new business in the U.S. will survive for five years is 35%. However, 81% of new entrepreneurs estimated that their business has at least a 70% chance of success, and 33% said the odds of failure were zero.[187] We tend to make forecasts that are unrealistically close to the best-case scenario and which could be improved merely by considering the outcome of similar cases. We need look no further than our own homes for an example: The average kitchen remodel in the U.S. was budgeted at $19,000; the average final cost was $39,000.[188]

Not only do we tend to be overconfident in our own abilities, we also tend to remember our successes and mistakes in a way that reinforces our belief in ourselves. We attribute our successes to our own talents and insights, while we attribute our failures to things that we cannot control such as bad luck or unforeseeable circumstances. This leads us to the belief that if we just tweak our strategy a little bit, it will work flawlessly the next time. This collision of pride and memory was recognized by Nietzsche: "'I did that,' says my memory. 'I could not have done that,' says my pride, and remains adamant. Eventually – the memory yields."[189] When our rational mind yields to our pride, we are doomed to repeat the mistakes of the past.

---

[187] Kahneman (2011), p. 256-257
[188] Kahneman (2011), p. 250
[189] Nietzsche, Friedrich. "Beyond Good and Evil." in *Basic Writings of Nietzsche*, edited by Walter Kaufmann. New York: The Modern Library, 1992, 4:68, p. 270

The convergence of overconfidence, optimism and the planning fallacy casts a great deal of doubt on an individual's ability to calculate the odds of success in an uncertain scenario. We must consider also that estimation errors will compound in situations where there are multiple independent steps. If Cheater Bob must take several discrete steps in order to successfully copy a neighbor's exam answers and he is overconfident at each step, the impact of his overconfidence is compounded. For example, suppose that in order to successfully cheat Cheater Bob needs to select a well-prepared student, arrange to sit next to that student during the exam, and view that student's exam without being detected. Suppose further that Bob, failing to consider the independent nature of each step, continues to estimate the probability of success of the overall project to be 90%. If he is overconfident, as most people are in situations of uncertainty, and his actual odds of success at each step are only 70%, his odds of being successful at the project as a whole is only .7 x .7 x .7 = 34.3%.[190] Bob is starting to look like, for lack of a better term, a fool.

Inter-temporal Choice

> *"But much more frequently he is seduced from his great and important, but distant interests, by the allurement of present, though often very frivolous temptations. This great weakness is incurable in human nature."*
> - David Hume, *Essays Moral, Political, Literary*

The tendency to prefer present gain over future gain has been noted by psychologists, economists and philosophers alike. In the 18th century Hume and Smith recognized that present benefits play upon our emotions while future benefits play upon our reason, and Shefrin addresses the same phenomenon in the 21st century:

[190] It should be noted that, although Cheater Bob's odds of success are now only 34.3%, this does not reduce the expected utility of his efforts from -55 in the prior example to -640 (.343 x 50 + .657 x -1000) under this scenario. This is due to the fact that, if Bob fails at the first or second step of the three-step process, he will probably not realize the -1000 utility outcome because he will fail the exam but he will not be caught cheating and expelled. If we assign a utility of -50 to failing the exam, Bob's expected utility is now -155.4 (.7 x .7 x .7 x 50 + .7 x .7 x .3 x -1000 + .7 x .3 x -50 + .3 x -50)

> The needs of the present make themselves felt through emotion. Those needs have a strong voice and clamor for immediate attention. In contrast, the needs of the future have a much weaker voice, expressing themselves through thought. Most people feel the urge to satisfy their immediate needs, but they only think about satisfying their future needs.[191]

What psychologists refer to as inter-temporal choice is merely what Hampton refers to as shortsightedness. Humans are hard-wired to prefer present gain to future gain because when we are presented with an inter-temporal choice our emotions override our reason and opt for the former. Shefrin shows that this bias is evident in the failure of many Americans to save for retirement, opting instead for sports cars, vacations and ever-larger digital TV sets.[192] As Hampton notes, this bias is also evident in the failure to give sufficient weight to the future benefits of cooperative interaction versus the short-term benefits of cheating. Where the Fool claims that violating covenants is "not against reason," he should consider the very real possibility that it is not reason, but emotion that is exerting its influence on his decision and driving him to focus on the short-term only. Reason would lead him to more carefully consider the long-term benefits of adherence.

Illusion of Control

> *"Wall Street's favorite scam is pretending luck is skill."*
> - Ron Ross, *The Unbeatable Market* (2002)

Human beings are motivated to control their environment, and the more difficult a particular task is, the stronger is the resulting feeling of satisfaction if one is able to control it.[193]

---

[191] Shefrin, Hersh. *Beyond Greed and Fear.* New York: Oxford University Press, 2002, pp. 141-142
[192] Shefrin, p. 142
[193] Jennings, Dennis L., Teresa M. Amabile, and Lee Ross. "Informational Covariation Assessment." in *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 211-230. New York: Cambridge University Press, 1982, p. 238

We have a psychological bias to believe that we have more control over our environment than we actually do, and if we perceive that the ability to control a situation is to our benefit, we will tend to attribute control where little or none exists.

Throughout my career as a professional investor I have had numerous opportunities to observe the illusion of control in practice. Wall Street was built on the lie that it is possible to exert a significant amount of control over investment outcomes, and given the extremely high rewards that can be reaped by convincing others of one's ability to exert such control, the motivation to believe in one's own control abilities is very strong. The truth is, however, that the evidence of a *lack of control* on Wall Street is overwhelming. Leaving the events of the global financial crisis of 2008 aside, numerous studies[194] have shown that fund managers usually underperform the market over the long term (that is, they cannot control their investment outcomes), and when they do outperform they attribute their performance to skill rather than luck.[195] These individuals, who are assumed to be familiar with statistical quantitative methods, apparently ignore the irrefutable fact that random chance dictates that some managers must outperform over any given time period due purely to luck. The fact that their outperformance does not persist provides strong evidence that outperformance is due to luck rather than skill; an inconvenient truth that Wall Street unsurprisingly chooses to ignore.

When the Fool constructs a strategy to violate a covenant, he naturally believes that he can control the outcome of his plan. However, just like the Wall Street fund managers, it is likely that the Fool has overestimated the amount of influence that he can exert on his

[194] For an explanation of why this is the case, see Armstrong, Frank. *The Informed Investor*. New York: Amacom, 2002: 85-98, and Goldie, Dan and Gordon Murray. *The Investment Answer*. New York: Business Plus, 2011: 41-49. For updated data on the returns of fund managers versus the markets, see the quarterly Standard and Poor's Indices Versus Active Funds (SPIVA) Scorecard. The outcome is always the same: as a group, fund managers underperform the market.

[195] Kahneman (2011), p. 215

environment, and he has almost certainly underestimated the amount of influence that random and uncontrollable events will exert on him. The failure to recognize the impact of random events in our lives is the most powerful of all of the behavioral biases under consideration, and it is the central theme of a brilliant book by Taleb,[196] whose ideas will further undermine the Fool's argument.

Randomness, Predictability and Probability Neglect

> *"Any time you try a decent crime, you got 50 ways you can (mess) up. (If) you think of 25 of them, you're a genius. And you ain't no genius."*
> - Mickey Rourke as Teddy Lewis in *Body Heat*[197]

We humans have a strong motivation to control our environment, and in many situations we can succeed in doing so. In situations where we cannot control our environment, our mind attempts to compensate for this inadequacy and we suffer from the illusion of control. We also have a strong motivation to *understand* our environment, but this is often a far more difficult task. Because the world around us is highly complex, we have evolved mental "shortcuts" to allow us to attempt to make sense of our environment in a quick and efficient manner. These shortcuts are useful in survival situations where rapid decisions must be made, but they can be harmful in many of the more complex situations that modern society places us in. They lead us to misunderstand probability and to underestimate the impact of randomness on our lives, which in turn leads us to believe that the world is far more predictable than it actually is.

Sunstein[198] coined the term "probability neglect" to describe the inability of the human mind to deal with small risks. We tend to grossly overestimate the likelihood of recent, dramatic

[196] Taleb, Nassim Nicholas. *Fooled by Randomness.* New York: Texere, 2004
[197] Rourke, Mickey, *Body Heat.* DVD. Directed by Lawrence Kasdan. Los Angeles: Warner Home Video, 1997.

and sensational events and underestimate the likelihood of unspectacular events or events that have never occurred. For instance, Slovik, et. al.[199] found that earthquake insurance purchases spike shortly after an earthquake occurs as the dramatic event is at the forefront of people's minds, then steadily decline as the memory recedes despite the fact that the risk of future earthquake remains constant. Furthermore, Taleb[200] shows that individuals tend to give too much weight to the most likely outcome and ignore the impact of abstract or unlikely outcomes. Ignoring unlikely but potentially disastrous outcomes is yet another example of how an individual can fail to properly calculate the weighted average probability of the utility of an expected outcome. When calculating his odds of successful violation, the Fool will account for the ways he has seen others get caught in the past, and he will probably imagine some spectacular ways that he could get caught this time, but he will be unable to properly account for the plethora of mundane or just plain silly events that could cause him to fail in his attempt at deception. The human mind is simply not designed to imagine all of the possible outcomes and then properly assign probability and corresponding utility to each of them. Furthermore, even if he was able to properly apply the techniques of weighted average probability, his failure to fully appreciate the role of randomness in his life would still hinder the Fool's attempts at deception. The impact of randomness is the central theme of Taleb's work[201], and its profound implications reach far beyond the scope of this paper. For our purposes it will suffice to summarize a few of the major points.

Psychologists use the term "hindsight bias" to describe the tendency to misapply current hindsight to past foresight. Events that have occurred appear to have been more predictable in

---

[198] Kahneman (2011), p. 144
[199] Kahneman, Slovik & Tversky (1982), p. 465
[200] See Taleb (2004), Ch 11
[201] See also Taleb, Nassim Nicholas. *The Black Swan.* New York: Random House, 2007

hindsight than they actually were before they occurred. This bias runs absolutely rampant on both Wall Street and Main Street as the human mind constructs narratives to explain how easily an event could have been foreseen. Consider how many times we have seen the news media report a sensational event such as a stock market crash, a terrorist attack or an upset in a sporting event and thought, "I *knew* that was going to happen!" After having such thoughts, we often create stories to explain the "obvious" causal nexus that led to this outcome and we can convince ourselves that these causes and the resulting outcome were evident from the beginning and that we knew it was going to happen all along. Yet we seldom pause to consider whether we were thinking in the same way before we knew how the event transpired.

College football provides us with a robust example of hindsight bias in action. In the first week of the season, most of the major programs schedule a "tune-up" game with a far inferior opponent in order to build confidence and prepare for games against better teams later in the season. However, each year, much to the dismay of the fan base, one or more of these major programs experiences the embarrassment of losing its tune-up game. After the loss, fans and the sports media invariably concoct narratives to make the loss seem like it was inevitable from the start; the quarterback has always been overrated, the coach was too easy on the players in practice, and the opposing team is far better than an average tune-up game opponent. Anyone could have seen that the loss was inevitable. What the fans fail to consider is that a similar narrative could also be constructed for all of the major programs that won their games that day. Their hindsight bias makes the loss that did occur seem obvious, but they fail to consider the dozens of losses that did not occur despite their being a fit for the same narrative. Our mind's insatiable desire to make sense of a complex world often fools us into believing that we have the ability to predict the future, but the world will continue to confound us.

Taleb also emphasizes the detrimental effects of incomplete information, a phenomenon Kahneman refers to as "WYSIATI", or, What You See Is All There Is.[202] WYSIATI is really a form of cognitive laziness. Humans are constantly in a situation of incomplete information, yet we fail to actively seek new information, we tend to make decisions using only information that we already have, and we discount the importance of information that we do not have. We suppress doubt and ambiguity, and we construct narratives to convince ourselves that we have all of the information necessary to make an informed decision. When consulting past experience, we assume that if something did happen, it *had to happen*; we do not consider alternative possible states of affairs. That is, we convince ourselves that we cannot be impacted by the effects of information that we do not have or random events that we did not (and could not possibly) consider. Once again, our minds are attempting to construct a coherent understanding of the world around us, but the task is far beyond our current capacities.[203]

It is not by accident that Taleb's book has the word "fool" in the title. The book is about a fortunate (but not necessarily Hobbesean) fool who imagines himself to be successful due to skill or some other determinate factor, but who is actually successful due to luck.[204] Ironically, this fool may actually benefit from being a fool because he takes tremendous risks without recognizing the scale of the risks he is taking. He is overconfident in his abilities. He fails to properly calculate probability. He thinks he has more control over his life's outcomes than he actually does and he ignores the distant future. His failure to appreciate the impact of

[202] Kahneman (2011), pp. 85-88

[203] Interestingly, a recent study suggests that an individual's political views can negatively impact her ability to correctly interpret statistical data that offer support to policies which she opposes. This "confirmation bias" leads us to be accepting of data that confirm our existing notions, but critical of data that contradict these notions. See Kahan, Dan M., Ellen Peters, Erica Dawson, and Paul Slovic. "Motivated Numeracy and Enlightened Self-Government." Yale Law School, Public Law Working Paper no. 307 (September 3, 2013), URL=< http://ssrn.com/abstract=2319992 or http://dx.doi.org/10.2139/ssrn.2319992>), p. 149

[204] Taleb (2004), pp. 1-3

randomness allows him to enter a risky situation without fear, and if it he obtains a favorable result he has benefitted from his own stupidity. Examples of such behavior can be found in every stock market bubble since the beginning of time. Stock Market Fool takes excessive financial risks she does not understand, overconfident of her ability to predict the unpredictable. If she is lucky enough to get out before the bubble bursts, she constructs a narrative. She claims, "I bought internet stocks before anyone was talking about them," even though she got the idea from reading an article on internet stocks in Business Week, "and I got out because I saw the end was near," even though she actually sold because she needed the money to pay a massive credit card bill. The problem is, Stock Market Fool has increased her level of overconfidence and her belief in the predictability of the market, and she is setting herself up for an even higher likelihood of a disastrous outcome the next time.

That which is true of Stock Market Fool is true of Hobbes's Fool as well. If he has enjoyed prior success in covenant-breaking he will likely underemphasize the risks he took in the past and become overconfident in his ability to succeed in the future. However, risk only exists in the future and time is the enemy of the Fool. His past success is likely due to luck rather than skill, and as the number of his attempts at violation increases, so do the odds of his luck running out.

Conclusion

Taleb and the other behavioral economists provide strong support for Kavka's reply to the Fool. The Fool challenges Hobbes by arguing that it is wise to adhere to covenants in

general, but to violate when profitable "golden opportunities" present themselves. Kavka[205] recognizes that the Fool's reply would be impossible to overcome if it were possible to definitively identify ahead of time which opportunities are truly golden and which are merely "Fool's gold." However, behavioral economics provides a wealth of evidence that it is very likely that overconfidence and an underestimation of randomness will lead the Fool to commit a large number of "false positive" errors, and he will see golden opportunities where none actually exist. Cognitive biases and ignorance of uncertainty are so pervasive and influential in our psychological constitution that we cannot rely on our own judgment to ascertain when an opportunity has a positive expected return. The world is highly complex and unpredictable, and *given a sufficiently large amount of trials, every unjust person will eventually realize some amount of downside by being discovered*. The fact that we can cite examples of people who have managed to get away with violations of justice over long time periods does not reduce the force of this argument. As long as the odds of successful violation are greater than zero, chance dictates that we will observe some individuals who manage to get away with deception over and over again, but it is impossible to identify ahead of time which individuals will succeed, because their success is attributable to luck rather than skill.[206] As their luck continues, their overconfidence will grow, their sense of control will strengthen, they will attribute their success to skill rather than chance, and they will engage in more reckless behavior. If they do not die or

[205] Kavka (1995), p. 26

[206] This observation has a useful analogy in the investment world. It is frequently observed that a small number of portfolio managers will outperform the market for several consecutive years, only to revert to average or below-average performance over a longer time period. In the marketing blitz that typically follows the period of outperformance, the portfolio manager is heralded as having exceptional skill and insight, yet his subsequent fall from grace indicates that the performance was most likely a result of luck rather than skill. The unfortunate consequence of this phenomenon is that the manager's clients are paying "skill prices" and receiving "luck results."

reform before their luck runs out, they will eventually be discovered. As Kavka says, "Time wounds all heels."[207]

A potential critique of this conclusion is offered by Rainbolt.[208] He claims that in order for an argument of this kind to work, it must be assumed that we all have an approximately equal ability to detect and conceal dispositions,[209] and he correctly observes that Gauthier does implicitly make this assumption. However, Rainbolt takes exception to this assumption. He claims that the assumption of equality in the ability to detect and conceal a disposition to injustice is implausible, and that in the case where someone (Snidely) is exceptionally good at concealing his own disposition and detecting those of others, SM is his rational strategy.[210]

The problem with this argument is that it is plausible in a world inhabited by beings who reason in a perfect manner, but it fails to account for any of the flaws in reasoning that were addressed by the behavioral economists above. Rainbolt does not claim that Snidely is seen to have this ability by an impartial observer with perfect information; he claims that Snidely is making a self-assessment. This claim is implausible due to the now-familiar human tendency to be overconfident and ignore randomness. That is, it is necessary to differentiate between Snidely believing that he has an ability to detect and conceal dispositions and Snidely actually having this ability. It is likely that Snidely thinks he is better than most at detection and concealment of dispositions, but he is probably about average. Although he or someone like him may have gotten away with an SM strategy in the past, this is not necessarily evidence of a talent for detection and concealment; it is probably the result of luck.

---

[207] Kavka (1995), p. 27

[208] Rainbolt, George. "Gauthier on Cooperating in Prisoner's Dilemmas." *Analysis* 49, no. 4, Oct, 1989: 216-220

[209] It should be noted that the Rainbolt article pre-dates Gauthier's rejection of his own dispositional account in "Assure and Threaten" (1994). I will therefore continue to refer to dispositions in addressing Rainbolt's criticism of Gauthier, but it should be noted that Rainbolt's critique of Gauthier's dispositional account can apply equally well to Gauthier's account of deliberative procedures in 1994 and beyond.

[210] Rainbolt, pp. 218-219

Rainbolt claims, "to assume equality of ability to detect and conceal dispositions …robs Gauthier's argument of its interest,"[211] but he is mistaken. Gauthier's implicit assumption may not be true, but it does not damage his larger argument. It is certainly the case that individuals possess differing abilities for detection and concealment, but an individual's lack of ability to objectively assess her own abilities, combined with the negative consequences of an overly-optimistic assessment, are evidence in favor of the adoption of a deliberative procedure (now used by Gauthier in place of dispositions) of observing covenants at all times versus adopting the SM strategy. As Pascal might say, we probably do not all have equal or even nearly-equal abilities at deception and concealment, but we are well-advised to act as if we do.

The Fool wants to convince us of the benefits of opportunistic violation of covenants, but the insights provided by behavioral economics demonstrate that he is definitely overconfident and probably bad at math. We are now left to re-examine Gauthier's three claims from Chapter 3 in light of the commentary from the current chapter. First, Gauthier claims that humans have the ability to willingly choose deliberative procedures. Kavka agrees, and he supports this notion under the heading of rules-based reasoning. Second, Gauthier claims that it is rational to choose a deliberative procedure in accordance with APO. We have seen that APO has its flaws, but Kavka provides a modified version that accounts for the importance of reciprocity and repeated interaction. Kavka has demonstrated that making a precommitment to observe covenants in all cases is rational because attempting to calculate risks on a case-by-case basis subjects an individual to errors and biases. Third, Gauthier claims that it is rational to adhere to a deliberative procedure once the procedure has been chosen. In Chapter 3 I assert that this claim

[211] Rainbolt, p. 220

is valid because cross-temporal consistency of behavior is psychologically beneficial, and this claim goes unchallenged by the philosophers we have examined in this chapter.

We now have what I believe to be a convincing argument that acting in accordance with justice is a rational strategy for a self-interested actor. The human ability to reason is flawed, but this does not mean that we cannot act rationally in deciding to adopt a deliberative procedure to act in a just manner. When we reason on a case-by-case basis, we are subject to a wide variety of cognitive biases which hinder our ability to make rational decisions. However, we have a much higher chance of performing truly rational actions if we adhere to a pre-determined rational rule. A permanent deliberative procedure in accordance with justice is a rational choice in a world of imperfectly rational humans.

CHAPTER 5: OUTSIDE THE SIMULATOR, INSIDE OURSELVES

Chapter 4 argued that adherence to a deliberative procedure in accordance with justice is a rational choice in a world of imperfectly rational humans. Given our overconfidence, our failure to recognize the impact of randomness in our lives, and our inability to accurately assess the odds of cheating successfully, we are well-advised to behave in a just manner, for, if we do, we can enjoy the benefits of mutual cooperation now and in the days and years to come. Yet this characterization of the benefits of justice is lacking in some respect. Our ordinary notions of just behavior lead us to believe that there are reasons for behaving in a just manner other than the fact that we probably cannot get away with acting otherwise; there seems to be more to justice than the market-based benefits addressed in the contractarian argument.

In this chapter, I will examine whether there are benefits to justice other than the market-based benefits addressed so far. My contractarian argument in favor of a deliberative procedure in accordance with justice from Chapter 4, which I will now refer to as the "imperfect reason argument," is dependent upon the notion that one should not cheat on covenants because it is impossible to be certain about one's ability to avoid detection. The search for additional benefits to justice will begin with a thought experiment that temporarily suspends this notion and asks the question, "What if I was 100% certain that I could not get caught?" I will attempt to demonstrate that there are indeed benefits to justice other than the market-based benefits, and I will examine whether there is intrinsic value in these additional benefits or in justice itself. Furthermore, I will consider the possibility that justice is more than a source of instrumental or

intrinsic value for us; I will claim that justice is a necessary and essential aspect of what it means to be a human being.

The Ring of Gyges

One of the core premises of my imperfect reason argument from Chapter 4 is that a cheater's luck will eventually run out. It may therefore be useful to conduct a thought experiment involving justice under which this premise is suspended. That is, we should ask, "How can we view the benefits (or lack thereof) of justice in a situation in which an individual could not possibly be caught violating a covenant?" The classic version of this thought experiment is found in the Ring of Gyges example from *The Republic*. In Book II, Glaucon tells the story of the Lydian Shepherd who finds a magic ring that bestows upon him the ability to become invisible at will. The Lydian shepherd uses the power of the ring to kill the Lydian king, seduce his wife and usurp his power over the kingdom. Clearly, the Lydian Shepherd has abandoned the observance of the rules of justice, and what Glaucon wants to know is, what reason could the Lydian Shepherd possibly have to do otherwise? That is, if an individual could behave in an unjust fashion without having any chance of being caught and subjected to the consequences, why should that individual behave justly?

> Indeed, every man believes that injustice is far more profitable to himself than justice. And any exponent of this argument will say he's right, for someone who didn't want to do injustice, given this sort of opportunity, and who didn't touch other people's property would be thought wretched and stupid by everyone aware of the situation...[212]

The reason for re-engaging the Lydian Shepherd at this point is not to revisit Socrates' response to Glaucon; this has already been addressed in Chapter 1, and I will not repeat it here.

[212] *The Republic*, 360 c-d, p. 1001

The Lydian Shepherd thought experiment is invoked merely as a tool in order to properly frame the inquiry into the value of justice. The Lydian Shepherd will show us that the benefits of justice are specific to the human condition as we actually find it, and not applicable to certain idealized situations where justice does not apply.

For example, Hume recognized that certain aspects of the environment in which we humans find ourselves are absolutely essential to our account of justice.[213] Our situation of limited scarcity, in which we are neither the victims of material distress nor the beneficiaries of superabundance, is one example. Hume explicitly states that if we were to find ourselves in a state of "extreme misery" such as famine, the conventions of property and justice that we have adopted would no longer be useful and would therefore not apply. Similarly, if we were to find ourselves in a state of superabundance, where every individual had everything she wanted and needed, we would have no need for justice, and our feelings of sympathetic approval for acts that are currently considered just would disappear. In other words, justice is only considered to be a virtue because it is useful to us in our current state.

The Lydian Shepherd, after obtaining the Ring, does not find himself in Hume's state of limited scarcity; his is a situation of superabundance. He can obtain whatever he wants, whenever he wants, so he has no need to enjoin the cooperation of others. The benefits of justice, in fostering a non-zero-sum game are useless to him. He is operating *outside the circumstances of justice*.

In addition to his abundance, the Lydian Shepherd also enjoys an imbalance of power which places him even further outside the circumstances of justice. One of Hobbes's primary assumptions in arguing for the value of justice was that no single human had sufficient power to

[213] Hume, *An Enquiry Concerning the Principles of Morals*, III, Part I, pp. 20-24

dominate all others. It is a fact of the human condition that even the weakest among us has the ability to kill the strongest with little difficulty, and this lack of self-sufficiency makes entering into a covenant a worthwhile undertaking. Hume also recognized that there would likely be no justice in situations of significantly unequal power:

> Were there a species of creatures, intermingled with men, which, though rational, were possessed of such inferior strength, both of body and mind, that they were incapable of all resistance…the necessary consequence, I think, is, that we should be bound, by the laws of humanity, to give gentle usage to these creatures, but should not, properly speaking, lie under any restraint of justice with regard to them…[214]

Due to the extreme power that he has over others, the Lydian Shepherd's relationship to normal humans is similar to that of normal humans with Hume's hypothetical inferior creatures. His power is such that he need not fear reprisals, so reciprocity is not relevant to him. He may choose to treat us gently, but such a choice would be that of a benevolent but all-powerful dictator, not that of a human being observing the rules of justice.

Thus, the Lydian Shepherd finds himself outside the circumstances of justice. Justice has ceased to be useful to him, so he has chosen to abandon its rules. Within the framework of the imperfect reason argument that I developed in Chapter 4, the Lydian Shepherd has made a rational choice. That is, for him, the benefits of justice no longer exceed the costs, so he would be ill-advised to continue to constrain himself according to its precepts. He can obtain whatever he wants and he need not account for the needs or reprisals of others. He has placed himself above the normal circumstances of human life, and he would view justice as a relic of his prior life which has no relevance to his current one.[215]

---

[214] Hume, *An Enquiry Concerning the Principles of Morals*, III, Part I, p. 25

[215] See Gauthier, David. "Three against Justice: The Foole, the Sensible Knave, and the Lydian Shepherd." in *Moral Dealing: Contract, Ethics and Reason*, 129-149. Ithaca: Cornell University Press, 1990), p. 147.

It is of crucial importance to note, however, that Glaucon's argument in favor of the actions of the Lydian Shepherd is far different from the one posed by Hobbes's Fool. Recall that the Fool is advancing the claim that we should selectively violate our covenants when it appears to be in our own self-interest to do so. Glaucon is not claiming that the Lydian Shepherd should violate covenants when it is to his advantage; he is claiming that the Lydian Shepherd has no reason to enter into a covenant in the first place. That is, the Lydian Shepherd should choose a deliberative procedure that recommends he avoid covenants and take whatever he needs to in order to fulfill his wants. He has no need to deceive those with whom he has made covenants; the situation he finds himself in allows him to act with impunity and it eliminates the need for covenants of any kind.

While the story of the Ring of Gyges is an interesting thought experiment, its use is limited simply because the situation of the Lydian Shepherd is not the situation in which we humans actually find ourselves. It is fair to say that the Lydian Shepherd would be foolish (at least from the standpoint of the imperfect reason argument) to obey the rules of justice, but this is not relevant to our inquiry. The Lydian Shepherd is operating outside the circumstances of justice, and this thought experiment will not apply to humans because we will never find ourselves in such a circumstance. The benefits of justice apply only in the actual circumstances of human life, and to ask justice to do more than this would be to ask for too much.

The Simulated Value of Justice

The Ring of Gyges example does not directly apply to our inquiry into the benefits of justice because it lies outside the circumstances of justice. This thought experiment does, however, lead us to consider one possible shortcoming of the imperfect reason argument.

Glaucon contends that the Lydian Shepherd is well-advised to abandon the rules of justice and take whatever he needs, regardless of whether he has to use force and fraud to do so. In the context of my imperfect reason argument, I would have to agree with Glaucon on this point, simply because the Lydian Shepherd seems to derive no material benefit from engaging in cooperation with other individuals. Yet, if we make this recommendation to the Lydian Shepherd, there is obviously something missing. It seems somewhat odd to argue that we should behave in a just way only because a weighted average probability statistical analysis demonstrates that it is in our best interest to do so, and that, if the statistics were to favor cheating, we would be well-advised to dismiss justice entirely. Although the statistical argument is valid, it is obviously failing to capture the entire essence of justice as we typically understand it.

Socrates and many others after him have argued that there are benefits to justice over and above the avoidance of the consequences of being caught behaving in an unjust fashion. In this section, I will address what those benefits are and what impact they may have on the efficacy of the imperfect reason argument. Socrates claims that behaving in a just way allows one to lead a life of fulfillment, self-confidence, social acceptance, and freedom from fear. Hume views just behavior as a route to peace of mind, integrity, and ultimately, happiness. Gauthier recognizes the value of participation in social activities, and he sees a deliberative procedure in accordance with justice as a necessary means to this end. None of these benefits were required for the formulation of the contractarian arguments of Chapters 3 and 4, but they are benefits nonetheless.

In order to properly understand this facet of the inquiry, it will be helpful to revisit a distinction made by Reeve[216] that was addressed in Chapter 1. Recall that Reeve claims that Glaucon is making an argument, not about justice, but about reputed justice. That is, all of the benefits of justice that Glaucon is interested in are "simulator accessible;" they can be attained via just behavior as well as via a *simulation* of just behavior. With Reeve in mind, it is fair to characterize the benefits of both Gauthier's contractarian argument of Chapter 3 and my imperfect reason argument of Chapter 4 as simulator accessible as well. That is, the prior arguments in favor of justice do not require an individual to act from any tuistic feelings or moral motivation to justice; the benefits necessary to validate just behavior under these models will be enjoyed by an individual who is merely going through the motions in a purely self-interested way as much as they will be enjoyed by someone who is genuinely engaged with the interests of others. The argument I will be making in this section is that, in addition to the simulator accessible (hereafter "SA") benefits of justice that are necessarily attained in the prior contractarian accounts of justice, there are non-simulator accessible (hereafter "NSA") benefits that may be attained as well.

Before moving on to the central claims of this section, however, two important aspects of NSA benefits must be clarified. First, it is important to note that NSA benefits are not the same as intrinsic benefits. Socrates attempts to demonstrate that justice is intrinsically valuable, and he largely fails. However, he is able to make a reasonable argument that justice is instrumentally valuable for NSA reasons. I do not intend to make the same mistake that Socrates does by claiming that justice is intrinsically valuable. The distinction between intrinsic and instrumental value can be quite difficult to draw, and it is not a necessary element of the argument. I do

[216] Reeve, p. 100

intend to draw a distinction between SA and NSA benefits, demonstrate that NSA benefits of justice do exist, and show that, although some NSA benefits do have intrinsic value, justice itself is an instrumental means to these intrinsic benefits and does not have intrinsic value of its own.

Second, it is important to recognize that justice will not bring NSA benefits to everyone. Thrasymachus, Hume's Sensible Knave and other antagonists do not put any value on such loftier goods. They value power, immediate gratification and material goods over friendship, happiness and integrity. However, this does not invalidate the claims that will be made in this section. Recall that Chapter 4 reached its conclusion (that justice is a rational strategy for a self-interested actor) without making any reference to these loftier benefits of justice. The arguments of Thrasymachus and other immoral interlocutors do nothing to weaken the force of this conclusion; it applies to them as well as to those who engage in justice for tuistic reasons. What I will attempt to demonstrate here is that, while everyone who behaves in a just way enjoys the SA benefits of justice, for many individuals there is a "bonus" in the form of NSA benefits as well.

The Martian Interpretation of Glaucon

Recall from Chapter 1 that Glaucon challenges Socrates to demonstrate that actually being just is superior to appearing to be just but acting unjustly. The reply that Socrates provides is quite simple; he merely points out that true justice is superior to feigned justice because the just person is the happier person. When Glaucon advocates the advantages of injustice such as power and wealth, he is using an incomplete notion of human good and ignoring some of the best things in life. The unjust tyrant lives in isolation from others and in constant fear of retaliation, and the unjust deceiver lives in constant fear of detection. The just person not only enjoys the

material benefits of justice that the unjust deceiver attains, she also enjoys freedom from fear and a genuine connection with other people, which allows her to attain a happiness that the unjust person will never know.

As noted by Reeve, Glaucon claims that the sole motivation to just behavior is the reputational benefit that is attained when one is perceived by others to be acting in a just manner. Glaucon believes that the best course of action for a self-interested individual is to appear to be just but to act in an unjust manner when it is to one's own benefit. In this way, Glaucon is advocating a course of action very similar to the one advocated by Hobbes's Fool. Glaucon is arguing, in Reeves' terminology, that all of the benefits of justice can be obtained by a mere simulation of justice.

However, Glaucon is mistaken on two levels. First, as demonstrated in my prior imperfect reason argument, a strategy of "apparent justice" is not a rational course of action because the probability of success is sometimes small and always uncertain, and the downside of being caught is potentially devastating. Second, even if an individual was able to get away with a mere appearance of justice with certainty (as the Lydian Shepherd could), this would still not be an advisable course of action because there are many benefits of justice that cannot be attained by the mere appearance of justice. That is, the class of benefits of actual justice is larger than the class of SA benefits. An individual who acts in a truly just manner will enjoy all of the benefits of simulated justice as well as many other NSA benefits that the pretender can never enjoy.

What exactly are these NSA benefits? The SA benefits of justice were addressed in the prior contractarian arguments and they are easy to see and to describe. Free market commerce, protection from physical harm, and confidence in the possession of one's property are all

benefits of cooperation in a non-zero-sum game, and they are all available to a truly just person as well as to one who is merely simulating justice. The NSA benefits of justice, on the other hand, tend to be based upon an individual's internal state of affairs and are therefore less obvious and more difficult to categorize. Fortunately, an analogy employed by Philippa Foot will help to shed some light on precisely what it is that simulated justice cannot provide.

Foot directly addresses the arguments of Thrasymachus and Glaucon regarding justice with an analogy about friendship in which she imagines what Martians would think if they viewed human acts of friendship from an unfamiliar third-person perspective. These Martians may imagine that it is better to pretend to be a friend and to enjoy the benefits of friendship without reciprocating, but to view friendship in this way is to miss the point:

> These Martians would see friendship very much as Plato's immoralists see justice. In itself acting as a friend is, the Martians suppose, disagreeable, like gymnastic exercise or medical treatment. For the run of humans it is, however, worthwhile for its rewards. Were it possible to get these rewards by gaining the reputation of being a friend without really accepting its duties, that is what any human would seek. The point of my analogy lies, of course, in the fact that these Martians would be *failing to understand* what friendship actually is in human life…A Thrasymachean view of friendship would instantly be recognized as wrong.[217]

Foot recognizes that there are benefits to friendship that are not simulator accessible. Merely going through the motions may give an individual some of the benefits of friendship, but the greatest benefits of friendship can be enjoyed only from the inside,[218] out of the view of Martians but within the empathetic understanding of other humans, and these benefits are available only to those who are truly friends and not just faking it. Likewise, to obtain a reputation for justice without actually being just will deprive an individual of justice's most profound and lasting benefits. Singpurwalla very effectively captures the essence of what the

[217] Foot, Philippa. *Natural Goodness.* New York: Oxford University Press, 2001: 101-2

[218] For more on the idea that the importance of justice is given to us "from the inside," see Thompson, Michael. "Three Degrees of Natural Goodness (Discussion Note, *Iride*)." URL=http://www.pitt.edu/~mthompso/three.pdf (retrieved August 14, 2012)

Martians are missing when she characterizes the NSA benefits of justice as a fulfillment of our need to be connected w/ others:

> Socrates thinks that we have a reason to behave justly because behaving justly is necessary for fulfilling a deeply important need that we all as social creatures have, namely, the need to be connected or unified with other people.[219]

The basic human need to be unified with others is something that we all intuitively recognize, but it is an aspect of justice and friendship that cannot always be seen from the outside. Socrates clearly recognizes this need, and he emphasizes the unhappy life of the tyrant to illustrate that even though an individual might be able to attain many types of goods without adhering to the requirements of justice, an unjust individual will be lacking the personal connections and unity with others that lead to the greatest kind of happiness.

The Sensible Knave

Where Socrates has Glaucon and Thrasymachus, and Hobbes has his Fool, Hume has the Sensible Knave. Like the other antagonists, the Knave advances an argument against justice which demands a reply. He claims:

> That *honesty is the best policy*, may be a good general rule; but is liable to many exceptions: And he, it may, perhaps, be thought, conducts himself with most wisdom, who observes the general rule, and takes advantage of all the exceptions.[220]

The Lydian Shepherd has no use for covenants because his power allows him to take what he wants with impunity. The Knave, like the Fool, has no such power. Like the Fool, the Knave is arguing that his lack of overwhelming power makes it rational for him to enter into a covenant and to violate the covenant when it is highly certain that he can get away with it. Yet, despite the

[219] Singpurwalla, p. 276
[220] Hume, *An Enquiry Concerning the Principles of Morals*, IX, Part II, p. 81

fact that the reasons for the Knave's claims are different from those of the Lydian Shepherd, the argument that Hume levies against the Knave includes elements of Socrates' reply to the Shepherd as well as Hobbes's reply to the Fool:

> Inward peace of mind, consciousness of integrity, a satisfactory review of our own conduct; these are circumstances very requisite to happiness, and will be cherished and cultivated by every honest man, who feels the importance of them.
>
> Such a one has, besides, the frequent satisfaction of seeing knaves, with all their pretended cunning and abilities, betrayed by their own maxims; and while they purpose to cheat with moderation and secrecy, a tempting incident occurs, nature is frail, and they give in to the snare; whence they can never extricate themselves, without a total loss of reputation, and the forfeiture of all future trust and confidence with mankind.
>
> But were they ever so secret and successful, the honest man, if he has any tincture of philosophy, or even common observation and reflection, will discover that they themselves are, in the end, the greatest dupes, and have sacrificed the invaluable enjoyment of a character, with themselves at least, for the acquisition of worthless toys and gewgaws[221]

The Knave is making the same mistake as Hobbes's Fool, as he fails to recognize his own overconfidence and the role that randomness plays in his ability to successfully deceive. However, Hume takes the argument against his interlocutor further than Hobbes does. Hume shows us that, not only does the Knave have a high likelihood of being caught; even if his deception is successful, the Knave will miss out on the higher goods that Socrates emphasizes, such as character and integrity.[222] That is, in his response to the Knave, Hume emphasizes the Knave's loss of both SA and NSA benefits.

Some scholars have argued that Hume dismisses the Knave too quickly.[223] The fact that the Knave is willing to sacrifice long-term integrity for the immediate gratification of his lower-

[221] Hume, *An Enquiry Concerning the Principles of Morals*, IX, Part II, p. 82
[222] See Kavka (1995), p. 26
[223] See Postema, Gerald. "Hume's Reply to the Sensible Knave." *History of Philosophy Quarterly* 5, no. 1 (Jan., 1988): 23-40 and Krause, Sharon. "Hume and the (False) Luster of Justice." *Political Theory* 32, no. 5 (Oct., 2004): 628-655

level desires signifies that his "toys and gewgaws" have more value for him than having an admirable character, and this is not necessarily irrational. Hume himself says it is, "not contrary to reason to prefer the destruction of the whole world to the scratching of my finger."[224] It would therefore be arbitrary for us to argue that the Knave is irrational in preferring material wealth to the loftier desires espoused by Socrates, and it is quite likely that the Knave would be unconvinced by any value-based argument we could throw at him.

While it is probably true that the Knave will not be convinced by Hume's argument, this fact does nothing to undermine the core of the argument itself. Hume's intention is not to convince the Knave of the error of his ways; the Knave is probably beyond convincing. By denying the value of integrity, the Knave demonstrates that he does not understand what justice is. Like Thrasymachus, he does not value being connected with others. Like Foot's Martians, he cannot see the difference between actually being a just person and simply faking it. Hume is merely trying to augment an already strong argument in favor of justice by appealing to those of us who do see value in the "enjoyment of a character."[225] While it is true that justice has only SA value for the Knave, it does not follow from this that it has only SA benefits for everyone. In fact, justice does have NSA benefits for most people who are not invisible shepherds, arrogant fools, ignorant knaves or Martians. Those of us who do appreciate the NSA benefits of justice are the proper audience for Hume's argument against the Knave.

Furthermore, it should be noted that even though the Knave does not value the NSA benefits of justice such as integrity and character over SA benefits such as toys and gewgaws, it is still rational for him to adopt a deliberative procedure of adherence to covenants for the reasons given in the imperfect reason argument of Chapter 4. The Knave will benefit from

[224] Hume, *A Treatise of Human Nature,* 2:2.3.4, p. 267
[225] *An Enquiry Concerning the Principles of Morals,* IX, Part II, p. 82

adherence to his covenants because he will enjoy the SA value of justice without having to fear the consequences of being discovered. For those of us who do value character and integrity, we can enjoy these NSA benefits as well as the SA benefits that accrue to the Knave. In some way, we may be able to think of the NSA benefits of justice as a "bonus" that is available to some of us in addition to the SA benefits that are available to everyone.

Gauthier and the Liberal Individual

The distinction between SA and NSA benefits is a consistent theme that runs from Socrates to Hume and beyond. Gauthier explicitly recognizes this distinction and he illustrates it brilliantly via his comparison of economic man and the liberal individual. Recall that, in making his main argument for the rationality of justice in *Morals by Agreement*, Gauthier assumes that humankind is self-interested and non-tuistic. He refers to this caricature of humanity as "economic man." Economic man is similar to Glaucon's concept of the just individual; he understands the need to accept constraints, but his emotions and affections are not engaged by these constraints.[226] He is not truly a just individual because he is motivated by considerations other than justice itself. He views his lack of overwhelming power and self-sufficiency as an evil because he derives no value from interaction with other humans and he would therefore prefer to be able to act unilaterally in the fulfillment of his needs and desires. Participation in social activities serves as a reminder of this weakness and of his dependence on others. He only derives instrumental value from interaction, and this limits the benefits of his participation in cooperative activities to SA benefits alone. As Gauthier states:

[226] Gauthier (1986), p. 328

> Morals by agreement may indeed be the only morality that economic man can understand, but their value to him is lessened by his indifference to many of the activities that they help to make possible.[227]

Gauthier recognizes that economic man is clearly not an accurate representation of an actual human, so he contrasts economic man with another, more realistic caricature known as the "liberal individual." The liberal individual is tuistic and she has an emotional, affective capacity; like actual humans she takes an interest in the interests of other individuals and she derives pleasure from social interaction. In Gauthier's terms, she values *participation*. Where economic man views constraints on his actions and his dependence on others negatively, the liberal individual places a positive value on the participation that she has with others, both instrumentally and intrinsically, and she also values the constraints of justice that make this participation not only possible, but necessary.[228] In *Moral Dealing*, Gauthier writes:

> Rather than chafe unwillingly under the constraints of justice, the liberal individual recognizes that an essentially just society provides the conditions necessary to realize her own good through free participation in fair cooperation with her fellows. To the liberal individual, human relationships in a just society are not exclusively or even primarily contractual, but they offer the respect for each individual's good, the assured mutuality of benefit, and the freedom from exploitation that voluntary, rational agreement would guarantee.[229]

The liberal individual enjoys the same contractual benefits of society that are enjoyed by the rational capacity of economic man, with the added benefit of having her emotional capacity engaged by the process of interaction itself.

Gauthier's discussion of the liberal individual has several implications for the continuation of our own discussion of NSA benefits and of justice in general. First, it is important to understand Gauthier's distinction between instrumental and intrinsic value. What

[227] Gauthier (1986), p. 326
[228] Gauthier (1986), p. 347
[229] Gauthier (1990A), p. 6

the liberal individual shows us is that contractarianism does not necessarily require that justice lead only to instrumental benefits. In fact, the instrumental rational foundation of justice is a necessary condition for the formation of the affective hold that it has on us; if justice was irrational we would not be able to link it to an intrinsically valuable concern for others.

Gauthier believes that participating in cooperative interaction and taking an interest in the interests of others are intrinsically valuable activities, however, he stops short of making a firm claim that justice itself is intrinsically valuable. He makes a vague claim that justice *may be* intrinsically valuable,[230] but offers no real proof of the distinction, and his more important point is that justice is an instrumental catalyst that provides us with some additional NSA benefits that are of intrinsic value. He envisions humankind as a middle ground between economic man, who values activities only instrumentally, and a "utopian man," who is free of scarcity and therefore values activities only intrinsically. Gauthier claims that the liberal individual values activities for both reasons, and it is in this context where justice can be found. For Gauthier, intrinsic and instrumental value are inter-dependent, and although many NSA benefits of justice are intrinsically valuable, justice itself is not.

Second, it should be recognized that Gauthier's liberal individual is the antithesis of the Sensible Knave. The Knave is, in a way, a version of Gauthier's economic man with poor reasoning ability; he thinks he can get away with violations of covenants, although in actuality he cannot. If we add proper reason to the Knave, we get economic man, who adheres to his covenants, but only for the value of SA benefits. If we add a recognition of the value of NSA benefits to economic man, we arrive at the liberal individual. Where the Knave fails to recognize the value of tuistic participation, the liberal individual embraces it. The liberal individual is

[230] Gauthier (1986), pp. 326-327

therefore able to benefit from justice in all of the same ways as the Knave and economic man, with the added benefit of deriving value from NSA activities as well.

Third, and most important, Gauthier's characterization of the liberal individual hints at a critical element of justice that has not yet been addressed, namely, the idea that justice is an essential aspect of being human. Gauthier recognizes that economic man is not an accurate depiction of the way we actually are, but his portrayal of the liberal individual is certainly reminiscent of humans as we actually find them. By inextricably associating justice, not only with participation in social activities, but with seeking and striving for intrinsic goods, Gauthier is suggesting that justice is woven into the very fabric of the human constitution. The liberal individual necessarily recognizes her need for seeking, striving and social participation, and justice is a necessary prerequisite for these activities. Gauthier claims that the liberal individual "does not see a self-sufficient life as fully human,"[231] which suggests that, for Gauthier, justice may be a prerequisite for consideration as a normal human being. This casual afterthought of Gauthier's demands further attention, and fortunately receives it in the work of Philippa Foot, for whom naturalism is the central idea of moral philosophy.

Justice and Natural Normativity[232]

In Chapter 1, I claimed that an understanding of justice must be based upon an understanding of human nature. As we will now see, it may even be the case that justice itself is actually a necessary part of our human nature. Via the liberal individual, Gauthier has argued that there is something more to justice than merely instrumental SA benefits, and he subtly

[231] Gauthier (1986), p. 325
[232] An earlier draft of this section benefitted significantly from the commentary of Professor Brook Sadler.

suggests that justice may be even more than that; it may actually be a natural and essential part of who we are. The controversial claim that justice is natural has been vigorously debated for centuries. This claim has been denied by Thrasymachus and more recently by his 19th century counterpart, Nietzsche, and it finds it most ardent advocate in Philippa Foot. By contrasting Foot's naturalistic account of justice with the accounts given by Nietzsche and Socrates' antagonists, I attempt to demonstrate that justice is indeed an essential element of human nature.

Thrasymachus and Nietzsche

Recall from Chapter 1 that Thrasymachus believes that any concept of justice is temporal; it derives its meaning and validity from the prevailing public opinion of the time and it is not based upon any underlying facts about human nature. He believes that justice cannot be viewed as good or bad without qualification; my justice benefits me and damages you, and your justice benefits you and damages me.[233] Individuals who obey the rules of justice purely out of respect for justice itself are placing an *unnatural* restraint on themselves, to their own detriment.

As noted by Chappell,[234] Socrates and Thrasymachus agree that justice must be justified by reference to some notion of human *flourishing* which must either be helped or hindered by an adherence to justice. Although they would agree that justice is beneficial only if it assists in the flourishing of the individual, they ultimately disagree on the efficacy of justice due to their disagreement regarding what it is that constitutes human flourishing. For Thrasymachus, flourishing consists in having as much wealth and power over others as possible, and he sees justice as an obstacle to these goals. Socrates, as noted in Chapter 1, sees a flourishing human

[233] See Chappell, T.D.J. "The Virtues of Thrasymachus." *Phronesis* 38, no. 1 (1993): 14
[234] Chappell, p. 5

life as one that involves happiness, philosophical reflection, clear conscience and the respect of others, and justice is clearly a necessary means to this end. Where Thrasymachus sees justice as a contribution to the flourishing of others, Socrates sees it as a contribution to the flourishing of oneself.

The argument posed by Thrasymachus is brief and it is ultimately circumvented by Socrates. However, the argument is revived and restated in a similar fashion centuries later in the works of Nietzsche.[235] Like Thrasymachus, Nietzsche provides his own definition of justice as a foil to the idea of justice as it is typically construed:

> At the risk of displeasing innocent ears I propose: egoism belongs to the nature of a noble soul – I mean that unshakable faith that to a being such as "we are" other beings must be subordinate by nature and have to sacrifice themselves. The noble soul accepts this fact of its egoism without any question mark, also without any feeling that it might contain hardness, constraint or caprice, rather as something that may be founded in the primordial law of things: if it sought a name for this fact it would say, 'it is justice itself.'[236]

For Nietzsche, as for Thrasymachus, justice (as it is commonly understood) is a contemptible practice that obstructs human flourishing. He argues that the idea that justice is absolute and unchanging across time and cultures is absurd. For him, true justice does not mean equality; it is a way of defending one's own vantage point and it is an ever-changing standard that adapts to suit the needs of particular individuals. It means different things to different people and it recognizes that different people should be treated differently. When we evaluate the justice or injustice of any action, it is not the action itself that is at the core of the evaluation, but rather the nature of the person who performs the action.[237]

---

[235] It must be acknowledged that the work of Nietzsche is subject to various interpretations. However, for the purpose of illustrating the argument in favor of the "justice is irrational" position, I will interpret Nietzsche from a literalist viewpoint.

[236] Nietzsche, *Beyond Good and Evil*, 9:265, p. 405

[237] This point is reminiscent of Aristotle's discussion of distributive justice in Book V of *Nichomachean Ethics*.

When Nietzsche addresses the common understanding of justice (justice as fairness or as the observance of covenants), he adopts the Thrasymachean view that adherence to this conventional notion of justice is inconsistent with the flourishing of the best individuals among us and is therefore foolish. Nietzsche believes that all of the moral virtues, including justice, are the result of what he refers to as "slave morality," in which the weak and powerless have turned nature on its head and imposed a suffocating moral constraint upon the great and powerful and made those individuals less than they would otherwise be. In *Genealogy of Morals*[238] he ridicules humble and fearful individuals for promoting their adherence to justice as a virtue and for attempting to convince themselves that their support for justice is not merely an admission of impotence. When he describes the way individuals of merit behaved prior to the imposition of slave morality, Nietzsche sounds remarkably similar to Thrasymachus:

> Human beings whose nature was still natural…hurled themselves upon weaker, more civilized, more peaceful races…[239]

> …the noble caste was always the barbarian caste: their predominance did not lie mainly in physical strength but in strength of the soul - they were more *whole* human beings (which also means, at every level, "more whole beasts").[240]

At this point Nietzsche takes the argument against justice farther than Thrasymachus was willing (or permitted) to go. He overtly supports the exploitation of the weak by the strong and he contends that placing one's interests on a par with another and refraining from doing what one wants and is able to do by force and cunning is a denial of flourishing and a move towards death and decay. He claims that the main purpose of society is to foster the development of a higher type of individual, or "overman," and that the practice of justice as it is currently espoused by the masses is a hindrance to this goal.

---

[238] Nietzsche, Nietzsche, Friedrich. "On the Genealogy of Morals." in *Basic Writings of Nietzsche*, edited by Walter Kaufmann. New York: The Modern Library, 1992, 1:13, p. 482
[239] Nietzsche, *Beyond Good and Evil*, 9:257, p. 391
[240] Nietzsche, *Beyond Good and Evil*, 9:257, p. 392

It is important to note that Nietzsche and Thrasymachus are not immoralists. They are not arguing for a complete abandonment of values; what they seek is a revaluation of values. Nietzsche clearly values struggle and suffering in the development of the overman. They both value wealth and power of the will. They have their own set of values that they implicitly and explicitly support, and they are making a claim about the way in which humans should live in order to flourish, given these values.[241]

In summary, Nietzsche strongly believes that adherence to conventional principles of justice is not a rational practice for all individuals; justice is instead the "silly good nature"[242] of the weak, naive and cowardly. While the weak have an interest in promoting justice in order to protect themselves from the ravages of the strong, an observance of justice is not in the best interests of the best individuals among us, and nature dictates that they should instead practice injustice whenever possible.

Philippa Foot

Foot, like Thrasymachus and Nietzsche, believes that justice is closely tied with a notion of human flourishing. However, the similarities between her ideas regarding justice and the ideas of Thrasymachus and Nietzsche largely end there. Foot's moral theory, as developed in *Natural Goodness,* can be characterized as an attempt to base morality in facts, and specifically in facts about human life. She describes her theory in terms of a response to two questions, namely:

[241] See Foot, Philippa. "Nietzsche: The Revaluation of Values." in *Virtues and Vices*, 81-95. Oxford: Clarendon Press, 2002
[242] Foot (2001), p. 100.

1) Can we develop a factual definition of what it means to be a good human being, and
2) Can we demonstrate that each of us has a reason to act in accordance with this definition of what it means to be a good human being?[243]

She begins her response to the first question by saying that she wants to "describe a particular type of evaluation and to argue that moral evaluation of human action is of this logical type."[244] The type of evaluation that she wants to describe is one in which facts about a particular subject matter are outlined, and the argument that she makes is that moral arguments are merely descriptions of facts about the subject of human life.

Her argument is a response to the non-cognitivist view that moral evaluations are not grounded in fact, but are instead emotive utterances or expressions of subjective values. Where the non-cognitivists claim that there is a gap between the values expressed by a moral judgment and any facts upon which one attempts to ground such a judgment, Foot denies the existence of such a gap.

Foot argues that human goodness is analogous to the goodness that can be seen in plant and animal life. Where a good tree is one with deep, strong roots and a good lion is one that is fast enough to catch sufficient prey to feed itself and its offspring, so a good human is one who has the ability to think clearly, to recognize patterns, and who recognizes virtues such as friendship, loyalty, and justice. Clearly, the move from the human ability to recognize patterns to the recognition of justice as a virtue is a bold one, but Foot describes this move as just part of the continuum of "natural normativity." She wants to show that, since we would view a human with no sense of justice whatsoever as being defective, and since humans need such virtues as

[243] This is a paraphrase of the questions posed by Gary Watson, cited on p. 53 of *Natural Goodness.*
[244] Foot (2001), p. 3

justice in order to survive and flourish, the term "good" as used to describe good moral actions in humans is no different from the term "good" when it is used to describe the characteristics of a plant or the actions of an animal. In each instance the term "good" is grounded in *facts* that are relevant to the survival and flourishing of a particular organism.

Thus, Foot's theory is attempting to integrate morality and justice with the survival and evolution of the human species. "Good" actions for any individual animal are determined by what will allow that individual to reproduce and to flourish:

> 'natural' goodness, as I define it, which is attributable only to living things themselves and to their parts, characteristics, and operations, is intrinsic or 'autonomous' goodness in that it depends directly on the relation of an individual to the 'life form' of its species.[245]

In the case of the human animal (an animal which is unique in that it has the ability to act based upon reasons), the actions that allow it to flourish can be described in factual *moral* terms.

Having addressed the first question, Foot then attacks the second: Can we demonstrate that each of us has a *reason* to act in accordance with this definition of what it means to be a good human being? She responds to this question with a brief logical progression: That which is good for human life is moral, those actions which are moral are also rational, and to act rationally is to act on reasons. In other words, goodness defines practical rationality, and not vice-versa. We can describe in factual terms what the good action is, and the fact that the action is good for human beings means that it is rational for a human being to perform the action.

She then anticipates the response of a committed skeptic who may yet ask, "What reason do I have for acting rationally?" Foot merely claims that this is an incoherent question:

[245] Foot (2001), pp. 26-27

> To ask for a reason for acting rationally is to ask for a reason where reasons must a priori have come to an end. And if (the skeptic) goes on saying 'But why *should* I?', we may query the meaning of this 'should'.[246]

Foot argues that we can develop a factual definition of what it means to be a good human being, and we can offer reasons why we should act in accordance with this definition. Stated concisely, to act well is to act rationally, and to act rationally is to act on reasons. For her, no further explanation is necessary.

Justice as a Part of Our Nature

It is clear that Foot was influenced by Nietzsche, and it will be helpful to outline some of the common threads that run through both of their philosophies before turning to the more important topic of their differences on the issue of justice. First, both Nietzsche and Foot are proponents of "naturalism" in the sense that humans should be viewed as a part of the continuum of the animal world. As mentioned before, Foot's moral theory has an evolutionary (or at least biological) tilt to it, and it is also possible to characterize Nietzsche as an evolutionist, although more in the Lamarckian sense as opposed to the Darwinian.

In addition, the recognition of humans as part of the animal kingdom allows both Nietzsche and Foot to describe moral action in terms of facts about human life and characteristics that promote the flourishing of the human species. As previously mentioned, Nietzsche's entire project is based upon the primary underlying value of a constant struggle towards a higher type of human. His is a philosophy that praises progress, strength, health, and general self-creation. Foot's theory is also explicitly grounded on the progress of the species:

[246] Foot (2001), pp. 65

> The way an individual *should be* is determined by what is needed for development, self-maintenance, and reproduction: in most species involving defence, and in some the rearing of the young.[247]

She sees the goal of morality as being the promotion of a general flourishing of the individual and the species, and, like Nietzsche, she argues that actions which inhibit the flourishing of the species are morally wrong.

While these similarities can be informative when tracing the genealogy of Foot's ideas, the profound differences between Nietzsche and Foot provide a far more robust basis for analysis. One obvious point about which Foot and Nietzsche disagree is the role that rationality has for morality. Foot bases her entire moral theory on the argument that reason and rationality are a necessary condition of morality. For her, morality and reason both play a part in what it means to be a member of the human species.

Nietzsche's position on rationality is entirely different: "*there is* a realm of truth and being, but reason is *excluded* from it!"[248] Nietzsche's critique of the capacity of reason and his preference for the aesthetic is well known. While he and Foot come to the same conclusion that the moral act is that act which encourages human flourishing, Nietzsche sees this flourishing not as a rational process, but as an aesthetic one. This particular aspect of Nietzsche's philosophy encourages Foot to refer to his evaluation as being aesthetic and not moral:

> Thus Nietzsche thinks of value as belonging only to a person who has created his own character in a pattern that cannot be prescribed for others, and it is here that his shift from a moral to an aesthetic form of evaluation becomes clear.[249]

Foot's claim here, that Nietzsche's evaluation is not a moral evaluation, is dubious. It is more accurate to describe Nietzsche's evaluation as a moral evaluation done in an aesthetic rather than

[247] Foot (2001), p. 65
[248] Nietzsche, *On the Genealogy of Morals*, 3:11, p.554.
[249] Foot, Philippa. "Nietzsche's Immoralism." in *Moral Dillemas*, 144-158. Oxford: Clarendon Press, 2002: 148

in a rational framework. Regardless of how it is characterized, however, Nietzsche's aesthetic perspective is in stark contrast to the purely rational perspective that Foot adopts throughout her moral theorizing.

Another profound difference between the ideas of Foot and Nietzsche is their respective positions on egoism. Early in her argument, Foot explicitly recognizes that some species *necessarily* operate in a group context, and the fact that they do so will have profound implications for the types of actions that will contribute to their flourishing:

> In most cases we speak of what each member of the species needs to be and to do in order that *it* should flourish. But of course what is needed may be needed in a group, like cooperation in a pack, or obedience to a leader, and what a member of the species is or does may advantage others rather than himself.[250]

Foot argues that, because humans necessarily behave in a cooperative fashion, those individual humans who do not behave cooperatively can be considered defective. She also emphasizes the point that things like friendship, loyalty, truthfulness, and especially justice are a necessary part of the human condition because adherence to these virtues has allowed us to flourish as a species. In other words, it is *rational* to observe these virtues in our dealings with others because, in doing so, it promotes the progress of the human species.

Because she recognizes collective behavior as being a necessary part of the human condition, Foot has to criticize egoism for failing to account for this fact. She characterizes the egoist as a free rider, that is, someone who selfishly enjoys the benefits of the group-directed behavior of others without contributing to the benefit of the group him- or herself. For Foot, the human who does not contribute to the well being of the human species as a whole (by practicing virtues such as friendship, loyalty, and justice) is *defective* in the same way that a wolf that does

[250] Foot (2001), footnote, p. 33

not aid in the hunt is defective. In the context of her moral theory, this defective behavior should be considered rationally and objectively wrong.[251]

While Foot explicitly denounces egoistic moral theories, egoism is a crucial part of Nietzsche's philosophy. Like Foot, Nietzsche recognizes that humans often operate in a group environment, but unlike Foot, he sees the "herd" as a hindrance to the development of a better type of individual because herd behavior leads to a reduction of individual creativity and critical thought:

> "The *over-all degeneration of man* down to what today appears to the socialist dolts and flatheads as their 'man of the future' – as their ideal – this degeneration and diminution of man into the perfect herd animal..."[252]

Nietzsche is the epitome of the rugged individualist. He praises the individual who makes her own rules and who questions the prevailing cultural and social norms. He rejects the idea that actions themselves can be characterized as good or bad, instead favoring the view that *individuals* are good or bad, and that any action performed should be judged in the context of the person who performs it. Stated in contemporary terms, he despises those who *outsource* the guidance of their behavior to an external authority, whether it be a boss, family member, or priest, and instead prefers those egoistic individuals for whom moral action is defined as "that which is right for *me*, *now*." [253]

Because he accepts egoism, Nietzsche emphasizes the flourishing of the individual as opposed to the group, and even at times encourages the flourishing of the individual *at the*

[251] At this point the determined egoist may still try to accommodate pro-group behavior in an egoistic framework by arguing that an individual defers to the group at times because it benefits the individual in the long run, or because it makes the individual feel good. A detailed analysis of the egoistic position is beyond the scope of this essay, however.

[252] Nietzsche, *Beyond Good and Evil*, 5:203, p. 308

[253] I emphasize the term "me" because "that which works now" will be different for any given individual versus another individual. The term "now" is emphasized because "that which works for me" will evolve over time.

expense of the group. His acceptance of self-interest as a virtue in the ascending individual is quite explicit:

> Self-interest is worth as much as the person who has it: it can be worth a great deal, and it can be unworthy and contemptible.[254]
>
> Faith in oneself, pride in oneself, a fundamental hostility and irony against 'selflessness' belong just as definitely to noble morality as does a slight disdain and caution regarding compassionate feelings and a 'warm heart.'[255]

It is evident that Nietzsche stands in stark opposition to Foot with respect to the role that collective behavior plays in the development of the species. Foot contends that group behavior is a necessary part of the human condition, while Nietzsche is convinced that it is precisely this group behavior that is holding us back and preventing humankind from realizing its full potential.

Given the apparent hopelessness of reconciling these two related but sharply opposed philosophies, we should instead focus our efforts on deciding which of them presents the better argument on the nature of justice. Since both Foot and Nietzsche base their ideas of morality upon facts about human life, the proper question to ask is, "Who has the facts right?"

Nietzsche's moral philosophy (as characterized above) clearly does not count justice among its virtues. For Nietzsche, justice is part of slave or herd morality and he sees justice as being antithetical to the ascendancy of the better type of individual. In a Nietzschean framework, justice is only favored by the weak because they lack the strength to protect themselves against the injustice of the strong. While the weak clearly have an interest in favoring justice as a virtue, according to Nietzsche, the strong have no interest whatsoever in promoting justice; those who have the power should simply do as they please. In this respect, Nietzsche is in agreement with

[254] Nietzsche, Friedrich. "Twilight of the Idols." in *The Portable Nietzsche*, edited by Walter Kaufmann. New York: Viking Penguin, 1976, 33, p. 533
[255] Nietzsche, *Beyond Good and Evil*, 9:260, p. 395

some of the antagonists from Plato's dialogues who argue that justice is a convenient rule for a certain type of human rather than a universal prescription.[256]

Foot, on the other hand, argues not only that justice must be a universal prescription, but that justice is a necessary part of whom we are, that is, that justice is part of being human. With the Martian analogy previously mentioned in this chapter, Foot discredits the position of Plato's antagonists and Nietzsche, and shows us via this analogy that justice, like friendship, cannot be evaluated merely in a "cost-benefit" framework in which we view the virtue of justice as a necessary sacrifice that must be undertaken in order for us to get what we want. Justice means much more than this to us in the conduct of our day-to-day lives. As social animals we need justice in order to cooperate with each other and to help each other feed, clothe, house, and educate ourselves. In other words, we need justice in order to flourish. A certain inherent sense of the goodness of justice is innate to us as humans; we necessarily place value on justice and we take pleasure in performing just acts and seeing them performed by others.[257] Stated in terms of Foot's overall moral theory, a person who does not recognize the inherent value of justice would most certainly be seen by the rest of us as being *defective*.

Returning to the question, "Who has the facts right?" it seems clear that Foot has given us a more accurate characterization of the facts as we experience them from day to day. To deny that justice is a sentiment that is common to us all is to deny the facts as we find them in everyday life as humans. In addressing her differences with Nietzsche on this point, she accuses him of poor psychological analysis: "Nietzsche seems to have fallen into the trap of working a modicum of psychological observation into an all-embracing theory which threatens to become

[256] For more on the Platonic antagonists, see Chapter 7 of *Natural Goodness.*

[257] Empirical studies of infant and toddler behavior have suggested that humans have an inherent preference for pro-social (cooperative) individuals over antisocial (uncooperative) individuals. See Hamlin, J. Kiley, Karen Wynn, Paul Bloom, and Neha Mahajan. "How Infants and Toddlers React to Antisocial Others." *Proceedings of the National Academy of Sciences* 108, no. 50 (Dec. 13, 2011): 19931-19936

cut off from facts that could possibly refute it."[258] One need only imagine a life without justice to see that such a life would certainly not allow us to flourish or attain ascendancy, and it is likely that we would not even describe such a life as fitting our definition of what it means to be human.[259] We should instead recognize that a sense of justice is a necessary human capacity akin to the ability to recognize patterns or to use tools. Foot's closing statement in Chapter 7 of *Natural Goodness* states the central claim of her argument quite explicitly: "My point is that it is only for a different species that Nietzsche's most radical revaluation of values could be valid. It is not valid for us as we are, or are ever likely to be."[260]

Conclusion

Gauthier's economic man is well-advised to adopt a deliberative procedure in accordance with justice because such a deliberative procedure will allow him to reap the benefits of cooperation in a non-zero-sum game, and because the expected return of cheating on his agreements is negative. Yet, the justice of economic man does not capture the entire essence of what justice means to actual humans. As Socrates, Hume and Gauthier's liberal individual have shown, there are NSA benefits to justice other than the benefits that lead economic man to behave in a just fashion. The benefits enjoyed by economic man make justice a worthwhile policy for everyone, and the additional benefits enjoyed by the liberal individual provide a bonus for most of us.

[258] Foot. "Nietzsche's Immoralism." p. 156
[259] Note that Foot's view of the natural state of humans is quite different from the state of nature described by Hobbes. Where Hobbes sees justice as a means of escape from the state of nature, Foot sees justice as absolutely essential to our natural state.
[260] Foot (2001), p. 115

While it is clear that these NSA benefits do exist, we must be careful not to equate the presence of NSA benefits with proof that justice itself has intrinsic value. Justice does lead to happiness, participation and a sense of integrity, and it can be said that all of these things have intrinsic value. Yet this does not demonstrate that justice *itself* has intrinsic value. Rather than continue Socrates' unsuccessful effort to demonstrate the intrinsic value of justice, we are going to have to settle for something less: Justice itself will not provide us with intrinsic value, but it is a short ride from the end of the line. That is, there is nothing standing between justice and intrinsic value, and justice is a very effective catalyst for the realization of other benefits that do have intrinsic value. However, to attribute more than this to justice would be to overstate the case.

Not only does justice provide us with NSA benefits and help us attain goods with intrinsic value, it also seems to be an essential part of human nature. Foot has made a convincing argument in favor of this position; any individual who claims to have no sense of justice in her interaction with others would be viewed as somehow defective in the context of what we consider to be a normal, thriving human life. In fact, it is worthwhile to consider the possibility that Nietzsche has got the causal relationship between justice and humankind completely backwards. He claims that justice was created by weaker individuals for the purpose of restraining stronger individuals. In his view, the aristocratic individual of years past lived without the need for justice, and the justice that is currently imposed on the better individuals among us is artificial and contrived. However, this line of argument seems to have reversed the causal relationship that exists between the practice of justice and human beings as we find them. Nietzsche is arguing that humans created justice when it is more accurate to say that justice is a

necessary aspect of the species we currently know as human.[261] That is, human beings as we find them in everyday life, whether they are mediocre or exceptional, owe their existence to various cooperative group behaviors, including justice. Without justice, we never would have been able to become the reasoning, aesthetic creature of the will that Nietzsche embraces. To dismiss justice as a contrivance is, therefore, to completely ignore the fact that justice helped to make us what we are as a species. The human species as we know it is more a product of justice than a creator of it.

[261] Foot uses an argument based upon natural normativity to claim that justice and other virtues are an integral part of human life. Haidt supports a related claim, but his is more explicitly based on evolution and moral intuition. See Haidt (2012).

CONCLUSION

The motivating goal of this thesis has been to demonstrate that justice, understood as the consistent observance of cooperative agreements, is a rational strategy for a self-interested actor, despite its costs. I want to lend credence to our intuitions about just behavior by showing that following these intuitions is in our best interests, and that the Madoffs of the world are acting irrationally when they willingly violate the covenants they have made.

My thesis begins, appropriately, at the origin of western philosophy. In *The Republic*, we are introduced to the argument in favor of justice and to the amoral interlocutor via the character Thrasymachus. The primitive claim made by Thrasymachus and the more sophisticated subsequent arguments of Glaucon and Adeimantus, (as well as Hobbes's Fool, the Sensible Knave, Nietzsche and the straightforward maximizer) represent the opposing position against which I have argued throughout the dissertation.

Although Socrates' argument in *The Republic* is unsatisfying in many ways, he does introduce us to some key concepts that are critical for the formulation of my case in favor of justice. First, in his refutation of Thrasymachus, Socrates implies that the practice of justice is not a zero-sum game. All of the amoral interlocutors from Thrasymachus to Nietzsche will claim, correctly, that the observance of justice comes with certain costs,[262] in the form of restraint from certain behaviors. In order for me to claim that justice is a rational strategy, it is necessary to show that there are benefits to justice that outweigh these costs. What

[262] These costs can either be "out-of-pocket costs," which involve the renunciation of goods that one already has, or "opportunity costs," which involve the renunciation of goods that one could have had.

Thrasymachus fails to understand is the fact that the benefits that one individual enjoys as a result of the practice of justice do not necessarily lead to costs for some other party. That is, cooperative practices involve more than the mutually agreeable division of a fixed basket of goods; the process of cooperation itself actually increases the overall value of the goods to be divided, so everyone is better-off under the mutual observance of justice than they would have been if they had all acted independently. Justice, understood as consistent cooperation, provides more benefits than costs.

Socrates also shows us the distinction between the instrumental value and the intrinsic value of justice. He dispenses rather quickly with Thrasymachus' attempt to denounce the instrumental value of justice, but Socrates finds it far more difficult to overcome the claim of Glaucon and Adeimantus that justice lacks intrinsic value. As I demonstrate in later chapters, the instrumental benefits are sufficient to substantiate the argument in favor of justice. However, if it can be proven that there are intrinsic benefits to justice as well, that will provide an important bonus to those who choose to act accordingly. I conclude in the final chapter that justice comes up short in this regard.

Perhaps the most important contribution of *The Republic* is Socrates' suggestion that an understanding of justice and its benefits must be based upon an understanding of the underlying nature and psychology of human beings. This is an insight that has been embraced by subsequent philosophers of justice from Aristotle to Gauthier and beyond. Unfortunately for Socrates, this insight leads to his undoing; because his account of human psychology is deeply flawed, his account of justice is flawed as well.

In the early modern period, Hobbes and Hume advance many of the same themes of justice that were introduced by Socrates. They each offer an account of justice that is based upon

human nature and psychology, and they each recognize that justice is a non-zero-sum game. They also offer arguments in favor of justice as a rational strategy without having to depend on any notion of intrinsic value; for both Hobbes and Hume, the instrumental benefits of justice are sufficient.

The accounts of justice offered by Hobbes and Hume each have their own merits and drawbacks, but I have chosen to pursue Hobbes's line of argument over Hume's because it provides a much better framework for the project at hand. The goal here is to demonstrate that justice, understood as the consistent observance of cooperative agreements, is a rational strategy for a self-interested actor. Hume's account, while somewhat compelling, is dependent upon an individual's regard for the greater good. It is an argument leveled more on the societal level than on the individual level, and it is even altruistic in some respects. Hobbes, in contrast, makes his argument in purely self-interested terms. He is concerned with convincing us that to "perform (one's) covenants made"[263] is beneficial on an individual level, and his account is purely instrumental and more game-theoretic than Hume's. In addition, Hobbes's reply to the Fool provides us with the ideal starting point from which to advance the claim that the benefits of just behavior outweigh the costs.

Gauthier, Kavka and other contemporary contractarians have used Hobbes as the starting point for their own forays into the rationality of justice, and Gauthier's work, in particular, represents a quantum leap forward in the argument. Gauthier's goal in *Morals by Agreement* and his subsequent modifications of that work is similar to my own: He wants to show that justice, or "social morality," is a part of rational choice.[264] His work in *MbA* goes a long way towards

[263] *Leviathan*, ch. 15, p. 89
[264] Gauthier (2013), p. 624

achieving this goal, and it requires only a few modifications to reach what I believe to be the finish line.

By recognizing the distinction between equilibrium and optimality, Gauthier forms his most important insight, namely, that when an individual adopts an optimizing strategy, she is operating on the level of metachoice. That is, she is making a choice about how to make choices. Gauthier also introduces the key concept of deliberative procedures, and he makes three claims with regard to these. First, he claims that humans have the ability to willingly choose deliberative procedures in general. Second, assuming that the first claim holds, it is rational to choose a deliberative procedure in accordance with agreed Pareto-optimization (APO). Third, once we choose the deliberative procedure in accordance with APO, it is rational to comply with that procedure in all instances. With some minor assistance from other scholars, Gauthier presents a convincing argument in support of the first and third claims. However, the second claim falls short, and in order to make a convincing argument in favor of the rationality of justice, it is necessary to remedy this shortcoming.

The question that remains, then, is, "what kind of deliberative procedure is it rational to choose?" If it can be demonstrated that it is rational to choose a deliberative procedure that advocates consistent adherence to covenants, then I have achieved my goal: I have shown that justice is a rational strategy for a self-interested actor, despite its costs. If, however, it is rational to choose a procedure that recommends violating covenants when it is advantageous to do so, I have failed. Hobbes's Fool is in favor of opportunistic violation, and the rebuttal of his claim represents the final phase of my argument.

The Fool recognizes that justice is a non-zero-sum game, and he would agree that it is rational to adhere to covenants in most situations. However, he believes that when an individual

is presented with a situation in which the potential benefits of violation are high and the odds of getting caught are low, it is in the individual's best interest to violate the covenant. In other words, the Fool wants to permanently adopt and consistently adhere to a deliberative procedure that dictates that he violate covenants when it is to his advantage to do so. He claims that my goal cannot be met: The consistent observance of cooperative agreements is not always in one's own best interest.

The problem with the Fool's claim is that it is very difficult for him to objectively determine in advance which violations will turn out to be in his own self-interest. Behavioral economics has shown that overconfidence and ignorance of uncertainty are so pervasive and influential in our psychological constitution that we cannot rely on our own judgment to ascertain when an opportunity to violate has a positive expected return. The world is highly complex and unpredictable, and given a sufficiently large amount of trials, the Fool will eventually be discovered. Rather than heed the Fool's advice, an individual is better-served by a deliberative procedure that recommends consistent adherence to covenants because, as Kavka puts it, "Time wounds all heels."

Armed with these ideas from behavioral economics, we can now be more secure in the claim that it is rational to choose a deliberative procedure that advocates consistent adherence to covenants. With these new insights, I believe that my originally stated goal has been reached, yet there is still something missing. Our commonsense notions of justice lead us to believe that there are reasons for the consistent observance of cooperative agreements other than the fact that we probably cannot get away with acting otherwise; there seem to be benefits to justice other than the market-based benefits addressed in my imperfect reason argument.

While these non-simulator accessible (NSA) benefits do exist, we must be careful not to equate the existence of NSA benefits with proof that justice itself has intrinsic value. Justice does provide us with many things that have intrinsic value, such as happiness, a feeling of integrity, and a sense of participation and kinship with others, and justice may even be an integral part of what it means to be a normal human being. But this does not demonstrate that justice *itself* has intrinsic value.

If it were possible to vindicate Socrates by closing the circle and demonstrating that justice is valuable in itself, this would be a very tidy ending to this story. Unfortunately, we are going to have to settle for something less. Justice itself will not provide us with intrinsic value, but this is not a tragedy and it does not undermine the imperfect reason argument. Justice can provide everyone with some instrumental benefits, it can provide most of us with a bonus in the form of NSA instrumental benefits, and it can bring us very close to intrinsic value, even if it cannot get us all the way there.

Intentional Omissions

Hobbes defines justice as the keeping of covenants made, and this dissertation has addressed the topic of justice in this narrow contractarian sense. However, outside of the contractarian framework, the term "justice" often invokes connotations that are very different from the ones addressed here. It is likely that a reader of this dissertation will object, with justification, that important aspects of justice have been omitted. As I stated at the outset, I have omitted these aspects of the wider definition of justice not because they are unimportant, but because each of them could occupy a separate dissertation in its own right. Now that my claims regarding the contractarian sense of justice have been made in their entirety, I will very briefly

outline some important aspects of justice that have not been addressed, and consider how these might impact the central claims made above.

One instance of covenant and contract that has not yet been addressed is that in which one party to the contract is able to impose his will on the other party. Recall that one of the main assumptions of Hobbes's state of nature is that all individuals are approximately equal, in the sense that they all have approximately the same ability to kill one-another. Within this Hobbesean framework, when individuals enter into covenants, they do so for their own protection, of their own accord, and without being coerced by another party that is operating from a position of superior power. If this assumption of equality does not hold, and one of the parties to a covenant is in a position of superior power or coercion relative to the other, then my argument in favor of the consistent observance of covenants will no longer hold.[265] That is, if the more powerful party to a covenant can impose his will on the other party, both parties now have a reason to consider violation. The more powerful party may want to consider violating the covenant because he has little reason to fear punishment, reciprocity is no longer a motivation for compliance, and opportunistic violation is probably a rational strategy. The weaker party, on the other hand, may want to consider violating the covenant because if he had to be coerced, the covenant was probably never in his best interest in the first place. While the weaker party does still fear being caught, if he has been forced to enter a covenant that is worse than no covenant at all, he is incurring a cost with no corresponding benefit, and he should consider violating.

While situations of coercion do demonstrate that there are instances in which it is in one's rational self-interest to violate an agreement, these examples do not invalidate the claims I have made here because these examples are found outside the scope of justice as I have defined it.

[265]A critique of this kind has been made from a feminist perspective, via the claim that women are not equal parties to the original contract. See Pateman, Carole. *The Sexual Contract.* Stanford: Stanford University Press, 2008

Throughout this dissertation, I have defined justice as the consistent observance of cooperative agreements, and coercion is certainly not cooperation.[266] Granted, situations of coercion and unequal power certainly do exist, and the inability of my imperfect reason argument to deal with these situations demonstrates a limitation of the applicability of the argument. However, these situations are not examples of justice, either in common parlance or under my own definition, and they do not invalidate the arguments advanced here.

Another element of justice that has been conspicuously absent from the discussion so far is justice as "the first virtue of social institutions."[267] In this dissertation, I have emphasized the idea that justice, as I have defined it, involves choices. When we select a deliberative procedure in accordance with justice, we are making a choice about how to make choices. However, justice and injustice in the context of social structures often involves no choices at all. For example, I may argue that I am the beneficiary of an unjust system, as I am a white, well-educated man who happened to be born into a stable middle-class family in the wealthiest society in the history of humankind. I have done nothing to deserve these advantages, yet I benefit from them on a daily basis, sometimes (but not always) at the expense of others, without any act of choice on my part. For me to forsake all of these benefits in the name of justice would certainly be irrational if considered within the framework of the main arguments of this dissertation. If I were to choose "justice" in this Rawlsian sense, it would no longer be in my own self-interest.

It is true that the justice of social structures is a very commonsense notion of what justice is. It is of central importance in the work of Rawls, and it is directly addressed by Gauthier in the later chapters of *MbA*. However, as I have stated previously, the topic under discussion in this

[266] Gauthier recognizes this distinction also. See Gauthier (1986), pp. 191-192.
[267] Rawls, John. *A Theory of Justice*. Cambridge: Harvard University Press, 1971, p. 3

dissertation is not justice in the Rawlsian sense of social institutions. I am specifically concerned with justice understood as the consistent observance of cooperative agreements. I am not positing any veil of ignorance and I am not directly concerned with finding the optimal distribution of the benefits of cooperative agreements. While this is an appealing topic of great consequence which I hope can benefit from some of the insights that I have provided here, I will remain silent on the justice of social institutions, as it is beyond the scope of the current project.

The Impact of Modernity and Apathy

Hobbes's contractarian ideas were first proposed in 17th century Europe. Most of the people of this time lived in small villages and sparsely populated areas. Mobility was very limited; an individual was very likely to live her entire life within a few miles of her place of birth, and she would interact with the same small group of people repeatedly throughout her lifetime. In such circumstances, it is easy to see why Hobbes is confident in his reply to the Fool. In order to benefit from the adoption of a deliberative procedure that recommends opportunistic violation, an individual in these circumstances would have to deceive the same people over and over without being detected and punished, or at least without inducing them to alter their behavior towards him. In such close quarters, it is highly unlikely that this strategy will be successful. In addition, the costs of being exposed as a heel in a world like Hobbes's are quite high. If I have lived my entire life in a small community, and the friends and customs I have developed in that community are the only ones I know, to be ostracized from that community would be devastating. The wounds inflicted on a heel in such a situation would be quite deep.

Our 21st century world is very different from Hobbes's Europe in many obvious ways. We are now very mobile; individuals in developed countries have the ability to move from city to city and even from country to country with relative ease. Urban life has dramatically increased the density of populations and led to far more frequent contact among strangers. We now interact on a daily basis with other individuals whom we will likely never encounter again, and the internet increases this virtual anonymity exponentially. The claims I have made throughout this dissertation regarding the rationality of the consistent observance of cooperative agreements are dependent upon certain circumstances of the human condition, including the transparency of an individual's character (Sayre-McCord, 1991), the opportunity for reciprocity (Axelrod, 1981), and the availability of information (Skyrms, 1998). It is certainly worth asking whether modernity pushes us outside these circumstances.

Modernity has probably not changed the transparency of the average person when viewed in the context of repeated face-to-face interaction; we still have the ability to spot a cheat when information about the other party is literally right in our face. However, modernity has changed the nature of the interaction itself in that many of our interactions are neither repeated nor face-to-face. This almost certainly decreases transparency and increases the ability of a deceiver to conceal her deception. Axelrod recognizes that, "It is easy to maintain the norms of reciprocity in a stable small town or ethnic neighborhood,"[268] but the increased mobility of modern society clearly impacts reciprocity in that it can provide anonymity and decrease the likelihood of future interaction. Thus, modernity may provide a defector not only with the opportunity to conceal his intentions, but also with the ability to violate an agreement without having to fear retaliation. If he gets caught, the heel can just pick-up and move.

[268] Axelrod (1981), p. 312

While modernity has tilted the odds in favor of the heel in some ways, it has made life more difficult for him in other ways. The flow of information today is better than ever before. Urban life and impersonal commerce can certainly make anonymity easier to achieve, but with the recent ubiquity of the internet, the pendulum has swung the other direction with respect to anonymity. The internet makes reciprocation easier by giving nearly everyone the ability to damage a heel's reputation from the convenience of one's own laptop. The availability of information regarding the past behavior of counterparties is also more abundant now than ever before (consider the feedback functions on EBay and the ubiquitous credit report that we all carry with us into the virtual marketplace). Also, as Kavka suggests,[269] surveillance technology, polygraphs, and DNA testing all assist rule-enforcers, which increases the heel's likelihood of detection and possibly the ultimate cost of that detection as well.

It is unclear, therefore, whether the net impact of modernity is a benefit or a hindrance to the heel. In addition to the considerations listed above, the heel will also need to contemplate whether the benefits of consistent cooperation are higher now than they were in Hobbes time, and also whether the benefits of violation are higher. However, one thing that has not changed is the fact that the heel will need to make his assessment using imperfect machinery. His ability to reason is clouded by overconfidence and an inability to properly assess probability; this has not changed since Hobbes's era, and it continues to offer support for the rationality of just behavior.

Yet despite the inadequacy of the human ability to reason, there may still be some hope for the heels among us. Recall that the contractarian argument begins with fear. The individual in Hobbes's state of nature seeks peace because he fears the predations and reprisals of his fellows. If he has no such fear, he has no reason to seek peace. Similarly, my argument has

[269] Kavka (1995), pp. 29-30

assumed that a heel fears being detected because the downside of being detected is significant; a violator of a cooperative agreement will face a harsh punishment if she gets caught. However, if the heel does not fear retribution even in the event that she does get caught, she likely has no reason to behave in a just fashion.

This leads to a most troubling thought: *Apathy* may be the heel's best friend. It has been said that the only thing necessary for the triumph of evil is for good men to do nothing.[270] If but few people care about the appropriate punishment of transgressors, then a heel who places no value on the enjoyment of character or the esteem of others may be best-served by a policy of opportunistic violation of cooperative agreements.

It requires no great leap of the imagination to posit that some amount of this harmful apathy exists in modern western society. During the global financial crisis of 2008-2009, we witnessed Wall Street bankers and irresponsible homebuyers nearly destroy the world economy. The bankers were incompetent at best and criminally negligent at worst, yet their punishment will amount to little more than foregoing a portion of the money they improperly appropriated from their clients, shareholders and the public. The fact that any one of these banks has a single remaining client is a source of never-ending shock to an informed observer and it provides evidence of the presence of apathy regarding heels. Furthermore, the banks are not the only guilty party in this affair. Thousands of American homebuyers committed blatant mortgage fraud by overstating their financial health in order to obtain mortgages on homes they could not afford, yet there has been no effort made to prosecute these individuals and no public sense of outrage on the part of other individuals who spent responsibly or who lost their homes due to legitimate and unpredictable financial difficulty. Again, a sense of apathy is evident.

[270] Quote attributed to Edmund Burke

This apathy problem does potentially pose a challenge for my claim that justice is a rational strategy for a self-interested actor, and it would be an important and intriguing topic of additional research. However, it should be noted that the mere presence of apathy does not necessarily invalidate my central claim. The validity of my claim will depend, not upon the mere presence of apathy, but upon the pervasiveness and reliability of apathy in a given society. Apathy can be viewed as being analogous to the "errors of other men"[271] that Hobbes describes in his reply to the Fool. Like human error, the apathy of others is unreliable as a means of security, and the would-be heel is still well-advised to adopt a deliberative procedure of consistent adherence to his cooperative agreements rather than hope for assistance from the consistent and pervasive apathy of his peers.

Final Thoughts

Most people do not need to read an entire dissertation in order to be convinced of the benefits of just behavior. Our moral intuition and our recognition of NSA benefits such as a sense of connectivity with others and the enjoyment of good character are enough to convince us that we should behave in a just manner, and our fear of being caught in violation serves as a deterrent in the event of temptation. For those of us who are already convinced of the benefits of justice, the essential lesson here is that apathy is the enemy, and we need to be concerned about the observance of justice by others. We will benefit from justice only to the extent that we are not taken advantage of by free-riders and parasites. If some members of our community are permitted to openly ignore the precepts of justice without having to fear any serious ramifications, this significantly erodes the value of just behavior for the rest of us. We must

[271] *Leviathan*, ch. 15, p. 92

understand that not being taken advantage of is an ongoing act of constant vigilance, and a policy of forgive and forget will not serve us well.

However, this thesis was not intended to address those of us who are already convinced of the conclusions it draws. It was written to address the bright boy from Queens and others like him who fail to appreciate the enjoyment of character that accompanies adherence to cooperative agreements. Whether or not the behavior of these individuals can legitimately be described as "defective" is a subject of debate and is certainly a matter of degree. Regardless, the point is clear: Those individuals who believe that a deceptive strategy of opportunistic violation is in their best interest are only deceiving themselves. They are probably overconfident in their deceptive skills, their ability to assess the odds of success is deeply flawed, and if they have not yet been detected in their deceptions, this is likely due to chance rather than their own talent. Justice is the strategy of self-interest, and time is the enemy of all heels.

BIBLIOGRAPHY

Annas, Julia. “Comments on John Doris’ ‘Lack of Character’.” *Philosophy and Phenomenological Research* 71, no. 3 (Nov. 2005): 636-642.

Aristotle. “Nichomachean Ethics.” in *The Basic Works of Aristotle*, edited by Richard McKeon, 927-1112. New York: Random House, 1941.

Aristotle. “Politics.” in *The Basic Works of Aristotle*, edited by Richard McKeon, 1113-1316. New York: Random House, 1941.

Armstrong, Frank. *The Informed Investor*. New York: Amacom, 2002.

Ashraf, Nava, Colin F. Camerer, and George Loewenstein. “Adam Smith, Behavioral Economist.” *Journal of Economic Perspectives* 19, no. 3 (Summer 2005): 131-145.

Ashford, Elizabeth and Tim Mulgan. “Contractualism." *The Stanford Encyclopedia of Philosophy*. *edited by* Edward N. Zalta (Fall 2012 Edition),
URL = <http://plato.stanford.edu/archives/fall2012/entries/contractualism/>.

Axelrod, Robert. “The Emergence of Cooperation among Egoists.” *The American Political Science Review* 75, no. 2 (June, 1981): 306-318

Axelrod, Robert, and William Hamilton. “The Evolution of Cooperation.” *Science*, New Series 211, no. 4489 (March 27, 1981): 1390-1396

Axelrod, Robert and Douglas Dion. “The Further Evolution of Cooperation.” *Science,* New Series 242, no. 4884 (Dec. 9, 1988): 1385-1390.

Barney, Rachel. "Callicles and Thrasymachus." *The Stanford Encyclopedia of Philosophy*, *edited by* Edward N. Zalta (Winter 2011 Edition),
URL = <http://plato.stanford.edu/archives/win2011/entries/callicles-thrasymachus/>.

Barney, Rachel. “Socrates’ Refutation of Thrasymachus.” in *The Blackwell Guide to Plato’s Republic*, edited by Gerasimos Santos, 44-62. Hoboken: Wiley-Blackwell, 2008.

Boxill, Bernard. “How Injustice Pays.” *Philosophy and Public Affairs* 9, no. 4 (Summer 1980): 359-371.

Bratman, Michael. “The Interplay of Intention and Reason.” *Ethics* 123, no. 4 (July 2013): 657-672.

Brown, Eric. "Plato's Ethics and Politics in *The Republic*." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Winter 2011 Edition),
URL = <http://plato.stanford.edu/archives/win2011/entries/plato-ethics-politics/>.

Chappell, T.D.J. "The Virtues of Thrasymachus." *Phronesis* 38, no. 1 (1993): 1-17.

Cohon, Rachel. "Hume's Moral Philosophy." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Fall 2010 Edition),
URL = <http://plato.stanford.edu/archives/fall2010/entries/hume-moral/>.

Cooper, John M. "Plato's Theory of Human Motivation." *History of Philosophy Quarterly* 1, no. 1 (January, 1984): 3-21.

Cudd, Ann. "Contractarianism." *The Stanford Encyclopedia of Philosophy. edited by* Edward N. Zalta (Winter 2013 Edition),
URL = <http://plato.stanford.edu/archives/win2013/entries/contractarianism/>.

Cuneo, Terence. "Reid's Ethics." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Spring 2011 Edition),
URL = <http://plato.stanford.edu/archives/spr2011/entries/reid-ethics/>.

Danielson, Peter. "Closing the compliance dilemma: How it's rational to be moral in a Lamarckian world." in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 291-322. New York: Cambridge University Press, 1991.

Dawkins, Richard. *The Selfish Gene*. New York: Oxford University Press, 1976.

Doris, John. "Heated Agreement: Lack of Character as Being for the Good." *Philosophical Studies* 148 (2010): 135-146.

Doris, John. "Persons, Situations, and Virtue Ethics." *Nous* 32, no. 4 (Dec, 1998): 504-530.

Feinberg, Joel. "Psychological Egoism." in *Ethical Theory: An Anthology*, 2nd ed., edited by Russ Schafer-Landau, 167-177. Chichester: Wiley-Blackwell, 2012.

Finkelstein, Claire. "Pragmatic Rationality and Risk." *Ethics* 123, No. 4 (July, 2013): 673-699

Flew, Anthony. "Three Questions about Justice in Hume's Treatise." *The Philosophical Quarterly* 26, no. 102 (Jan. 1976): 1-13.

Foot, Philippa. "Moral Beliefs." *Proceedings of the Aristotelian Society* New Series 59 (1958-1959): 83-104.

Foot, Philippa. *Natural Goodness.* New York: Oxford University Press, 2001.

Foot, Philippa. “Nietzsche’s Immoralism.” in *Moral Dilemmas,* 144-158. Oxford: Clarendon Press, 2002.

Foot, Philippa. “Nietzsche: The Revaluation of Values.” in *Virtues and Vices,* 81-95. Oxford: Clarendon Press, 2002.

Frede, Dorothea. "Plato's Ethics: An Overview." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Fall 2013 Edition),
URL = <http://plato.stanford.edu/archives/fall2013/entries/plato-ethics/>.

Gauthier, David. “Assure and Threaten.” *Ethics* 104 (July 1994): 690-721.

Gauthier, David. “David Hume, Contractarian.” *The Philosophical Review* 89, no. 1 (Jan 1979): 3-38.

Gauthier, David. *The Logic of Leviathan.* Oxford: Clarendon Press, 1969.

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

Gauthier, David. “Rational constraint: Some last words,” in *Contractarianism and Rational Choice,* edited by Peter Vallentyne, 323-330. New York: Cambridge University Press, 1991.

Gauthier, David. “Rational Cooperation.” *Nous* 8, no. 1 (March 1974): 53-65.

Gauthier, David. “Thomas Hobbes: Moral Theorist.” *The Journal of Philosophy* 76, no. 10 (Oct. 1979): 547-559.

Gauthier, David. “Three against Justice: The Foole, the Sensible Knave, and the Lydian Shepherd.” in *Moral Dealing: Contract, Ethics and Reason,* 129-149. Ithaca: Cornell University Press, 1990.

Gauthier, David. “Twenty-Five On.” *Ethics* 123, No. 4 (July, 2013): 601-624

Gauthier, David. “Why Contractarianism?” in *Contractarianism and Rational Choice,* edited by Peter Vallentyne, 15-30. New York: Cambridge University Press, 1991.

Goldie, Dan and Gordon Murray. *The Investment Answer.* New York: Business Plus, 2011.

Hacker-Wright, John. “What Is Natural about Foot’s Ethical Naturalism?” *Ratio* 22 (Sept. 3, 2009): 309-321.

Haidt, Jonathan. “The Intuitive Dog and Its Rational Tail.” in *The Righteous Mind,* 27-51. New York: Pantheon, 2012.

Hamlin, J. Kiley, Karen Wynn, Paul Bloom, and Neha Mahajan. “How Infants and Toddlers React to Antisocial Others.” *Proceedings of the National Academy of Sciences* 108, no. 50 (Dec. 13, 2011): 19931-19936.

Hampton, Jean. “Equalizing concessions in the pursuit of justice: A discussion of Gauthier’s bargaining solution.” in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 149-161. New York: Cambridge University Press, 1991.

Hampton, Jean. “Hobbes and Ethical Naturalism.” *Philosophical Perspectives* 6 (1992): 333-353

Hampton, Jean. *Hobbes and the Social Contract Tradition*. New York: Cambridge University Press, 1986.

Hampton, Jean. “The Knavish Humean.” in *Rational Commitment and Social Justice*, edited by Jules L. Coleman and Christopher W. Morris, 150-167. Cambridge: Cambridge University Press, 1998.

Hampton, Jean. “Two faces of contractarian thought.” in *Contractarianism and Rational Choice*, edited by Peter Vallentyne, 31-55. New York: Cambridge University Press, 1991.

Harvey, Martin. “Hobbes and the Value of Justice.” *The Southern Journal of Philosophy* 42 (2004): 439-452.

Hayes, Peter. “Hobbes’s Silent Fool: A Response to Hoekstra.” *Political Theory* 27, no. 2 (April 1999): 225-229.

Hobbes, Thomas. *Human Nature*, edited by J.C.A. Gaskin. New York: Oxford University Press, 1994.

Hobbes, Thomas. *Leviathan*, edited by Edwin Curley. Indianapolis: Hackett Publishing Company, 1994.

Hoekstra, Kinch. “Hobbes and the Fool.” *Political Theory* 25, no. 5 (Oct. 1997): 620-654

Hoekstra, Kinch. “Hobbes on the Natural Condition of Mankind.” in *The Cambridge Companion to Hobbes’s Leviathan*, edited by Patricia Springborg, 109-127. New York: Cambridge University Press, 2007.

Hoekstra, Kinch. “Nothing to Declare?: Hobbes and the Advocate of Injustice.” *Political Theory* 27, no. 2 (April 1999): 230-235

Home, Henry, Lord Kames. *Essays on the Principles of Morality and Natural Religion*. Liberty Fund, The Online Library of Liberty. URL=<http://files.libertyfund.org/files/1352/Home_0995_EBk_v7.0.pdf>.

Hume, David. *An Enquiry Concerning the Principles of Morals*, edited by J.B. Schneewind. Indianapolis: Hackett Publishing Company, 1983.

Hume, David. *A Treatise of Human Nature*, edited by David F and Mary J Norton. New York: Oxford University Press, 2000.

Irwin, T.H. “Aristotle on Reason, Desire and Virtue.” *The Journal of Philosophy* 72, no. 17 (October 2, 1975): 567-578.

Jennings, Dennis L., Teresa M. Amabile, and Lee Ross. “Informational Covariation Assessment.” in *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 211-230. New York: Cambridge University Press, 1982.

Joseph, H.W.B. “Aristotle’s Definition of Moral Virtue, and Plato’s Account of Justice in the Soul.” *Philosophy* 9, no. 34 (April, 1934): 168 – 181.

Kahan, Dan M., Ellen Peters, Erica Dawson, and Paul Slovic. “Motivated Numeracy and Enlightened Self-Government.” Yale Law School, Public Law Working Paper no. 307 (September 3, 2013).
URL=< http://ssrn.com/abstract=2319992 or http://dx.doi.org/10.2139/ssrn.2319992>.

Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Strauss and Giroux, (2011).

Kavka, Gregory. “Hobbes’s War of All against All.” *Ethics* 93, no. 2 (Jan, 1983): 291-310.

Kavka, Gregory. “Morals by Agreement, by David Gauthier.” *Mind*, New Series 96, no. 381 (Jan, 1987): 117-121.

Kavka, Gregory. “The Rationality of Rule-Following: Hobbes’ Dispute with the Fool.” *Law and Philosophy* 14, no. 1 (Feb. 1995): 5-34.

Krause, Sharon. “Hume and the (False) Luster of Justice.” *Political Theory* 32, no. 5 (Oct. 2004): 628-655.

Kraut, Richard. "Aristotle's Ethics." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Winter 2012 Edition),
URL = <http://plato.stanford.edu/archives/win2012/entries/aristotle-ethics/>.

Kuhn, Steven. "Prisoner's Dilemma." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Spring 2009 Edition),
URL = <http://plato.stanford.edu/archives/spr2009/entries/prisoner-dilemma/>.

Lichtenstein, Sarah, Baruch Fischhoff and Lawrence D. Phillips. “Calibration of probabilities: The state of the art to 1980.” in *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic and Amos Tversky, 306-334. New York: Cambridge University Press, 1982.

Locke, John. *Two Treatises of Government*, edited by Peter Laslett. New York: Cambridge University Press, 1988.

London, Alex J. “Moral Knowledge and the Acquisition of Virtue in Aristotle’s ‘Nichomachean’ and ‘Eudemian Ethics’.” *The Review of Metaphysics* 54, no. 3 (March 2001): 553-583.

MacLean, Frederick and Tim Slattery. “The Collision of Pride and Memory.” *The Light Magazine*, September, 2010.

Martinich, A.P. *Hobbes*. New York: Routledge, 2005.

Moehler, Michael. “Why Hobbes’ State of Nature is Best Modeled by an Assurance Game.” *Utilitas* 21, no. 3 (Sept. 2009): 297-326.

Nietzsche, Friedrich. “Beyond Good and Evil.” in *Basic Writings of Nietzsche*, edited by Walter Kaufmann. New York: The Modern Library, 1992.

Nietzsche, Friedrich. “On the Genealogy of Morals.” in *Basic Writings of Nietzsche*, edited by Walter Kaufmann. New York: The Modern Library, 1992.

Nietzsche, Friedrich. “Twilight of the Idols.” in *The Portable Nietzsche,* edited by Walter Kaufmann. New York: Viking Penguin, 1976.

Pack, Spenser J. and Eric Schliesser. “Smith’s Humean Criticism of Hume’s Account of the Origin of Justice.” *Journal of the History of Philosophy* 44, no. 1 (2006): 47-63.

Pateman, Carole. *The Sexual Contract*. Stanford: Stanford University Press, 2008.

Plato. “The Republic.” in *Plato, Complete Works,* edited by John M. Cooper, translated by G.M.A. Grube and C.D.C. Reeve. Indianapolis: Hackett Publishing Company, 1997.

Postema, Gerald. “Hume’s Reply to the Sensible Knave.” *History of Philosophy Quarterly* 5, no. 1 (Jan., 1988): 23-40.

Postema, Gerald. “Whence Avidity? Hume’s Psychology and the Origins of Justice.” *Synthese* 152, no. 3 (Oct. 2006): 371-391.

Rainbolt, George. “Gauthier on Cooperating in Prisoner’s Dilemmas.” *Analysis* 49, no. 4 (Oct, 1989): 216-220.

Raphael, D.D. “Hume and Smith on Justice and Utility.” *Proceedings of the Aristotelian Society* New Series 73 (1972-1973): 87-103.

Rawls, John. *Lectures on the History of Moral Philosophy.* Cambridge: Harvard University Press, 2000.

Rawls, John. *A Theory of Justice.* Cambridge: Harvard University Press, 1971.

Reeve, C.D.C. “Glaucon’s Challenge and Thrasymacheanism.” *Oxford Studies in Ancient Philosophy* 34 (May 29, 2008): 69-103.

Reid, Thomas. *Essays on the Active Powers of Man, Essay V, Chapter V*. Early Modern Texts. URL<http://www.earlymoderntexts.com/rea5.html>.

Rescorla, Michael. "Convention." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Spring 2011 Edition),
URL = http://plato.stanford.edu/archives/spr2011/entries/convention/.

Sayre-McCord, Geoffrey. “Deception and reasons to be moral.” in *Contractarianism and Rational Choice,* edited by Peter Vallentyne, 181-195. New York: Cambridge University Press, 1991.

Schneewind, J.B. “The Misfortunes of Virtue.” *Ethics* 101, no. 1 (Oct. 1990): 42-63.

Shefrin, Hersh. *Beyond Greed and Fear.* New York: Oxford University Press, 2002.
Shields, Christopher. “Plato’s Challenge: the Case against Justice in Republic II.” in *The Blackwell Guide to Plato’s Republic*, edited by Gerasimos Santos, 63-81. Hoboken: Wiley-Blackwell, 2008.

Singpurwalla, Rachel. “Plato’s Defense of Justice in *The Republic*.” in *The Blackwell Guide to Plato’s Republic*, edited by Gerasimos Santos, 263-279. Hoboken: Wiley-Blackwell, 2008.

Skyrms, Brian. “The Shadow of the Future.” in *Rational Commitment and Social Justice,* edited by Jules Coleman and Christopher Morris, 12-21. New York: Cambridge University Press, 1998.

Smith, Adam. *The Theory of Moral Sentiments,* edited by D.D Raphael and A.L. Macfie. Oxford: Oxford University Press, 1976.

Smith, Adam. *The Wealth of Nations,* edited by Edwin Cannan. New York: Modern Library, 2000.

Sorell, Tom. “Hobbes’s Moral Philosophy.” in *The Cambridge Companion to Hobbes’s Leviathan*, edited by Patricia Springborg, 128-153. New York: Cambridge University Press, 2007.
Taleb, Nassim Nicholas. *The Black Swan.* New York: Random House, 2007.

Taleb, Nassim Nicholas. *Fooled by Randomness.* New York: Texere, 2004.

Thompson, Michael. “Three Degrees of Natural Goodness (Discussion Note, *Iride*).” URL=http://www.pitt.edu/~mthompso/three.pdf (retrieved August 14, 2012).

Trivers, Robert. “The Evolution of Reciprocal Altruism.” *The Quarterly Review of Biology* 46, no. 1 (March 1971): 35-57.

Vallentyne, Peter. “Contractarianism and the assumption of mutual unconcern.” in *Contractarianism and Rational Choice,* edited by Peter Vallentyne, 71-75. New York: Cambridge University Press, 1991.

Vanderschraff, Peter. “Game Theory, Evolution, and Justice.” *Philosophy and Public Affairs* 28, no. 4 (Autumn 1999): 325-358.

Vanderschraff, Peter. “The Invisible Foole.” *Philosophical Studies* 147 (2010): 37-58.

Verbeek, Bruno and Christopher Morris. "Game Theory and Ethics." *The Stanford Encyclopedia of Philosophy, edited by* Edward N. Zalta (Summer 2010 Edition),
URL = <http://plato.stanford.edu/archives/sum2010/entries/game-ethics/>.

Woozley, A.D. “Hume on Justice.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 33, no. 1 (Jan 1978): 81-99.

Yi, Byeong-Uk. “Rationality and the Prisoner’s Dilemma in David Gauthier’s Morals by Agreement.” *The Journal of Philosophy* 89, no. 9 (Sept. 1992): 484-495.