

# UNLEASHING THE LEVIATHAN

: Against Common Interpretations and a Contemporary  
Decision-Theoretic Reconstruction of Hobbes's Moral and  
Political Philosophy

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Hun Chung

January 2012

**© 2012 Hun Chung**

**ALL RIGHTS RESERVED**

# ABSTRACT

## UNLEASHING THE LEVIATHAN

: Against Common Interpretations and a Contemporary Decision-Theoretic  
Reconstruction of Hobbes's Moral and Political Philosophy

Hun Chung, Ph.D.

Cornell University 2012

My dissertation consists of two parts. Part I “Unleashing the Leviathan” attempts to free Hobbes’s moral and political philosophy from three commonly held interpretations. In chapter 1, I free Hobbes from the preference-satisfaction theory of the good. The preference-satisfaction theory of the good claims that what is good for each individual is simply to satisfy his or her current preferences or desires. I show that this is not the case for Hobbes, since the entire system of Hobbes entirely rests on the assumption that self-preservation is objectively each and every individual’s greatest good. In chapter 2, I free Hobbes from the Humean conception of instrumental rationality. This purely instrumental conception of rationality assumes that no preferences or desires can properly be said to be irrational in themselves, and that the role of reason or rationality can only be confined to: (a) informing the agent with true beliefs about the world, and (b) revealing the most effective means that could satisfy the current ends (whatever they are) that the agent happens to have. I show that this is not the case for Hobbes, since, for Hobbes, self-preservation is the very aim of

rationality. In chapter 3, I free Hobbes from psychological egoism. Psychological egoism is a theory of human psychology that claims that all human beings are motivated solely by their own self-interest. I argue that not only was Hobbes not committed to psychological egoism in any of its plausible formulations, I also argue that, unlike what people commonly think, psychological egoism is not even needed for Hobbes's political philosophy.

In Part II "Reinvigorating the Leviathan", I provide a contemporary reconstruction of Hobbes's moral and political philosophy in the lights of formal decision theory and modern game theory. In chapter 4, I reinterpret Hobbes's theory of the good as a version of what is, now, known as an ideal-advisor theory of the good. An ideal-advisor theory of the good maintains that what is really good for a given individual is to satisfy the type of preferences that his/her fully-rational self would form on behalf of his/her actual self. In chapter 5, I provide a formal representation of Hobbes's theory of the good in terms of contemporary utility theory. In Chapter 6, I analyze Hobbes's state of nature. There, I argue against conventional attempts to understand Hobbes's state of nature as a game of Prisoner's Dilemma. As an alternative, I provide three game-theoretic models that utilize the tools of modern Bayesian game-theory. Not only do these three Bayesian game-theoretic models respect what is actually written in Hobbes's original text, but they also show how universal conflict can inevitably emerge as a stable equilibrium due to uncertainty, without assuming psychological egoism.

## **BIOGRAPHICAL SKETCH**

Hun Chung was born in Seoul, South Korea. He earned his B.A. in Philosophy (*Summa Cum Laude*) from Seoul National University in 2006. After studying one semester in the Masters Program in Philosophy at Seoul National University, he came to the Sage School of Philosophy at Cornell University to pursue his doctoral studies in the fall of 2007. After finishing all requirements for Ph.D. candidacy, Hun Chung earned his M.A. in Philosophy in May, 2011, and earned his Ph.D. in Philosophy in January, 2012 from the Sage School of Philosophy at Cornell University.

# **DEDICATION**

I dedicate this dissertation to my parents:  
Changyin Chung and Yon Suk Lee.

## ACKNOWLEDGMENTS

My interest in economics and game theory was piqued after reading David Gauthier's *Morals by Agreement*. Gauthier's *Morals by Agreement* gave me a sense of how formal theory could be applied to moral and political philosophy in an interesting way; it showed how economic models could help understand and provide new insights to traditional questions of moral and political philosophy.

Gauthier expressed that his inspiration came from Thomas Hobbes. I soon discovered that there were other philosophers, such as Jean Hampton and Gregory Kavka – generally known as the *Hobbesian Contractarianists* – who took a very similar approach to that of Gauthier. I wished to know the Hobbesian contractarianists well.

Naturally, my research interests as a graduate student started to ramify into three different directions; (a) reading the works of contemporary Hobbesian contractarianists, (b) reading the works of Thomas Hobbes himself, and (c) studying formal and economic theory.

However, after studying these three areas simultaneously for some time, I started to feel that there were some problems with the Hobbesian contractarianists; their understanding of Thomas Hobbes as well as their application of formal game theory were not entirely accurate. Correcting these mistakes became one of my central motivations to write a dissertation on Thomas Hobbes.

The main title of my dissertation, “Unleashing the Leviathan”, was designed to metaphorically capture such intention. The word “*Leviathan*” refers to a biblical sea monster; but it was refers to the title of Thomas Hobbes's most significant work in moral and political philosophy. The title, “Unleashing the Leviathan”, is supposed to represent my attempts to free Hobbes from a number of interpretations that are commonly attributed to his philosophical views (especially by the Hobbesian contractarians.)

Writing and completing my dissertation was a long and arduous journey. During the process, I have benefited tremendously by many people's advice and guidance.

My dissertation supervisor was Professor Nicholas Sturgeon. Professor Sturgeon was one of the strictest supervisors that I have ever met in my academic career. He was extremely careful in the details, and every time I received comments on my draft from him, it was obvious that he had read my draft sentence-by-sentence multiple times. Not only did he teach me how to read historical texts in philosophy very carefully, he always provided extremely important criticisms which I have overlooked previously. This made me reorganize the entire structure of my dissertation multiple times.

Professor Sturgeon was not entirely antagonistic towards applying formal theory to the interpretation of Hobbes. However, he was generally skeptical about the fruitfulness of extensively applying formal theory to philosophy; he thought that such formal apparatus should be kept to a bare minimum. There were many occasions in which Professor Sturgeon thought that I was using too much formal theory.

Professor Richard Miller, who is the second member of my dissertation committee, also expressed similar worries, but, even more forcefully than Professor Sturgeon. It might have been Professor Miller's real opinion or he might have simply been playing the devil's advocate, but, it seemed that Professor Miller thought that there was virtually nothing that can be gained by applying formal theory to philosophy; he seemed to think that everything that could be gained in philosophy could be gained by doing philosophy alone.

These worries of Professor Sturgeon and Professor Miller led me to rethink about my basic approach at the fundamental level: What philosophical insights could be achieved by applying formal theory to philosophy? Why even bother to apply game theory to the interpretation of Hobbes? To respond to these skeptical worries, I had to constantly justify and explain why my basic approach was philosophically meaningful. And, as a result, I believe that I have become more intellectually mature and sophisticated than what I would have been otherwise. By rethinking and responding to the criticisms presented by Professor

Sturgeon and Professor Miller, I was able to avoid both traps of being either too slavish or too dismissive towards formal approaches to philosophy.

Professor Harold Hodes, the third member of my dissertation committee, gave me challenges from a different angle. As a mathematical logician, Professor Hodes was already quite familiar with taking a formal approach to philosophy and did not raise any fundamental questions concerning its philosophical significance. Rather, Professor Hodes concentrated on ensuring that the formal apparatus that I used in my dissertation was clear, precise, and met the standard of rigorousness required for a doctoral thesis. Professor Hodes helped me clarify and correct many notational errors and raised many insightful criticisms which led me to clarify the formal parts as best as I can. Professor Hodes has also put in tremendous effort to make my English writing sound less Korean.

Professor Derk Pereboom, who is the fourth member of my dissertation committee, also provided tremendous help by deliberately playing the devil's advocate. Professor Pereboom was a student of Jean Hampton – one of the main Hobbesian contractarianist which this dissertation criticizes – when he was graduate student at UCLA. During many conversations, Professor Pereboom tried to defend his past teacher as best as he could against my criticisms. Such conversation led me to rethink in what way my approach to Hobbes is distinctive than that of the Hobbesian contractarianists.

I would like to thank all of my dissertation committee members that I have mentioned above. All of their encouragements, criticisms, and guidance made my graduate years at the Sage School of Philosophy at Cornell University intellectually fulfilling. If there are any remaining faults or errors that are left in this final draft, I would like to emphasize that the fault is entirely on my part.

I would also like to express my special thanks to Professor John Roemer at Yale University, Professor Michael Sandel at Harvard University, and Professor Kyung-Sig Hwang at Seoul National University.

Professor Roemer was not a member of my dissertation committee. However, after a

couple of email correspondences, he asked me to send a couple of chapters of my dissertation. We were able to meet soon after, and we discussed about my work as well as my research plans. Professor Roemer encouraged me to keep doing research in the intersection of economics and philosophy.

I met Professor Michael Sandel in 2005 when I was still an undergraduate at Seoul National University. Professor Sandel visited South Korea to give a series of lectures in political philosophy. We kept in touch ever since. In the later stages of my graduate years, I was able to give a presentation of some parts of my dissertation in his presence, and he gave me very helpful suggestions and criticisms in return.

Professor Kyung-Sig Hwang is a professor ethics in Seoul National University. He was the very first person who led me to study ethics and political philosophy.

I would like to thank all of these people who provided support outside my graduate program at Cornell.

Last but not least, I extend my special thanks to my wife, Sunghwa Chung, for her endless support and criticisms.

This dissertation represents, not a final product, but an ongoing research in the intersection of economics and philosophy.

Hun Chung

Ithaca, New York

January 2, 2012

# Contents

<b>I</b>	<b>UNLEASHING THE LEVIATHAN</b>	<b>7</b>
<b>1</b>	<b>Freeing Hobbes From the Preference-Satisfaction Theory of the Good</b>	<b>11</b>
1.1	The Preference-Satisfaction Theory of the Good . . . . .	12
1.2	The Key Text . . . . .	15
1.3	Our Strategy: Two Ways to Show that Hobbes was not Committed to the Preference-Satisfaction Theory of the Good . . . . .	16
1.3.1	First Strategy: <i>Self-Preservation as Objectively One's Greatest Good</i>	17
1.3.2	Second Strategy: <i>It is Bad to Satisfy the Fool's Preferences</i> . . . . .	21
1.4	Interpretation of the Key Text: The Distinction between Real Good and Apparent Good . . . . .	25
1.5	Taking Care of One Last Worry . . . . .	29
1.6	Did Hobbes Hold an Idealized Preference-Satisfaction Theory of the Good?	31
1.6.1	The Motivation . . . . .	31
1.6.2	Grounds to Think that Hobbes had held an Idealized Preference-Satisfaction Theory of the Good . . . . .	33
1.6.3	Hobbes's Conception of Practical Rationality and the Reason Why Hobbes did <i>Not</i> Hold an Idealized Preference-Satisfaction Theory of the Good. . . . .	35
<b>2</b>	<b>Freeing Hobbes from the Humean Conception of Instrumental Rationality</b>	<b>41</b>

2.1	David Hume’s Conception of Instrumental Rationality . . . . .	41
2.2	Was Hobbes Committed to the Humean Conception of Instrumental Rationality? . . . . .	45
2.2.1	Hobbes Did Not Endorse Clause (b) . . . . .	46
2.2.2	Hobbes Did Not Endorse Clause (a) . . . . .	48
2.3	Reconciliation with Hobbes’s General Project . . . . .	52
<b>3</b>	<b>Freeing Hobbes From Psychological Egoism</b>	<b>56</b>
3.1	The Motivation Behind Attributing Psychological Egoism to Hobbes . . . . .	56
3.2	What is Psychological Egoism? - Some Clarifications . . . . .	58
3.3	Was Hobbes a Psychological Egoist? . . . . .	61
3.3.1	Was Hobbes a Psychological Hedonist? . . . . .	61
3.3.2	<i>Hobbes’s Dictum and Tautological Egoism</i> . . . . .	63
3.3.3	What Psychological Egoism is for Hobbes and Whether He Endorsed it . . . . .	65
3.3.4	Was Hobbes a <i>Subjective Egoist</i> ? . . . . .	70
3.4	Psychological Egoism is Not Needed for Hobbes’s Political Philosophy . . . . .	78
<b>II</b>	<b>Reinvigorating The Leviathan</b>	<b>83</b>
<b>4</b>	<b>Reconstructing Hobbes’s Theory of the Good as an Ideal-Advisor Theory of the Good</b>	<b>86</b>
4.1	Systematizing Hobbes’s Theory of Personal Good . . . . .	87
4.2	Reconstructing Hobbes’s Theory of Personal Good as an Ideal-Advisor Theory	93
<b>5</b>	<b>A Contemporary Decision Theoretic Reconstruction of Hobbes’s Theory of the Good</b>	<b>103</b>
5.1	Measurement Theory and the Assignment of Numbers . . . . .	104

5.1.1	<i>A General Introduction</i> . . . . .	104
5.1.2	<i>Why Bother with Utility Theory in Interpreting Hobbes?</i> . . . . .	111
5.2	An Ordinal Representation of Hobbes’s Theory of Real Good . . . . .	113
5.2.1	<i>An Ordinal Representation for <math>Bob_i</math></i> . . . . .	113
5.2.2	<i>Clarifying the Meaning of Utility Functions</i> . . . . .	126
5.3	An Expected Utility Representation of Hobbes’s Theory of Real Good . . .	131
5.3.1	<i>Clarifying the Meaning of Expected Utility Theory</i> . . . . .	132
5.3.2	<i>An Expected Utility Representation (Von-Neumann and Morgenstern’s Framework)</i> . . . . .	136
5.3.3	<i>A Note on One’s Attitudes Towards Risk</i> . . . . .	157
5.3.4	<i>An Expected Utility Representation (Savage’s Framework)</i> . . . . .	163
5.4	An Additive Utility Representation of Hobbes’s Theory of Real Good . . .	189
5.5	The Role of Laws of Nature: Solving the Epistemic Problem . . . . .	208
<b>6</b>	<b>A Bayesian Game-Theoretic Reconstruction of Hobbes’s State of Nature</b>	<b>212</b>
6.1	Hobbes’s State of Nature: State of War . . . . .	212
6.2	The Five Axioms of The State of Nature (which leads to a state of universal war) . . . . .	214
6.3	Hobbes’s State of Nature as a “Prisoner’s Dilemma (PD)” Game . . . . .	221
6.3.1	The Four Desiderata of Hobbes’s State of Nature . . . . .	221
6.3.2	The PD (Prisoner’s Dilemma) Game . . . . .	222
6.3.3	Some Common Misunderstandings of the PD Game . . . . .	225
6.4	Why Hobbes’s State of Nature is <i>Not</i> a PD Game . . . . .	232
6.5	Other Alternative Models: <i>The Stag Hunt</i> and <i>The Iterated PD Game</i> . . . .	236
6.5.1	Problems with Modeling Hobbes’s State of Nature as a Game of <i>Stag Hunt</i> . . . . .	237

6.5.2	Problems with Modeling Hobbes's State of Nature as an <i>Iterated PD Game</i> . . . . .	238
6.6	Modeling Hobbes's State of Nature with Bayesian Game Theory . . . . .	241
6.6.1	First Bayesian Model of Hobbes's State of Nature: Uncertain About the Other Person's Type . . . . .	243
6.6.2	Second Bayesian Model of Hobbes's State of Nature: Is it Possible for One to Credibly Signal One's Type? . . . . .	254
6.6.3	Third Bayesian Model of Hobbes's State of Nature: Uncertain About the Other Person's Beliefs . . . . .	258
6.7	Concluding Remarks . . . . .	261
	<b>Bibliography</b>	<b>268</b>

# List of Figures

5.1	How Function $U$ Represents One's Preference-Ordering on the Real Line . . . . .	120
5.2	Function $V$ as a Strictly Increasing Transformation of Function $U$ . . . . .	121
5.3	Reduction of Compound Lotteries . . . . .	138
5.4	Cardinal Preferences of Each Type of Bob . . . . .	154
5.5	Risk-Averse . . . . .	161
5.6	Risk-Loving . . . . .	162
5.7	Risk-Neutral . . . . .	162
6.1	The First Bayesian Game-Theoretic Model of Hobbes's State of Nature . . . . .	250
6.2	The Second Bayesian Game-Theoretic Model of Hobbes's State of Nature . . . . .	254
6.3	The Third Bayesian Game-Theoretic Model of Hobbes's State of Nature . . . . .	258

# List of Tables

5.1	Some Common Scale Types . . . . .	109
5.2	Allais Paradox . . . . .	140
5.3	The Arms Race . . . . .	166
5.4	The Goodness of Chains of Consequences . . . . .	190
6.1	The PD Game . . . . .	223
6.2	The Story of the Prisoner's Dilemma . . . . .	223
6.3	The Stag Hunt . . . . .	237
6.4	Different Types of Bob and Jill's Preferences in the State of Nature . . . . .	244
6.5	The Consequences that Each Type of Bob and Jill will Face by Performing Their Respective Actions . . . . .	245
6.6	Summary of the Four Games Played by Each Type of Bob and Jill . . . . .	249

## **Part I**

# **UNLEASHING THE LEVIATHAN**

# General Introduction for Part I

This dissertation consists in two large parts. In Part I, “Unleashing the Leviathan”, I attempt to free Hobbes’ moral and political philosophy from a number of commonly held interpretations for his views. In Part II, “Reinvigorating the Leviathan”, I attempt to reconstruct Hobbes’s theory of the good as well as his description of the state of nature in the lights of contemporary decision theory and game theory. Part I consists in three chapters, which are all designed to free Hobbes from a certain kind of interpretation that is commonly given to his views (usually by people who are generally known as the Hobbesian contractarianists.<sup>1</sup>)

My arguments as well as my interpretive reconstruction of Hobbes’ theory of human psychology and his theory of good will be based primarily on textual evidence that can be found from *Leviathan*<sup>2</sup>, *De Homine*.<sup>3</sup>, and *On the Citizen*<sup>4</sup>, which are all regarded as the representative works of Hobbes’s moral and political philosophy. There are basically three commonly held interpretations that are attributed to Hobbes’s views.

The first common understanding of Hobbes is that Hobbes has proposed the simplest form of what is known as “the preference-satisfaction theory of good”. The preference-satisfaction theory of good claims that what is good for each individual at any given time is simply to satisfy his or her current preferences or desires. The preference-satisfaction theory

---

<sup>1</sup>See Gauthier 1969, 1984, Hampton 1986, Kavka 1986

<sup>2</sup>Thomas Hobbes, *Leviathan* (with selected variants from the Latin edition of 1668), (edited, with introduction, by Edwin Curley), Hackett (1994). All citations from *Leviathan* are made from this text.

<sup>3</sup>Thomas Hobbes, *Man and Citizen (De Homine and De Cive)*, (edited by Bernard Gert), Hackett (1991). All citations from *De Homine* are made from this text.

<sup>4</sup>Thomas Hobbes, *On the Citizen* (edited by Richard Tuck and Michael Silverthorne), Cambridge (1997). All citations from *On the Citizen* are made from this text.

of good is quite problematic, especially because it seems that people occasionally do prefer things that are not actually good for them. For example, I might prefer to play video games rather than to study for an important exam that is scheduled the very next day. However, it seems hardly true that playing video games would really be good for me or would be in my best interest simply because I had preferred so. However, many people commonly think that Hobbes was committed to such view in its most simplistic form.<sup>5</sup>

The second common understanding of Hobbes is that Hobbes was a precursor and an advocate of what is now known as “the Humean theory of instrumental rationality”. The Humean theory of instrumental rationality is a view that is often attributed to David Hume on the basis of what he wrote in *A Treatise of Human Nature*. The view claims that no preferences or desires can properly be said to be irrational in themselves, and that the role of reason or rationality can only be confined to: (a) informing the agent with true beliefs about the world, and (b) revealing the most effective means that could satisfy the current ends (whatever they are) that the agent happens to have. It is controversial whether Hume actually held this view. However, it is very common to think that *Hobbes* was committed to such view of rationality and reason.<sup>6</sup>

The third common understanding of Hobbes is that Hobbes was a “psychological egoist”.<sup>7</sup> Psychological egoism is a theory of human psychology that claims that all human

---

<sup>5</sup>See [Gauthier, 1984, Hampton, 1986, Railton, 1986b] I will quote the specific passages where these authors attribute the preference-satisfaction theory of good to Hobbes later on. There are certain versions of the preference-satisfaction theory of good that claims that what is good for an individual is to satisfy his or her *idealized preferences*. I have put “in its most simplistic form” to indicate that I am not referring to such an idealized preference-satisfaction theory of the good.

<sup>6</sup>See [Gauthier, 1969, Hampton, 1986] Again, I will quote the specific passages where these authors attribute the Humean conception of instrumental rationality to Hobbes later on.

<sup>7</sup>See [Butler, 1983, Hume, 1975, Broad, 1950]. [Kavka, 1986] thinks that some textual evidence does suggest that Hobbes was a psychological egoist, but thinks that only a weakened version of psychological egoism, which he calls “Predominant Egoism” is needed for Hobbes’s political philosophy to work. [McNeilly, 1966] thinks that Hobbes was at least committed to psychological egoism in his earlier works. [Hampton, 1986, pp. 20-24] interprets Hobbes as a psychological egoist who maintains that all of our desires are *caused by a “self-interested” bodily mechanism*, and opposes the idea of interpreting Hobbes as a psychological egoist who claims that all of our desires have *self-regarding content*. In other words, according to Hampton, Hobbes does allow people to have certain kinds of other-regarding desires. However, according to Hampton, these other-regarding desires play absolutely no role in Hobbes’s political argument that it is not entirely unreasonable to regard Hobbes as a psychological egoist when one is trying to understand his political philosophy. Gert [1967,

actions are ultimately motivated by self-interest or self-love alone. Not only do many people think Hobbes was a psychological egoist, but they also think that psychological egoism is essential to his political philosophy. That is, many people think that Hobbes cannot properly explain why it is the case, according to his view, that the state of nature would dissolve into a state of war of all against all without relying on psychological egoism. In other words, psychological egoism, according to the conventional interpretation, is absolutely necessary for Hobbes's entire political philosophy to work.

The main purpose of Part I is to argue that none of these three interpretations is, strictly speaking, true (or at least that there are reasonable interpretations of Hobbes that free him from such accusations). The first chapter is intended to free Hobbes from the preference-satisfaction theory of the good. The second chapter is intended to free Hobbes from the Humean conception of instrumental rationality. The third chapter is intended to free Hobbes from psychological egoism.

---

1991, "Introduction" in *Man and Citizen*]denies that Hobbes was a psychological egoist and claims that he can, at best, seen as merely, what he calls, a "tautological egoist."

# Chapter 1

## Freeing Hobbes From the Preference-Satisfaction Theory of the Good

It has been commonly thought that Hobbes had held the preference-satisfaction theory of the good. The preference-satisfaction theory of the good maintains that what is *good* for an individual at any given time is to satisfy his or her current preferences whatever they happen to be. Coupled with this interpretation is the interpretation that Hobbes was committed to a purely instrumental conception of practical rationality that foreshadows that of David Hume. One can quite easily see that the two doctrines make a very natural combination. If one thinks that the satisfaction of just any kind of preferences is good for the individual, then it seems pretty natural to think that the only role that reason and rationality can play is to inform the individual with the best available means to satisfy the type of preferences that he/she happens to have at the current moment; reason and rationality do not tell what the individual *should* rightly prefer, and are only, as Hume puts it, “slave of the passions.” The whole purpose of this chapter and the next chapter is to show that, despite the common trend to interpret Hobbes in these two ways, Hobbes was committed to neither of these doctrines.

## 1.1 The Preference-Satisfaction Theory of the Good

As I have explained, the preference-satisfaction theory of the good maintains that what is good for a given individual at a given time is the satisfaction of the individual's current desires and preferences whatever they are. The preference-satisfaction theory of the good is also (perhaps better) known in philosophy as "the desire-satisfaction theory of the good" or "subjectivist theory of value". Economists generally favor the term "preference" while it seems that philosophers tend to use the term "desire".

The major difference between the two terms is that, unlike "desire", "preference" has a comparative notion built-into its very meaning. At first, such a difference might not seem to be that significant, but it actually has a very important practical implication. For instance, suppose that there are three objects, x, y, and z that one desires; here, suppose that one prefers x to y to z. If so, then the desire-satisfaction theory of the good will claim that it would be good for one to obtain any of the three objects. However, the desire-satisfaction theory of the good would be silent on the issue of which object would be *better* for one to obtain if one wasn't able to obtain all three. By contrast, the preference-satisfaction theory of the good would claim that it would be best for one to obtain x if one had a choice among x, y and z, and that it would be better for one to obtain y if one had to choose between y and z. In short, it might be argued that preference is a more useful concept than desire in the sense that it informs us with a *ranking* of the various objects under consideration. I will use the name "preference-satisfaction theory of the good" to denote the stance that is usually known as "desire-satisfaction theory of the good" to philosophers.<sup>1</sup>

As I have mentioned, the preference-satisfaction theory of the good claims that it is generally better for people to get what they prefer. Note that it does not say anything about what people *should* prefer. If John prefers chocolate ice cream to vanilla ice cream and Jane prefers vanilla ice cream to chocolate ice cream, then the preference-satisfaction theory of

---

<sup>1</sup>For a further discussion concerning the distinction between the two terms, "desire" and "preference", see, Broome, "Introduction: ethics out of economics" in [Broome, 1999, section 1.5]

good claims that it is *better for John* to have chocolate ice cream and that it is *better for Jane* to have vanilla ice cream.

Here, we can see that the preference-satisfaction theory of the good is committed to a certain form of *relativism*; that is, it claims that anything that is good (or better) is *good (or better) for a specific individual*, not something that is *just simply good (or better) in itself*. Then, the theory supplements this relativistic notion of the good by providing an account of how the relative good of a given individual is determined; according to the theory, an individual's relative goodness is determined by the individual's current preferences. In other words, according to the preference-satisfaction theory of the good, a person's preferences work something very similar to a magic wand; the very fact that one prefers something automatically *makes* that thing good for the person who prefers it.

So, we can say the preference-satisfaction theory of good is a conjunction of the following two claims:

[PREFERENCE-SATISFACTION THEORY OF THE GOOD]:

1. (Relativism): what is good is *good for a specific individual*, not what is *simply good* in itself.
2. (Subjectivism): the good of a given individual at any given time is *determined by his* or her *current preferences (or desires)*.

Most commentators think that Hobbes had adopted the preference-satisfaction theory of the good in this particular formulation. Consider how Peter Railton characterizes Hobbes's theory of good:

Perhaps the simplest relational theory of goodness is that of Hobbes, who held that to call something good is always to speak of someone's good, and that the

only sense in which something can be good for someone is that he desires it.  
[Railton 1986b, contained in (2003), p.49]

Similarly, consider how David Gauthier characterizes Hobbes' theory of good:

Perhaps the classic philosophic formulation of a conception of value both subjective and relative was offered in the seventeenth century by Thomas Hobbes. ... Hobbes links subjectivism with relativism – the view that value is dependent on appetite or preference with the view that value is relative to each individual.  
[Gauthier 1984, p. 51]

Jean Hampton also writes:

...Hobbes is clearly defining 'good' as "what we desire", and 'bad' as "what we are averse to." ... What is good is simply what we desire, and what we hate simply what is bad. That this is a *baldly subjectivist ethical understanding of 'good'* is something Hobbes seems to not only admit but also welcome. ... Hobbes is saying that, strictly speaking, when we use the word 'good' we must use it relative to an individual or set of individuals at a particular time. [Hampton 1986, p. 29 emphasis added]

However common it is to attribute the preference-satisfaction theory of good to Hobbes, attributing the preference-satisfaction theory of good to Hobbes has two major problems.

First of all, the preference-satisfaction theory of good is independently a very implausible theory. People very often prefer to do things that are not actually good for them; a drug addict has a very strong preference to take drugs, but it would be quite absurd to claim that giving drugs to this drug addict would *really be good* for him/her.<sup>2</sup> Therefore, the fact that Hobbes

---

<sup>2</sup>From this, it has also been argued that the preference-satisfaction theory of good cannot be a basis for social goodness or social welfare which is the prime objective that welfare economics aims to achieve. See Broome, "Introduction: ethics out of economics" contained in [Broome, 1999, pp. 3-8], where Broome gives a short but a very convincing argument on why the preference-satisfaction theory of the good is false.

was committed to the preference-satisfaction theory of the good, which is independently implausible, undermines the plausibility of Hobbes's moral philosophy.

Second, regardless of whether or not the preference-satisfaction theory of the good is independently plausible, there are numerous pieces of textual evidence that suggest that Hobbes did not actually adopt the preference-satisfaction theory of the good, as I will argue. The main purpose of this chapter is to show that Hobbes did not actually adopt the preference-satisfaction theory of the good as we normally understand it.

## 1.2 The Key Text

Writers usually attribute the preference-satisfaction theory of the good to Hobbes on the basis of the following passage:

But whatsoever is the object of any man's appetite or desire that is it which he for his part *calleth good*; and the object of his hate and aversion, *evil*; and of his contempt, *vile* and *inconsiderable*. For these words of good, evil, and contemptible are ever used with relation to the person that useth them, there being nothing simply and absolutely so, nor any common rule of good and evil to be taken from the nature of the objects themselves, but from the person of the man . . . [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 7] <sup>3</sup>

The passage is generally interpreted as supporting both *relativism* and *subjectivism* which we have seen in the previous section, which are the two major components of the preference-satisfaction theory of the good.

---

<sup>3</sup>There is also another passage that people often cite to support their attribution of the preference-satisfaction theory of good to Hobbes.

*Continual success* in obtaining those things which a man from time to time desireth, that is to say, continual prospering, is that men call FELICITY; [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 58]

Here, we can see that Hobbes is identifying the continuous satisfaction of a person's own desires with the person's continual prospering - which is another way to say that the person is achieving his or her own good - the result of which is normally considered to be felicity (or happiness).

I do not intend to question whether or not Hobbes had held a relativistic view concerning people's good. I do think that the above passage supports such interpretation.<sup>4</sup>This means that Hobbes' theory of the good meets at least one part of the two constitutive claims of the preference-satisfaction theory of the good. So, in order to show that Hobbes was not committed to the preference-satisfaction theory of the good, we would have to show that Hobbes's theory of the good was *not subjectivist* (albeit being relativistic) as defined in section 1. In other words, we would have to show that Hobbes did not think that the satisfaction of just any kind of preferences or desires is really good for a given individual. Can this be done?

Right now, the odds are against us. In the above passage, Hobbes claims that whatever is the object of a person's desire is, *for his or her part, called good*. Here, Hobbes seems to be claiming that it is *better* for one to get what one prefers regardless of what that happens to be. That is, Hobbes seems to be saying that what is good for a given individual is determined by the individual's current preferences. This is subjectivism. And if so, it makes Hobbes committed to the preference-satisfaction theory of the good.

### **1.3 Our Strategy: Two Ways to Show that Hobbes was not Committed to the Preference-Satisfaction Theory of the Good**

Then, how are we to show that Hobbes was not committed to subjectivism, and was, thereby, not committed to the preference-satisfaction theory of the good? There are two basic ways

---

<sup>4</sup>For, in the passage, Hobbes states that there cannot be anything that is simply and absolutely good in itself, and that whatever is good is good *in relation to a particular person and the person's particular circumstances*. Such relativistic view is quite prevalent in Hobbes' works. Consider,

... therefore one cannot speak of something as being *simply good*; since whatsoever is good, is good for someone or other. ... Therefore good is said to be relative to person, place, and time.  
[Hobbes 1991: *De Homine*: Chapter XI, Section 3]

So, I think that it is safe to say that the theory of good that Hobbes adopts is at least relativistic.

to do this:

1. Show that there is something that Hobbes thought to be *objectively good* for a given individual regardless of whether or not he/she prefers or desires it.
2. Show that there are certain things that Hobbes thought to be *objectively bad* for a given individual even when he/she prefers or desires it.

### **1.3.1 First Strategy: *Self-Preservation as Objectively One's Greatest Good***

We start out with the first strategy: to show that there is something that Hobbes thought to be objectively good for a given individual independent of the individual's actual preferences or desires. If one reads Hobbes carefully, it is not hard to discover that the major assumption that permeates Hobbes's entire moral and political philosophy is that *self-preservation* is each and every individual's *greatest good*.

Moreover, *the greatest of goods* for each is his own *preservation*. For nature is so arranged that all desire good for themselves. Insofar as it is within their capacities, it is *necessary* to desire *life, health*, and further, insofar as it can be done, *security* of future time. [Hobbes 1991: *De Homine*, Chapter XI, Section 6, emphasis added]

Here, we can see what Hobbes deems to be the three major components that constitute one's self-preservation: it is one's life, health, and security. Hobbes claims that whenever it is within one's capacities, it is necessary to desire one's self-preservation.

The passage clearly shows that Hobbes thought that self-preservation is each and every individual's greatest good. What the passage does *not* show is that Hobbes thought that self-preservation is *objectively* each and every individual's greatest good. That is, it might be the case that the reason why Hobbes thought that self-preservation is every human being's greatest good is because every human being simply desires his/her self-preservation more

than anything else. If so, the fact that Hobbes thought that self-preservation is each and every individual's greatest good does not show that Hobbes was not committed to subjectivism. It might have very well been true that Hobbes thought that what *makes* self-preservation the greatest good for everybody is the fact that everybody desires it.

However, I claim that this is not the case for Hobbes. That is, I claim that, for Hobbes, self-preservation is *objectively* each and every individual's greatest good. In order to show that, for Hobbes, the greatest goodness of self-preservation is objectively, and not subjectively, determined, it is sufficient to show that Hobbes did not think that everybody, *as a matter of fact*, prefers securing his/her self-preservation more than anything else, and that Hobbes thought that self-preservation is the greatest good for *even these types of people who do not desire their self-preservation very strongly*.

First of all, it is quite clear that Hobbes thought that there exist people who do not prefer securing their self-preservation more than anything else. The most noticeable example is what Hobbes refers to as *vain-glorious men*.

Also, because there be some that taking pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their security requires, [Hobbes 1994: *Leviathan*, Chapter XIII, Section 4]

Vain-glorious men ... are inclined to rash engaging... [Hobbes 1994: *Leviathan*, Chapter XI, Paragraph 12]

For Hobbes, *glory* is defined as the pleasure that one feels by self-recognizing that one has superior power over other people.<sup>5</sup> *Vain-glorious* men are of a type of people who are obsessed with glory, which makes them pursue power and superiority over others much more than what their self-preservation requires. Simply put, vain-glorious people are of a type who desire glory much more than their self-preservation.

One should note that Hobbes did not think that *everybody* is vain-glorious in this way.

---

<sup>5</sup>“Joy arising from imagination of a man's own power and ability is that exultation of the mind which is called GLORYING...” [Hobbes 1994: *Leviathan*, Chapter VI, Section 39]

In the state of nature there is in all men a will to do harm, but not for the same reason or with equal culpability. One man practices the equality of nature (...) this is the mark of *modest man* (...) Another, supposing himself superior to other, wants to be allowed everything (...) that is the sign of *an aggressive character*. In his case, the will to do harm derives from *vainglory*. [Hobbes 1997: *On the Citizen*, Chapter 1, Section 4 emphasis added]

In other words, according to Hobbes, only *some* people are vain-glorious; that is, only some people prefer glory much more than their self-preservation. However, as we have just seen above, Hobbes claims that self-preservation is the greatest good for *everybody*. This means that, for Hobbes, self-preservation is the greatest good for even the vain-glorious people, who do not desire their self-preservation more than anything else. This suggests that, for Hobbes, the greatest goodness of self-preservation does not really depend on its being desired or preferred by anybody.

If Hobbes was truly a subjectivist, and was, thereby, committed to the preference-satisfaction theory of the good, he would have claimed that glory is the greatest good for the vain-glorious people, who desire, more than anything else, their glory. It can be quite easily shown that Hobbes did not think this way. The fact that, for Hobbes, self-preservation is *objectively* each and every individual's greatest good can be further confirmed by Hobbes's attitude towards those who do not desire their own self-preservation strongly enough.

The passion whose violence or continuance maketh *madness* is either great vain-glory, which is commonly called *pride* and *self-conceit*, or great dejection of mind. [Hobbes 1994: *Leviathan*, Chapter VIII, Paraphraph 18, emphasis on "madness" is mine]

Here, Hobbes lists what he thinks as the two major passions that cause people to act in ways that are detrimental to their own self-preservation; according to Hobbes, the two major passions are either *vain-glory* or *great dejection*. Please note what Hobbes says about the

people who are being influenced by any of these two passions; Hobbes thinks that people who are being influenced either by vain-glory or great dejection are suffering *madness*.

I believe that this is a point that many scholars of Hobbes have either completely ignored or have not taken very seriously. According to Hobbes,

... all passions that produce strange and unusual behavior are called by the general name of madness. [Hobbes 1994: *Leviathan*: Chapter VIII, Paragraph 20]

Hobbes is saying here that anybody who displays *strange and unusual behaviors* indicate that he/she is suffering madness. In order to understand what Hobbes meant by “strange and unusual behaviors”, we should first understand what Hobbes considers to be *normal behavior*. As we have seen above, Hobbes thinks that it is necessary<sup>6</sup> for human beings to desire their own self-preservation more than anything else. So, for Hobbes, normal behavior consists in behaving in ways that are consistent with the achievement of one’s long-term self-preservation. This means that anybody who is behaving in ways that show that he/she is not giving utmost priority to securing his/her long-term self-preservation, and is actually acting in ways that are inconsistent with its very achievement, is, for Hobbes, a person who is displaying strange and unusual behavior, and, is, therefore, a person who is suffering madness.

In other words, for Hobbes, even if somebody happens to value glory more than his/her self-preservation, this fact does not automatically make the achievement of glory *really better* for him/her than the achievement of his/her long-term self-preservation. Rather, the fact that the person values glory more than his/her self-preservation only signifies that the person is irrational or mad<sup>7</sup>: that he/she is desiring *the wrong thing*.

---

<sup>6</sup>Again, “Moreover, *the greatest of goods* for each is his own *preservation*. For nature is so arranged that all desire good for themselves. Insofar as it is within their capacities, it is *necessary* to desire *life, health*, and further, insofar as it can be done, *security* of future time.” [Hobbes 1991: *De Homine*, Chapter XI, Section 6, emphasis added]

<sup>7</sup>Somebody might point out that there is a difference between being irrational and being mad; a person might behave in irrational ways, now and then, without being clinically diagnosed to be mad. However, for Hobbes, there is no significant difference between being in the two states.

This shows that, for Hobbes, the fact that self-preservation is the greatest good for everybody does not really depend on its being desired or preferred more than anything else by everybody. Within Hobbes's moral system, self-preservation is already assumed to be *objectively* the greatest good for each and every individual. This completely defies the main spirit of subjectivism - which claims that a person's good is completely determined by what the person actually desires or prefers. In short, Hobbes was not committed to the preference-satisfaction theory of the good.

### **1.3.2 Second Strategy: *It is Bad to Satisfy the Fool's Preferences***

In the previous section, I have shown that Hobbes was not committed to the preference-satisfaction theory of the good by showing that there is something the value of which Hobbes did not think to depend on its being desired or preferred by anyone; namely, self-preservation, which Hobbes deems to be objectively the greatest good for each and every individual. In this section, I will show that Hobbes was not committed to the preference-satisfaction theory of the good by showing that Hobbes did not think that the sheer fact of desiring or preferring something automatically makes that something good for the individual.

Actually, I have already shown two examples of this in the previous section. The vain-glorious person and the person who is subjected to great dejection - the two types of people that Hobbes thinks to be suffering madness. The former prefers glory to his/her self-preservation, and the latter prefers self-destruction to self-preservation. The fact that Hobbes thought that the preferences of these two types of people are mad clearly indicates that Hobbes did not think that something is good (or better) for a given individual simply because he/she desired (or preferred) it.

However, there is another famous example that shows that Hobbes did not think that the sheer fact of desiring or preferring something automatically makes that something good for the individual. The example is what is famously known as *Hobbes's fool*.

In *Leviathan*, the fool is an imaginary opponent of Hobbes who questions the rationality of performing one's own part of a mutual agreement, when the other party has already performed his/her part.

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." He does not therein deny that there be covenants, and that they are sometimes broken, sometimes kept, and that such breach of them may be called injustice; but he questioneth whether injustice, taking away the fear of God (for the same fool hath said in his heart there is no God), may not sometimes stand with that of reason which dictateth to every man his own good. ... you may call it injustice, ..., yet it can never be against reason... [Hobbes 1994: *Leviathan*, Chapter XV, Section 4]

What the fool is arguing in this passage is roughly this: If one expects to gain at the other's expense by not performing one's own part of a covenant, and if one can also expect that one can do this without ever being punished for performing such actions, then such actions cannot be properly called irrational (even if we may conventionally call such action to be unjust). According to Hobbes, "This specious<sup>8</sup> reasoning is nevertheless false. [Hobbes 1994: *Leviathan*, Chapter XV, Section 4]".

Hobbes provides two main reasons why he thinks preferring to perform such actions, which ostensibly seem to be in one's own best self-interest, is, nonetheless, not really good

---

<sup>8</sup>The term "specious" has changed its meaning in the last several centuries. Although Edwin Curley thinks that the term "specious" is used pejoratively in this context, such pejorative connotation was not always associated with the term in its historical usage. (For example, the term is not used pejoratively by David Hume. I thank Nick Sturgeon for pointing this out for me.) In the *glossary* of Hobbes [1994], the term "specious" is defined as "plausible, apparently sounding convincing, but in reality sophistical or fallacious; fair, attractive but lacking genuineness."

for oneself:

first, that when a man doth a thing which, notwithstanding anything can be foreseen and reckoned on, tendeth to his own destruction (howsoever some accident which he could not expect, arriving, may turn it to his benefit), yet such events do not make it reasonably or wisely done. Secondly, ..., 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 can cannot reasonably reckon upon as the means of his security; and therefore, if he be left or cast out of society, he perisheth; and if he live in society, it is by the errors of other men, which he could not foresee nor reckon upon; and consequently [he has acted] against the reason of his preservation... [Hobbes 1994: *Leviathan*, Chapter XV, Section 5]

The wording is slightly convoluted. But, here is the main point. According to Hobbes, anyone would realize that not doing one's part of a mutual agreement, when one already knows that the other party has performed his or her own part, is irrational, if he or she properly takes into account *the long-term effects* that one's actions would bring. This is because one cannot reasonably expect that one would be able to conceal one's actions forever; sooner or later other people will find out that one has cheated, and by knowing that one has cheated in the past, other rational people would be quite unlikely to cooperate with one or let one participate in some cooperative venture in the future. Since one cannot properly continue to live without engaging in at least some cooperative activities with other people, being excluded from most cooperative activities in this way would very likely undermine one's own self-preservation, which, according to Hobbes, is objectively the greatest good of all mankind.

Of course, there might be some exceptional cases where one could completely get away

with one's cheating-behaviors, and there might also be some exceptional cases where other people, who already know that one is a cheater, would, nonetheless, let one participate in their cooperative activities. However, according to Hobbes, such exceptional cases cannot be a proper basis to ground one's decisions for action.<sup>9</sup>

Anyway, if one were to decide to cheat based on such exceptional cases, then this means that one thinks that such exceptional cases provide reasons for one to assume that people in general cannot detect others' cheating-behaviors in a reliable way, or that, even if people generally are reliable cheater-detectors, they are usually stupid or irrational enough to give additional chances to pre-identified past-cheaters and would prefer to bear themselves the risk of getting cheated once again. This is totally unreasonable.

Therefore, if one had decided to cheat rather than to keep one's part of the covenant, then one's decisions were *irrational*, even if, by some accidental fluke, one were to get away with one's cheating-behaviors without punishment. In short, what Hobbes is claiming in the passage above is that it is actually *bad* for the person (in the long-run) to cheat and take advantage of other people's initial cooperation even when the person strongly prefers to do so. This is the reason why Hobbes calls such person a "fool."<sup>10</sup>

In this section, we have seen that Hobbes thought that there are certain things that are objectively bad for a given individual even if the individual strongly prefers or desires it. This again completely defies the main spirit of subjectivism - which claims that anything that is the object of one's current preferences or desires are good for one. Again, Hobbes was not committed to the preference-satisfaction theory of the good.

---

<sup>9</sup>We can say that, here, Hobbes is assuming that nobody has the Gyges' ring introduced in Plato's *Republic* Book II (See Cooper [1997, p. 1000]), and is being very optimistic about people's ability to detect as well as their willingness to punish cheaters. I thank Nick Sturgeon for pointing this out to me.

<sup>10</sup>According to Curley, the position that Hobbes ascribes to the "foole" is very close to the one Grotius ascribes to Carneades in *De jure belli ac pacis* (*On the Law of War and Peace*.) [Hobbes 1994: *Leviathan*, p. 90, footnotes 2 and 3]

## 1.4 Interpretation of the Key Text: The Distinction between Real Good and Apparent Good

In the previous sections, we have seen two reasons to think that Hobbes did not hold the preference-satisfaction theory of the good. For one thing, there is something that Hobbes thought to be objectively the greatest good for each and every individual - namely, his/her self-preservation - regardless of whether or not the person desires or prefers his/her self-preservation more than anything else. Secondly, there are some things that Hobbes thought to be objectively bad for a given individual even when the individual strongly desires or prefers it; namely, satisfying the preferences of the fool as well as the preferences of the people who are overwhelmed by a passion for glory or dejection.

What all this shows is that Hobbes did not hold the preference-satisfaction theory of the good as it is commonly perceived. But, then, how does all of this fit with the key text that have made so many people attribute the preference-satisfaction theory of the good to Hobbes. Let's go back to key text that we have seen in section 2.

But whatsoever is the object of any man's appetite or desire that is it which he for his part *calleth good*; and the object of his hate and aversion, *evil*; and of his contempt, *vile* and *inconsiderable*. For these words of good, evil, and contemptible are ever used with relation to the person that useth them, there being nothing simply and absolutely so, nor any common rule of good and evil to be taken from the nature of the objects themselves, but from the person of the man . . . [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 7]

In the passage, Hobbes claims that whatever is the object of a person's desires or preferences is, for his/her part, *called good*. The passage can be interpreted in any of the following three ways:

*Interpretation 1.* We might think of Hobbes as *defining* what goodness *is*. Such interpre-

tation implies that Hobbes was committed to the preference-satisfaction theory of the good. It is evident that people who think that Hobbes held the preference-satisfaction theory of the good adopts this interpretation. However, doing so contradicts with what we have established in the previous sections; namely, that Hobbes did not hold the preference-satisfaction theory of the good. So, if one adopts this interpretation, one has no choice but to claim that Hobbes was inconsistent in his views about value. This may be true. But, I believe that saying that some philosopher's view is inconsistent should be left as a last resort after one has successfully eliminated all other alternate interpretations.

*Interpretation 2.* We might think that what Hobbes is advancing in the above passage is *not a normative ethical theory* that tries to provide a substantial theory of what goodness (for a particular person) consists in, but rather a *meta-ethical theory* that intends to explain what is going on when we are using ethical terms and, thereby, *call* something “good” or “evil”. According to this interpretation, what Hobbes is saying in the above passage is this: when somebody calls an object *X* good, what he/she is doing is simply *expressing his/her positive attitudes* (namely, his/her *desires*) towards *X*, and is *not asserting anything*. This would imply that Hobbes was an *expressivist* or a *non-cognitivist* about value. There are some people who have persuasively advanced such interpretation.<sup>11</sup> However, such an interpretation implies that Hobbes thought that sentences such as “*X* is good for *Y*” lacks any definite truth-value. This conflicts with many parts of Hobbes's original text which jointly suggest that, for Hobbes, the sentence, “self-preservation is the greatest good for individual *Y*”, is not simply truth-valueless, but, *objectively true*.

*Interpretation 3.* Lastly, we might interpret the passage this way. When somebody desires or prefers something and, thereby, *calls* that thing “good”, this indicates that the object *seems* or *appears* to be good to him/her. However, this leaves open the possibility that the person might be mistaken about the goodness of the object.

---

<sup>11</sup> See Darwall [2000]. Gregory Kavka also calls such interpretation “the straightforward interpretation” and considers it as a possibility, but later discards it in favor of what he calls “the subtle interpretation”. See Kavka [1986, pp. 292-297]

For instance, suppose that somebody took some kind of herbal medicine whenever he had a flu believing that the herbal medicine would invigorate his body and give him strength to overcome the disease. However, suppose that medical researchers have later confirmed that the herbal medicine is strongly carcinogenic. In this situation, we can say that when the person *called* the herbal medicine *good* and desired to take it whenever he caught a flu, the herbal medicine *appeared good* to him; but, as a matter of fact, the herbal medicine was *not really good* for him.

This kind of interpretation invokes a distinction between what is *really good* and what is merely *apparently good* for a given individual. And it is not hard to verify that the distinction between real good and apparent good is a distinction that Hobbes explicitly makes in several places in his works.

... good (like evil) is divided into *real* and *apparent*. [Hobbes 1991: *De Homine*, Chapter XI, Section 5]

Whence it happens that inexperienced men that do not look closely enough at the long-term consequences of things, accept what appears to be good, not seeing the evil annexed to it; afterwards they experience damage. And this is what is meant by those who distinguish good and evil as *real* and *apparent*. [Hobbes 1991: *De Homine*, Chapter XI, Section 5]

By using Hobbes's own distinction between real good and apparent good, we might interpret the key passage, which many scholars have hitherto interpreted as implying Hobbes had held the preference-satisfaction theory of the good, as defining a person's *apparent good* as opposed to a person's *real good*.

That is, according to this interpretation, what Hobbes is saying in the key passage is that whatever is the object of a person's desires or preferences is the person's *apparent good*; however, whether that object is *really good* for the person is still left as an open question.

If obtaining the object is consistent with securing the person's long-term self-preservation,

which Hobbes deems to be objectively the greatest good for everybody, then the satisfaction of such desire would be really good for the person. However, if the obtaining of the object puts the person's long-term self-preservation at significant risk, then the satisfaction of such desire would only be apparently, but not really, good for the person.

This third interpretation has several advantages over the other two interpretations. First of all, none of the other two interpretations (i.e. interpretation 1 and interpretation 2) can make sense of Hobbes's explicit distinction between real good and apparent good.

Interpretation 1 commits Hobbes to the preference-satisfaction theory of the good, which implies that Hobbes thought that it is really good to satisfy just any kind of desire or preferences that a person happens to have. If this is so, it is unclear why Hobbes had explicitly distinguished between what is really good and what is merely apparently good for a given individual, and claimed that there are occasions where the two do not coincide.

Interpretation 2 makes Hobbes an expressivist or a non-cognitivist for value-terms such as "good." If interpretation 2 is true, then Hobbes would have to think that such sentences as, "X is apparently good for individual Y, but X is not really good for individual Y.", lacks any objective truth-value. However, for Hobbes, such sentence does have an objective truth-value; if Y happens to prefer X, but if obtaining X is inconsistent with achieving Y's long-term self-preservation, then the sentence would be objectively true; if Y happens to prefer X, and if obtaining X is consistent with achieving Y's long-term self-preservation, then the sentence would be objectively false. Interpretation 2 would not be able to accommodate such objective truth-conditions of statements concerning an individual's good. On the other hand, interpretation 3, which interprets the key text as defining Hobbes's notion of apparent good, encounters none of these problems.

Second, interpretation 3 shows us a way to make Hobbes's views about value and goodness consistent. We have seen in the previous section that there is much textual evidence that jointly suggest that Hobbes did not hold the preference-satisfaction theory of the good as he is commonly assumed to hold. By interpreting the key passage as Hobbes's attempt to define

“apparent good” (not “real good”), we are able to make the key passage consistent with these other parts of Hobbes’s original text which conflict with the view that Hobbes had held the preference-satisfaction theory of the good.

More specifically, according to interpretation 3, what Hobbes was saying in the key passage was simply that whatever is the object of one’s current desires or preferences is *apparently* good. This means that whether such object is *really* good for the individual is yet undetermined. In order to see whether such object is really good for the individual, we would need to see how the object relates to the individual’s long-term self-preservation.

So, if a person happens to prefer glory even at the very expense of his/her long-term self-preservation, satisfying his/her preferences would, for Hobbes, still be good (and thereby called “good”) albeit apparently. However, this does not mean that satisfying the person’s preferences would be really good for him/her, and the fact that the person did not prefer his/her self-preservation strongly enough does not undermine the fact that his/her long-term self-preservation is objectively the greatest good for him/her. In short, another advantage of interpretation 3 is that we do not have to accuse Hobbes of holding inconsistent doctrines at the same time.

## 1.5 Taking Care of One Last Worry

Before I conclusively claim that Hobbes did not hold the preference-satisfaction theory of the good, I would like to take care of one last qualm that is left. Let’s go back to the last part of the key passage:

For these words of good, evil, and contemptible are ever used with relation to the person that useth them, there being nothing simply and absolutely so, nor any common rule of good and evil to be taken from the nature of the objects themselves, but from the person of the man... [*Hobbes 1994: Leviathan*, Chapter VI, Paragraph 7]

One might think that Hobbes is, here, advocating the preference-satisfaction theory of good by concentrating on the part where Hobbes claims that “there is no common rule of good and evil (..) but from the person of the man.” Hobbes can be seen here as relying on the individual’s current preferences to determine what is good for him or her.

But note that Hobbes is not merely claiming that “there is no common rule of good and evil” but that “there is no common rule of good and evil *to be taken from the nature of the objects themselves.*” In other words, what Hobbes is denying here is merely that the objective properties of external objects, *taken by themselves*, can constitute what is good for somebody without their being related to that person in some relevant way. However, the relevant relation in question need *not* be *the satisfaction of one’s current preferences.*

Here is an example. Suppose that I have diabetes and you don’t. Then a specific food that would be objectively good for your health might actually be very bad for mine. We can see here that the specific food is not good or bad *in itself*: rather, the food is objectively good *for you* and objectively bad *for me* in relation to our health. This is *relativism.*

And as I have already explained, relativism does not imply the preference satisfaction theory of the good. This is because relativism does not imply subjectivism. In other words, the fact that a specific food is good for you and bad for me could have nothing to do with satisfying our current preferences whatsoever; I, who have diabetes, might strongly prefer to eat that specific food which would be objectively bad for me, and you, who do not have diabetes, might strongly prefer to not eat that specific food even when it is quite obvious that it would be objectively good for you. However, the fact that we respectively have these types of preferences does not change the fact that the specific food is objectively bad for me while it is objectively good for you.

This is what I believe is behind Hobbes’s distinction between real good and apparent good. That is, there might be things that are really good for somebody despite the fact that they are apparently bad (i.e. the person does not prefer such things), and there might be things that are really bad for somebody despite the fact that they are apparently good

(i.e. the person does prefer such things.) What the last part of the key passage is saying is that what kind of things are really good for somebody might differ from person to person (i.e. relativism.) However, this does not mean that whether something is good or bad for a given person is determined by the person's current preferences. I claim that Hobbes did not advocate a preference-satisfaction theory of good.

## **1.6 Did Hobbes Hold an Idealized Preference-Satisfaction Theory of the Good?**

I believe that many readers would now have been persuaded that Hobbes did not hold the preference-satisfaction theory of the good as it was introduced in section 1. However, I expect that many people would still be quite reluctant to accept that Hobbes had thought that one's long-term self-preservation is *objectively* one's greatest good. This is because the fact that Hobbes had thought that there is something that is objectively each and every individual's greatest good seems to completely defy the main spirit of Hobbes's entire project.

### **1.6.1 The Motivation**

As it is well-known, one of the primary aims of Hobbes was to present an alternative system of moral and political philosophy that was completely purged of any vestiges of Aristotelean metaphysics - which assumes there to be objectively the highest good for all human beings which everybody should rightly pursue as their final end. Instead, Hobbes intended to present a system of ethics and political philosophy where all normative conclusions can be properly derived and reduced to facts about human psychology and physical motions. The basic intent was to align moral and political philosophy with the then burgeoning natural sciences in Hobbes's own time.

So, it is not a surprise that many people would be shocked by my claim that the primary

assumption of Hobbes's moral and political philosophy is that self-preservation is *objectively* each and everybody's greatest good; this seems to make Hobbes's moral and political philosophy very close to that of Aristotle's.

Of course, people will not want to deny that self-preservation is each and every individual's greatest good in Hobbes's moral and political philosophy as securing one's self-preservation is the central notion that permeates through out all of Hobbes's works. What they would want to deny is that, for Hobbes, the greatest goodness of self-preservation is *objectively determined*. That is, if one were to preserve Hobbes's general intention to provide a reductive analysis of normative concepts that was purely based on facts about human psychology and natural physics, it would be nice if we could find a way to say that, for Hobbes, the greatest goodness of self-preservation is somehow determined by what everybody, as a matter of fact, desires or prefers.

This would require two things: (i) showing that Hobbes was committed to the preference-satisfaction theory of the good, and (ii) showing that Hobbes thought that everybody as a matter of fact desires or prefers his/her self-preservation more than anything else. However, we have seen that Hobbes clearly thought that there are people who do not desire or prefer their long-term self-preservation as much as they rightly should (e.g. the madman and the fool); Hobbes clearly thought that satisfying the preferences of these people would not be really good for them. In short, I have shown that Hobbes did not endorse any of these two requirements.

To this, somebody who wants to retain the general spirit of Hobbes's reductive analysis might claim that it is, strictly speaking, not the satisfaction of people's *actual* or *current* preferences which Hobbes thought to be really good for people, but rather, what Hobbes thought to be really good was the satisfaction of people's *rational* or *idealized* preferences - that is, the type of preferences which people would form if they were perfectly informed about the relevant facts and were free from any psychological distractions. And it might also be argued that Hobbes thought that everybody would prefer to secure their own long-

term self-preservation more than anything else when people are perfectly informed about the relevant fact and are free from any psychological distractions in this way.

In short, one might try to argue that Hobbes was committed to an *idealized preference-satisfaction theory of the good*, and that the reason why Hobbes thought that self-preservation is each and every individual's greatest good is because he thought that everybody *would* prefer their own self-preservation more than anything else if they were *ideally rational*.

Of course, this would make Hobbes's theory of the good undoubtedly normative; so it does not completely preserve Hobbes's intention to reduce normative concepts to factual ones. However, it still retains a certain form of *subjectivism* by saying that whatever is good for a given person is good *by the very fact that the person would desire or prefer it in certain idealized circumstances*. Such interpretation, on the surface, seems to be in better accordance with Hobbes's general spirit of providing a reductive analysis for value terms than simply assuming that self-preservation is objectively the greatest good for everybody.

### **1.6.2 Grounds to Think that Hobbes had held an Idealized Preference-Satisfaction Theory of the Good**

It should be noted that there are actually textual grounds to interpret Hobbes as proposing an idealized preference-satisfaction theory of the good. Consider:

And because in deliberation the appetites and aversions are raised by foresight of the good and evil consequences and sequels of the action whereof we deliberate, the good or evil effect thereof dependeth on the foresight of a long chain of consequences, of which very seldom any man is able to see to the end. . . . so that he who hath by *experience* or *reason* the greatest and surest prospect of consequences *deliberates best* himself, and is able, when he will, to give the best counsel unto others. [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 57 emphasis added]

Here, we can see that Hobbes is saying that people who have the greatest and surest prospects of consequences are generally the best deliberators. And, we can see that Hobbes thinks that one acquires such information concerning the nature as well as the likelihoods of various consequences that occurs by performing a given course of action by one's past experiences and reasoning powers. If we combine this with Hobbes's distinction between real good and apparent good that we have seen previously, we can see that there is room for us to interpret Hobbes as saying that what is truly (or really) good for a given individual is what that individual *would* desire or prefer if he/she were fully informed about the relevant facts, which may be acquired by his/her past experiences and reasoning capabilities.

Note that, here, what is really good for a given individual is not fixed or predetermined independent of the individual's desires or preferences; what is really good for a given individual is what he/she would happen to prefer or desire provided that he/she has full information about the relevant facts. From this, we can explain why Hobbes deems self-preservation to be each and every individual's greatest good as follows. According to Hobbes, self-preservation is each and every individual's greatest good because everybody *would* prefer his/her long-term self-preservation more than anything else once he/she becomes fully aware of the relevant facts. In short, we might interpret Hobbes as being committed to what might be called as a full-information account of the good<sup>12</sup>, which is one particular version of an "*idealized preference-satisfaction theory of the good.*"

---

<sup>12</sup>Noticeable people who have proposed theories that are quite similar in their general spirit in recent times are: [Railton, 1986a,b, Smith et al., 1989, Firth, 1952, 1955, Brandt, 1955, 1998] What all of these theories have in common is that they interpret value – whether it is individual goodness or moral rightness – as a *dispositional property* that invokes a certain kind of *positive psychological reaction to ideally suitable subjects*. The theories differ in what sort of evaluative properties that their theories are trying to propose a dispositional analysis of – is the property moral rightness? [Firth, 1952, 1955, Brandt, 1955, 1998]; or is the property individual goodness? [Railton, 1986a,b, Brandt, 1998]; or is the property simply what people generally value? [Smith et al., 1989] – as well as in what they consider to be "ideally suitable subjects" whose psychological reactions determine the truth or falsity of the type of evaluative statements in question – is it one's own psychological reaction when one is in idealized circumstances that matters?; or is it only some outsider who could provide professional advice that counts?; does the ideally suitable subject have to have full-information or would it be sufficient (or even preferable) to have more limited information?; what other traits (such as the level of imaginative capacities, perceptual awareness, impartiality, etc.) do the theories require ideally suitable subjects possess? and so on.

### 1.6.3 Hobbes's Conception of Practical Rationality and the Reason Why Hobbes did *Not* Hold an Idealized Preference-Satisfaction Theory of the Good.

The problem with this strategy is that there is really no way to make such interpretation *non-circular* in a *non-question-begging* way. This is because, within Hobbes's entire system, *reason* and *rationality* are defined in a way to have self-preservation *as its very aim*, and relevant facts which Hobbes advises one to incorporate in one's deliberation process include certain type of *moral facts* which already assumes that self-preservation is every human being's greatest good. This needs some explanation.

Within Hobbes's system, the two major components of one's rationality are *prudence* and *reasoning*. Whereas prudence is one's ability to project roughly reliable predictions about future consequences based on one's past experiences alone, reasoning is one's ability to generalize these past experiences into conditional statements (e.g. "if one eats ice cream, then one will later get fat.") and to use these conditional statements as basic premises in one's deliberation and syllogistic thinking (e.g. P1: I do not want to get fat. P2: If one eats ice cream, then one will later get fat. C: Therefore, I should not eat ice cream now.") . When this process of reasoning is performed in a systematic way by starting out with correctly defined concepts, the whole process results in what Hobbes calls "scientific knowledge."

By this it appears that *reason* is not, as sense and memory, born with us, *nor gotten by experience only, as prudence is*, but attained by industry, first in apt imposing of names, and secondly by getting a good and orderly method in proceeding from the elements, which are names, to assertions made by connexion of one of them to another, and so to syllogisms, which are the connexions of one assertion to another, till we come to a knowledge of all the consequences of names appertaining to the subject in hand; and that is it men call SCIENCE.

[Hobbes 1994: *Leviathan*, Chapter V, Section 17 emphasis added]

In short, for Hobbes, the very culmination of one's successful use of prudence and reasoning capabilities is scientific knowledge. Hobbes basically repeats the same characterization of science in the following:

When the discourse is put into speech, and begins with the definitions of words, and proceeds by connexion of the same into general affirmations, and of these again into syllogisms, the end or last sum is called the conclusion, and the thought of the mind by it signified is that *conditional knowledge, or knowledge of the consequence of words*, which is commonly called SCIENCE. [Hobbes 1994: *Leviathan*, Chapter VII, Section 4, emphasis added]

After one has arrived at scientific knowledge in a specific area, one can then re-use this scientific knowledge again as basic premises in one's deliberation and syllogistic thinking. So, when Hobbes advises one to base one's desires and preferences on full information (that is, the greatest and surest prospect of consequences), what he is advising one to do is to base one's desires and preferences on various fields of scientific knowledge.

Among the several specific fields of scientific knowledge, there is a field, according to Hobbes, that specifically concerns how to achieve what is good for individuals and societies in general. Hobbes calls such scientific field "moral science" or "moral philosophy" the major contents of which are "the laws of nature."

A LAW OF NATURE (*lex naturalis*) is a precept or general rule, found out by reason, by which a man is forbidden to do that which is destructive of his life or taketh away the means of preserving the same, and to omit that by which he thinketh it may be best preserved. [Hobbes 1994: *Leviathan*, Chapter XIV, Section 3]

And *the science* of them [*the laws of nature*] is the true and only moral philosophy. For moral philosophy is nothing else but the science of what is good

and evil in the conversation and society of mankind. [*Hobbes 1994: Leviathan*, Chapter XV, Section 40 emphasis added]

So, when Hobbes recommends one to form desires and preferences on the basis of full information (i.e. the greatest and surest prospects of consequences), what is included in the list of full information are the laws of nature which, according to Hobbes, are the major contents of moral science and moral philosophy.

However, as we can see in the quoted passage right above, the laws of nature are none other than a list of general rules or precepts that are designed to help one secure one's long-term self-preservation. This means that when Hobbes claims that what is (not just *apparently* good, but) *really* good for somebody is to satisfy the type of desires and preferences which he/she would form after he/she has fully incorporated moral science and the laws of nature into his/her deliberation process, we can see that Hobbes is *already assuming* that securing one's long-term self preservation is each and everybody's greatest good.

That is, since moral science is a field of scientific knowledge which concerns what is really good for people, and since, the contents of moral science, which are the laws of nature, are a set of prescriptions that are specifically designed to tell people what the most effective means to secure their own self-preservation are, we can see that saying that one should form one's preferences on the basis of moral science and the laws of nature already assumes that securing one's long-term self preservation is each and everybody's greatest good.

Furthermore, the fact that moral science and the laws of nature, which specifically aim for the achievement of long-term self-preservation for individuals and society, are the final culmination of one's combined rational faculties indicate that, for Hobbes, self-preservation is the very *goal* of rationality itself. We can even go further and say that, within Hobbes's system, the very concept of practical rationality itself is defined in terms of self-preservation, which Hobbes deems to be each and everybody's greatest good. That is, we can say that, for Hobbes, practical reason and practical rationality is none other than one's ability to recognize

that long-term self-preservation is one's greatest good as well as one's ability to correctly choose the right kind of means to best achieve it.

The fact that this is, indeed, Hobbes's working notion of practical rationality can be (indirectly) confirmed by looking at what kind of preferences Hobbes deems to be *irrational*. As we have seen, for Hobbes, one's preferences are *irrational* whenever, for whatever reason, one does not prefer to act in ways that best secure one's long-term self-preservation. For Hobbes, there are two ways for a person to form irrational preferences.

One way is for one to rightly desire one's long-term self-preservation more than anything else, but to either be insufficiently informed or misinformed about the best available means to achieve it. An example of this would be the case of *the fool* which we have discussed previously. We can say that Hobbes's fool does desire his/her self-preservation; this is why he/she enters into a mutual covenant that aims to provide mutual protection. However, Hobbes's fool is either misinformed or insufficiently informed in the sense that he/she does not properly realize that failing to reciprocate towards his/her partner's cooperation in a mutual covenant would very likely jeopardize his/her chances to secure his/her self-preservation in the long-run despite its short-term gain. Let us call a set of preferences that are irrational in this way - that is, by being based on insufficient (or mis-) information, "unconsidered."

Another way for one to form irrational preferences, according to Hobbes, is for one to be swayed by the wrong kind of passions themselves, (such as vain-glory or self-dejection), which prevent one from desiring one's long-term self-preservation strongly enough. Hobbes calls any desire or passion, which momentarily directs one's attention away from strongly desiring one's long-term self-preservation, "perturbations."

Emotions or *perturbations* of the mind are species of appetite and aversion (. . .)  
They are called perturbations *because they frequently obstruct right reasoning*.  
They obstruct right reasoning in this, that they militate against the *real good* and  
in favor of *the apparent* and most *immediate good*, which turns out frequently

to be evil when everything associated with it hath been considered. (...) Therefore, although *the real good* must be sought in the long term, which is the job of reason, appetite seizeth upon a present good without foreseeing the greater evils that necessarily attach to it. Therefore appetite perturbs and impedes the operation of reason; whence it is rightly called a *perturbation*. [Hobbes 1991: *De Homine*, Chapter XII, Section 1 emphasis added]

However, when the influence of these wrong passions go over the limits of local-level perturbations and one starts to become completely overwhelmed by their overarching influence, this is the point where one is, now, suffering what Hobbes calls “madness.” When one is suffering madness, the irrationality of one’s preferences consists in one’s desiring *the wrong thing* - that is, desiring something (such as glory) at the very expense of one’s long-term self-preservation. Let us call a set of preferences that are irrational in this way - that, by being based on the wrong kind of basic passion, “unbalanced.”

So, within Hobbes’s system, one’s preferences are irrational whenever one prefers to take a course of action that does not best achieve one’s long-term self-preservation, and one’s actual preferences can be irrational either because they are *unconsidered* or *unbalanced*. From this, we can say that, for Hobbes, one’s preferences are *rational* whenever they are *well-considered* and *well-balanced*, and this is so whenever (a) one has correctly identified what course of action best achieves one’s long-term self-preservation (which makes one’s preferences well-considered) and (b) one has properly directed one’s motivational states to perform the course of action that has been so identified (which makes one’s preferences well-balanced.)

So, it is true that Hobbes thinks that people would necessarily desire or prefer their own long-term self-preservation more than anything else when they are fully rational and are fully informed about the relevant facts; but this is so because what Hobbes considers to be “the relevant facts” already include the moral fact that self-preservation is each and every

individual's greatest good, and also because Hobbes's working notion of practical rationality defines rationality as the ability to understand that long-term self-preservation is one's greatest good as well as the ability to be properly motivated to do what has been identified as the best means to achieve it.

This contradicts the main spirit of an idealized preference-satisfaction theory of the good. For if Hobbes really intended to propose an idealized preference-satisfaction theory of the good, he would have had to claim that obtaining glory would be actually better for anybody who preferred glory over his/her self-preservation given that he/she were fully rational and were fully informed about the relevant facts. Rather, as it is clearly shown in his discussion of madness, for Hobbes, the fact that somebody prefers glory over his/her self-preservation simply shows that that the person is *irrational*. I believe that this shows that Hobbes did not hold an idealized-preference satisfaction theory of the good, at least, in its pure descriptive form; within Hobbes's system, self-preservation is objectively the greatest good for everybody.

## Chapter 2

# Freeing Hobbes from the Humean

## Conception of Instrumental Rationality

Until now, we have freed Hobbes from one conventional interpretation of his moral and political philosophy; namely, that Hobbes was a defender of the preference-satisfaction theory of the good. In the previous chapter, we have seen that Hobbes was not committed to the preference-satisfaction theory of the good in any of its variations. However, there is another common interpretation of Hobbes that is coupled with seeing him as an advocate of the preference-satisfaction theory of the good; namely, that Hobbes's conception of rationality was purely instrumental which foreshadows that of David Hume's. This is another conventional interpretation that I would like to free Hobbes from.

### 2.1 David Hume's Conception of Instrumental Rationality

As I have just mentioned, many commentators interpret Hobbes as being committed to an *instrumental view of rationality* which foreshadows that of David Hume. It is quite controversial whether Hume had actually adopted such view of rationality.<sup>1</sup> However, it is quite

---

<sup>1</sup>Nick Sturgeon's unpublished manuscript, "Hume on Reason and Passion", deals with this issue with careful detail in depth.

common to attribute such instrumental conception of rationality to Hume, and, regardless of whether Hume had actually held this view or not, I would like to show that at least *Hobbes* wasn't committed to such view.

It is best to understand what is known as the "Humean conception of practical rationality (i.e. instrumental rationality)" by looking at the passages that Hume wrote himself. According to Hume,

We speak not strictly and philosophically when we talk of the combat of passion and of reason. Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than the serve and obey them. [Hume 1984: *A Treatise of Human Nature*, Book II, Part III, Section III: p. 462]<sup>2</sup>

'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledge'd lesser good to my greater, and have a more ardent affection for the former than the latter. [Hume 1984: *A Treatise of Human Nature*, Book II, Part III, Section III: p. 463]

The thesis that Hume is interpreted as advancing in the first quoted passage is, what is generally known as, the "No-Combat Thesis"; namely, the thesis that there is no way for one's desires or preferences to contradict the dictates of reason and rationality. In other words, according to Hume, it is not the role of reason or one's other rational faculties to *evaluate* and *oversee* whether one's desires or preferences are rational or reasonable. One's desires and preferences are simply things that the individual just happens to have whose contents which the individual's reason or rationality has no control over.

---

<sup>2</sup>The page numbers refer to the Penguin (1984) edition.

The result of this, according to Hume, is that there cannot be any preferences or desires that can properly be called *irrational*, and, thereby, be seen as in conflict with one's rationality. According to Hume, *no preferences are irrational* – it is not irrational to prefer the destruction of the whole world to the scratching of my finger; it is not irrational to prefer my total ruin to the least amount of uneasiness that a totally unrelated foreigner might feel; and it is not even irrational to prefer what I believe to be worse for me to what I believe to be better for me – and it is not the job of one's reason and rational faculties to encourage one to form rational preferences – because, strictly speaking, preferences can *neither be rational nor irrational*.

Then, according to Hume, what are the roles of reason and rationality in the process of deriving one's preferences? According to Hume, there are two major roles that reason and rationality play in this area: one is to help the agent form *true beliefs* about the world, and the other is to help the agent choose *the most effective means* to satisfy the agent's current preferences or desires (whatever they happen to be) in the light of these true beliefs.

This means that there are two basic ways that one's preferences or desires might be (rather imprecisely) called irrational or unreasonable. One way is for one's preferences to be based on a false belief; such as when one prefers not to go to the bathroom at the middle of the night because one falsely believes that there is a scary monster living in the bathroom that only shows up at night. Another way is for one to prefer to use a specific means for a given end which is highly ineffective in accomplishing the end that one has in mind; such as when one prefers to use a toothbrush to open a bottle of beer.

'tis only in two senses, that any affection can be call'd unreasonable. First, When a passion, such as hope or fear, grief or joy, despair or security, is founded on the supposition or the existence of objects which really do not exist. Secondly, When in exerting any passion in action, we chuse means insufficient for the design'd end, and deceive ourselves in our judgment of causes and effects. [Hume

1984: *A Treatise of Human Nature*, Book II, Part III, Section III, p. 463]

However, even though we commonly call these two types of preferences as instances of irrational or unreasonable preferences in our ordinary life, according to Hume, it is not, strictly speaking, the preferences themselves that are irrational or unreasonable, but only the accompanying beliefs that are so.

In short, a passion must be accompany'd with some false judgment, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgment. [Hume 1984: *A Treatise of Human Nature*, Book II, Part III, Section III, p. 463]

So, according to the Humean conception of practical rationality, it might be irrational for Jill, who is currently living a moderately decent life, to prefer to commit suicide by swallowing a chewing gum. But, the reason why we can (rather imprecisely) call such preference irrational is *not* because preferring to commit suicide when one is living a moderately decent life is *itself* irrational, but rather because Jill has not chosen an effective means to achieve her given end, which is to kill herself. If Jill had preferred to kill herself and end her life by using a very effective means (e.g. jump off a 40-story building), then Hume would say that there is nothing irrational about her preferences. In other words, according to the Humean conception of practical rationality, no preferences (not even a preference to kill one's life without having any good reason to do so) can, strictly speaking, be irrational or unreasonable in itself.

Let's summarize the Humean conception of practical rationality into the following formula.

[HUMEAN CONCEPTION OF PRACTICAL RATIONALITY (I.E. INSTRUMENTAL RATIONALITY)]

: The Humean conception of practical rationality (i.e. the purely instrumental view of rationality) is the conjunction of the following two claims:

(a) No preferences or desires, by themselves, are irrational or rational.

(b) (The Role of Reason and other Rational Faculties)

: the basic role of reason and other rational faculties are confined to the following two roles;

(i) To inform the agent with true beliefs (about the world), and

(ii) To inform the agent with the most effective means to achieve a given end in the light of these true beliefs (about the world).

\* That is, reason and rationality say nothing about the agent's ends themselves.

## **2.2 Was Hobbes Committed to the Humean Conception of Instrumental Rationality?**

Now, the question is: "Did Hobbes endorse the Humean conception of instrumental rationality as formulated above?" Many people think that he did. Consider Jean Hampton's early interpretation:

So, for Hobbes, (...) Rationality would therefore be regarded by him as having instrumental value; a rational man would be one whose reason would serve his desires well by determining correctly how those desires could be satisfied. (...) Moreover, (...), Hobbes makes his instrumentalist notion of reason crystal clear (...) [Hampton 1986, p. 35]

David Gauthier's interpretation of Hobbes is similar.

If we accept also Hobbes' contention that 'those actions are most reasonable, that conduce most to their ends' (E.W. iii, p. 133), we can then derive from 'a man must do X to secure what he wants', the further conclusion 'a man, if rational, will do X.' (...) In the state of nature *reason is the slave of the passions*; hence to refer to a right to do what one naturally endeavours to do is otiose. [Gauthier 1969, p. 21, 35, emphasis added]

Hobbes certainly did not deny that one's reason and rational faculties can play such instrumental role in one's deliberation process. In fact, one of the major reasons why Hobbes advises one to bring one's prudence, reasoning, and scientific knowledge to bear on one's deliberation process is, as we have seen, to enable the agent to determine his/her final preferences on the basis of "the greatest and surest prospects of consequences." And, here, we might say that the reason why Hobbes advises one to combine "the greatest and surest prospects of consequences" with one's deliberation process is because, having such information will, generally, reduce the chances of one's final preferences to be based on *false beliefs*, and such information will also, very likely, reveal *what the most effective means* or *what the most effective course of action* could be to achieve the end that one has in mind. In other words, Hobbes *does* think that one's reason and rational faculties can perform the specific roles that are described in clause (b) of the formula of "the Humean conception of practical rationality" stated above.

### **2.2.1 Hobbes Did Not Endorse Clause (b)**

However, as we have already seen, for Hobbes, the roles that reason and rationality play are *not confined* to the two activities of preventing the formation of false beliefs and revealing the most effective means to a given end. In the previous sections, I have already shown that Hobbes thought that any preferences the satisfaction of which are inconsistent with the achievement of one's long-term self-preservation are *irrational*. (As we have seen, this gen-

erally happens when the agent is either (1) insufficiently (or mis-) informed or (2) perturbed by the wrong kind of basic passions.)

On the face of it, this already seems to contradict clause (a) of the above formula – which states that no preferences or desires, by themselves, are irrational. However, we would need to verify that when Hobbes talks about irrational preferences or instances of madness, he is not merely talking about preferences that are based either on false beliefs about the world or on false beliefs about the effective means for a given end, but is also talking about preferences, which, he thinks, are irrational in *themselves*.

But, before we deal with clause (a), let's first start with clause (b). According to clause (b), the role of rationality and reason is confined to providing the agent with true beliefs and revealing what is the best means to achieve whatever end that the agent just happens to have. To put it in another way, clause (b) claims that the role of reason and rationality does not extend to directing one's desires and preferences towards a specific direction; that is, reason and rationality say nothing about the agent's ends themselves.

However, we can see that Hobbes did not endorse clause (b) by reminding ourselves of what Hobbes said about moral science and the laws of nature. According to Hobbes, the final culmination of one's reasoning abilities is scientific knowledge. When the matter is about what is good for individuals and societies in general, the scientific field which provides answers to these questions is what Hobbes calls moral science or moral philosophy. In short, moral science, which is the final culmination of one's practical rationality, informs one about what is good for individuals and societies.

And, since the laws of nature, which constitute the major contents of moral science, are none other than a list of general rules or precepts that are specifically designed to teach the best ways to secure one's long-term self-preservation, we can clearly see that Hobbes actually thought that reason and rationality should properly direct the agent to aim for his/her long-term self-preservation.

This shows that Hobbes did not accept clause (b) of the formula of the Humean concep-

tion of instrumental rationality – which, in effect, claims that reason and rationality can say *nothing* about the agent’s ends themselves. According to Hobbes, reason or rationality *does* say what one’s basic end *should be*; and that is to achieve one’s long-term self-preservation.

### 2.2.2 Hobbes Did Not Endorse Clause (a)

To this, one might object that the fact that the laws of nature aim for self-preservation and that these laws are discovered by right reason does not, strictly speaking, imply that it is the proper role of reason and rationality to direct one’s ends. Rather, “the laws of nature” is a perfect example that shows how the basic role of reason and rationality is confined to the instrumental role described in clause (b). That is, according to this interpretation, the laws of nature, which are discovered by right reason, merely tell people *the most effective means* to achieve what they most basically *want*: namely, their own self-preservation.

To take this approach, one has to assume that Hobbes thinks that it is *a matter of fact* that *everybody* has a basic desire for self-preservation which is stronger than any other desire they happen to have and that the sole job of reason is to choose the most effective means to achieve this end which is *already given*. This seems to be the view adopted by those who attribute the Humean instrumentalist conception of rationality to Hobbes.

If we accept Hobbes’s view that man is a self-maintaining engine then . . . Men want, and necessarily want, to preserve themselves. Therefore, whatever can be shown to be a condition of human preservation, is thereby shown to be a means to man’s end. From premises of the form ‘X is a necessary means to self-preservation’, Hobbes can derive conclusions of the form ‘a man must do X to secure what he want’. . . . we can then derive from ‘a man must do X to secure what he wants’, the further conclusion ‘a man, if rational, will do X’. [Gauthier 1969, p. 21]

Here, we could specify the variable X by “the laws of nature”. According to Gauthier’s inter-

pretation, Hobbes thinks that we are “*self-maintaining engines*” who just cannot help but to desire our own self-preservation, and the main role of reason and rationality is to identify the most effective means – which is to follow the prescriptions of the laws of nature – to secure this already given end. So, when Hobbes seems to be advancing a normative claim that “one should follow the prescriptions of the laws of nature”, he is, strictly speaking, *not* recommending one to desire self-preservation as a specific end, but rather is only recommending one to choose the necessary means to secure what one already wants the most.

If this is the case, then, what’s so wrong with people who seek glory at the expense of their own self-preservation, who, according to Hobbes, are *irrational* and *mad*? Isn’t Hobbes, here, criticizing the specific ends to which glory-seeking people are committed? According to Hampton, what Hobbes thinks is wrong with glory-seeking people is not that they are pursuing the *wrong end*, but rather that they are seeking their own self-preservation *badly*.

So how can Hobbes have an instrumentalist conception of rationality when he is prepared to label as irrational those people who don’t act to pursue their self-preservation? He can have such a conception if that label’s meaning is roughly equivalent to ‘imprudent,’ that is, if the label is critical of such people not because they are pursuing an object rather than self-preservation but because they are perceived to be pursuing self-preservation *badly*. (...) Hobbes’s condemnation would thus convict them of an error in their [instrumental] reasoning, not an error in what they were desiring. [Hampton 1986, p. 36]

This is incorrect. In order for Hampton’s interpretation to be correct, she would have to show that Hobbes thinks that it is perfectly rational to seek glory (given that it is pursued effectively) if one *truly desires* glory and that it is OK to commit suicide (again, given that it is performed effectively) if one *truly desires* to conclude one’s life. However, as we have seen, when Hobbes characterizes certain people as being mad or irrational, he is specifically

criticizing the *abnormal passions* and *desires* that underlie their extreme behaviors. Recall how Hobbes defines “madness”:

In sum, all *passions* that produce strange and unusual behavior are called by the general name of madness. [Hobbes 1994: *Leviathan*, Chapter VIII, Paragraph 20 emphasis added]

Again, that *madness* is nothing else but *too much appearing passion* . . . [Hobbes 1994: *Leviathan*, Chapter VIII, Paragraph 23 emphasis added]

Here, it is apparent that what is Hobbes criticizing here are the passions themselves. And it seems that it would be too much of a stretch to think of what Hobbes calls “madmen” as being merely *imprudent* in our ordinary sense of the term as Hampton suggests. As we have already seen previously, somebody might happen to have an extremely strong desire for glory, and this person might very intelligently choose the most effective means to satisfy this basic desire. This person is certainly *not imprudent* in the sense suggested by Hampton; nonetheless, the person will still count as mad by Hobbes’s own standards, as we have quite clearly seen in the previous chapter.

So, when Hobbes is labeling glory-seekers as “mad” or “irrational”, what he is really criticizing is not merely that these people are choosing *ineffective means* to secure their own self-preservation, but also that these people are being overwhelmed by *the wrong kind of desires* – namely, the desire for (vain) glory – and, thereby, not desiring their own self-preservation *strongly enough*.

This shows that Hobbes did not endorse clause (a) of the above formula of the Humean conception of practical rationality – which claims that no desire or preferences are, by themselves, irrational. What Hobbes is criticizing about glory-seekers is their *ends themselves*, not just the means.

Moreover, since it is clear that Hobbes acknowledges the existence of mad people who do not desire their own self-preservation strongly enough, it is not entirely correct to interpret

Hobbes as regarding human beings as what Gauthier calls “self-maintaining engines”, since this implies that *everybody, as a matter of fact*, has a basic desire for self-preservation which is *stronger than* any other desire they have or would have.

Not only does Hobbes think that it is possible for somebody to prefer glory over self-preservation, he also thinks that there is a certain portion of the human population who actually do prefer glory over self-preservation in the state of nature where government authority is absent.<sup>3</sup>

So, it seems that Bernard Gert’s interpretation of Hobbes’ theory of reason was generally going in the right direction when he claimed:

For Hobbes, reason provides a genuine guide to conduct, one applicable to all rational men; it is not merely a method whereby each man attempts to harmonize or maximize his particular passions. *That is, for Hobbes reason is not, or at least should not be, the slave of the passions, rather the passions are to be controlled by reason.* This is not to deny that “every man by reasoning, seeks out the means to the end which he propounds to himself” but reason does more than this, it has an end of its own, avoidance of violent death. [Gert, “Introduction” in *Hobbes 1991: Man and Citizen*, p. 13 emphasis added]

The more surprising fact is that, in a later article, even Jean Hampton retracts her early interpretation of Hobbes and concedes that Hobbes did think that certain desires or preferences are irrational *in themselves*.

Elsewhere I argued that there was no inconsistency between these passages [i.e. passages where Hobbes apparently criticizes the preferences of glory-seeking people] and that theory of value [i.e. the desire(preferance)-satisfaction theory of good]. *But now I have second thoughts.* [Hampton 1992, p. 340 emphasis

---

<sup>3</sup>As I will argue in a later chapter, Hobbes thinks that it is precisely this fact that makes it inevitable for the state of nature to dissolve into a state of war of all against all.

added]

... Hobbes does appear to criticize certain basic desires themselves, and not merely action from them, as irrational [Hampton 1992, p. 342]

In sum, based on numerous pieces of textual evidence as well as the reasons provided above, we can see that Hobbes did not endorse any one of the two clauses (a) and (b) in the formula of the Humean conception of practical rationality. That is, Hobbes did think that there are certain desires or preferences that are irrational in themselves. Moreover, Hobbes also thought that the role of reason and rationality is not and should not be confined to merely that of informing true beliefs about the world as well as the most efficient means to achieve a given end, but also to *guide the agent's end towards the right direction*. Therefore, we can conclude that Hobbes was not committed to the Humean conception of instrumental rationality.

### **2.3 Reconciliation with Hobbes's General Project**

Until now, I have tried to free Hobbes from two widely held interpretations; namely, the preference-satisfaction theory of the good and the Humean conception of instrumental rationality. In section 1.6, I have explained that one of the major motivations to interpret Hobbes's moral and political philosophy in these two ways is in order to respect Hobbes's general spirit to provide a reductive analysis of normative concepts, and free moral and political philosophy from any vestiges of Aristotelian metaphysics.

However, we have seen that, based on numerous pieces of textual evidence, it is plausible to think that, within Hobbes's moral and political philosophy, long-term self-preservation is objectively each and every individual's greatest good, which one's reason and rationality should rightly aim and seek to achieve, and, hence, Hobbes was committed to none of the

two widely held interpretations. Such interpretation makes much better sense of Hobbes's distinction between real good and apparent good, Hobbes's discussion of madness and irrational preferences, and Hobbes's notion of reason, scientific knowledge, and the laws of nature. Such interpretation makes everything coherent except for one thing. Saying that self-preservation is objectively each and every individual's greatest good would make Hobbes's ethical system normative from its very first premise and would seem to defy Hobbes's general intention to provide an alternate moral and political philosophy to that of Aristotle's. I would like to respond to this worry as follows.

First of all, in the history of western philosophy, there are very few philosophers who had proposed a philosophical system that was completely consistent. And the fact that there are inconsistencies in some parts of a philosopher's works doesn't always damage the greatness of a philosopher's thoughts taken as a whole. I believe that Hobbes was no exception. There are many parts of Hobbes's writings which apparently contradict one another. The best thing we, as modern scholars, can do is to make these apparently contradicting elements coherent as best as we can. If there are any remaining inconsistencies after we go through such process, we cannot help but to admit the existence of such inconsistencies and make the most out of a philosopher's works.

As we have seen, Hobbes was not completely consistent with pursuing his ambitious project of providing an ethical system that can be completely reduced to non-normative facts. This doesn't mean that Hobbes's works are no longer worth reading. I believe that there are still many insights that Hobbes's moral and political philosophy provide that are relevant even for today; such as his claim that external enforcement of some sort is a necessary evil to prevent a sub-optimal social state and his justification for the existence of governments which relies on it. And, in order to make use of these insights and make sense of many other parts of Hobbes's original text, I believe that acknowledging that Hobbes did not entirely pursue his reductive project in a completely consistent manner is a relatively a small price to pay.

Second, if we pause and think about it for a moment, we can see that the fact that self-preservation is assumed to be objectively the greatest good for each and every individual is not completely devastating for the reductive project which Hobbes had originally intended. This is because, compared to Aristotle's notion of the highest good (which assumes the full realization of people's rational capabilities), the assumption that self-preservation is objectively the greatest good seems to be very weak.

That is, if one is supposed to enjoy anything that is good in life, it is evident that one must first have a life in the first place. In this sense, the fact that, within Hobbes's moral system, self-preservation is objectively the greatest good for everybody, which does not depend on its being desired or preferred by anybody, might be interpreted as Hobbes's attempt to present a *minimum threshold* for any desire or preference to count. And, once, this issue is settled, we might say that, for Hobbes, what is really good for any individual is the satisfaction of his/her desires or preferences *given that it is consistent with the achievement of his/her long-term self-preservation*. If we interpret Hobbes this way, we would be able to retain most of Hobbes's original intentions to provide a reductive analysis of normative concepts with only a slight modification. This is the basic plan that I have in mind for chapters 4 and 5.

Third, as we have seen, I believe that it is, now, quite clear that Hobbes did not think that the satisfaction of just any kind of desires or preferences is really good for an individual. So, the simplistic version of the preference-satisfaction theory of the good is not a viable interpretative option for us to take. The only other options would be to interpret Hobbes as holding either an idealized preference-satisfaction theory of the good or an expressivist view about value terms. As we have seen, the first option makes Hobbes's theory of the good normative as well. So, it does not have any advantage over our current interpretation that assumes that, for Hobbes, self-preservation is objectively the greatest good, in terms of respecting Hobbes's general intention to provide a fully reductive analysis of normative concepts. The second option simply ignores Hobbes's general intention to provide a fully reductive analysis of normative concepts that would render normative sentences to bear ob-

jective truth-value.

In other words, none of the other alternate interpretations can fully respect Hobbes's general intention to provide a fully reductive analysis of normative concepts either. If this is the case, then the fact that our current interpretation conflicts with Hobbes's general intention cannot be a reason to deny it, since none of the other viable alternate interpretations respect Hobbes's general intention either. And, if one considers how, unlike the other two interpretations, our current interpretation makes many specific elements of Hobbes's moral and political philosophy consistent, I believe that our current interpretation, albeit being imperfect, is still the best way to see Hobbes's moral and political philosophy.

## Chapter 3

# Freeing Hobbes From Psychological Egoism

### 3.1 The Motivation Behind Attributing Psychological Egoism to Hobbes

Many commentators interpret Hobbes as a *psychological egoist*.<sup>1</sup> Psychological egoism is a view that claims that *all* human actions are motivated, at bottom, exclusively by self-interest; it claims that everybody, in the end, are egoists. This is a very strong claim. It does not merely claim that *only some* or *the majority* of human actions come from considerations of self-interest; but rather that *all* human actions – including the ones that apparently seem to be acts of benevolence as well as those acts that seemingly stem from one’s moral conviction

---

<sup>1</sup>See Butler [1983], Hume [1975], Broad [1950]. Kavka [1986] thinks that some textual evidence does suggest that Hobbes was a psychological egoist, but thinks that only a weakened version of psychological egoism, which he calls “Predominant Egoism” is needed for Hobbes’s political philosophy to work. McNeilly [1966] thinks that Hobbes was at least committed to psychological egoism in his earlier works. Hampton [1986, pp. 20-24] interprets Hobbes as a psychological egoist who maintains that all of our desires are *caused by a “self-interested” bodily mechanism*, and opposes the idea of interpreting Hobbes as a psychological egoist who claims that all of our desires have *self-regarding content*. In other words, according to Hampton, Hobbes does allow people to have certain kinds of other-regarding desires. However, according to Hampton, these other-regarding desires play absolutely no role in Hobbes’s political argument that it is not entirely unreasonable to regard Hobbes as a psychological egoist when one is trying to understand his political philosophy.

– are ultimately motivated solely by a concern for one’s own exclusive personal good.

The major reason why so many people tend to attribute psychological egoism to Hobbes comes from the following passages:

...of the voluntary acts of every man the object is some *good to himself*. [Hobbes 1994: *Leviathan*, Chapter XIV, Section 8]

For no man giveth but with intention of good to himself, because gift is voluntary, and of all voluntary acts the object is to every man his own good... [Hobbes 1994: *Leviathan*, Chapter XV, Section 16]

For Hobbes, an act is *voluntary* if and only if it proceeds from one’s *will*. Here, “the will” is simply one’s final intention to perform the most preferred course of action (which is revealed after deliberation) that is available to that specific individual. So, what the passages above suggest is that, according to Hobbes, everybody, as a matter of fact, aims for their own exclusive personal good whenever they perform their most preferred course of action. As we can see, this sounds pretty close to psychological egoism.

In addition, commentators tend to attribute psychological egoism to Hobbes because they think that psychological egoism is a necessary foundation for Hobbes’s political philosophy; they think that we cannot get Hobbes’s political philosophy without psychological egoism. Hobbes’s justification for the existence of government power relies on Hobbes’s assumption that, absent government enforcement, the state of nature will result in a state of war of all against all. Here, one might naturally ask, “Why is it the case that people in the state of nature, according to Hobbes, cannot live harmoniously and cooperate with one another without external enforcement?” According to conventional interpretation, Hobbes’s answer to this question is: “Because all human beings are, by nature, *selfish*.”

This assumption of universal selfishness is the main ground that Hobbes’s state of nature has been so commonly modeled as a game of *Prisoner’s Dilemma* by contemporary Hobbes scholars who are influenced by modern game theory. What results from people having this

sort of egoistic psychology in the state of nature is a state of war of all against all – a sub-optimal state, which everybody finds undesirable and very much prefers to escape. The only way to escape this dire situation, according to Hobbes, is to establish a sovereign: a political government with unlimited and absolute power.

So, many people think that, for Hobbes, the justification for political government relies on the fact that, without it, the state of nature inevitably dissolves into a state of war, and the main reason why this is so is because people naturally have a strictly egoistic psychology. In short, psychological egoism is the major cornerstone upon which the entire system of Hobbes's political philosophy is erected; for many people, without assuming psychological egoism, Hobbes's political philosophy simply does not work.

However, as I have briefly suggested, psychological egoism is a very extreme and contestable doctrine. Many people think that it is false, and for good reasons. So, to the very extent that Hobbes's political philosophy relies on psychological egoism, we can say that it is based on a very weak foundation. If there is a way to build up Hobbes's political philosophy from a less contestable theory of human psychology without relying on psychological egoism, I believe that this will significantly bolster the general plausibility of Hobbes's political philosophy. This is one of my aims.

There are two things that I intend to argue in this chapter: first, I will argue that Hobbes was not actually a psychological egoist in any plausible interpretation of this doctrine, and second, I will argue that psychological egoism is not really needed for his political philosophy.

## **3.2 What is Psychological Egoism? - Some Clarifications**

So, what is *psychological egoism*? Here is psychological egoism stated in its most general form:

**[PSYCHOLOGICAL EGOISM]:** Everybody is ultimately motivated at bottom only by his/her own self-interest.

There are some things that we need to get clear about in order to understand psychological egoism properly.

First, psychological egoism is a doctrine about people's *ultimate* or *most basic motivations for action*. It intends to provide an answer to the question: "Why did *X* act in that way?" According to psychological egoism, there is a unique answer to all such questions: namely, "In order to promote his/her *self-interest*."

Psychological egoism does not deny that people can sometimes act in *seemingly benevolent ways*. It also does not deny that people can point to certain altruistic reasons to explain their seemingly benevolent actions. What psychological egoism denies is that those altruistic reasons were what *really* or *ultimately motivated* such people. According to psychological egoism, regardless of whether one is consciously aware of it or not, one's actions – even those actions that are seemingly benevolent – are ultimately motivated at bottom by one's own self-interest.<sup>2</sup> This leads to our next characteristic.

Second, one doesn't necessarily have to be *consciously aware* of the fact that one is ultimately motivated by one's self-interest in order for one's actions to qualify as instances of psychological egoism. It might be true that everybody is, in fact, ultimately motivated by self-interest, but such motivation might be so deeply embedded within people's subconscious states that not everybody is consciously aware that his/her actions are motivated in this way. Or people might simply be what Gauthier calls "self-maintaining engines"<sup>3</sup> who are just programmed (like machines) to pursue their own self-interests without always be-

---

<sup>2</sup>Drawing from contemporary psychological learning theories (such as that of Hull and Skinner), in one of his earlier papers, Michael Slote suggests that there could be an empirical basis for psychological egoism such that all higher-order drives and motives (e.g. altruistic and benevolent motives) are functionally dependent on a certain number of basically "selfish" unlearned primary drives and motives. [See Slote 1964]

<sup>3</sup>See (Gauthier 1969, p. 21)

ing consciously aware that they are motivated in this way. Psychological egoism would still be true if either one of these two doctrines (or a combination of both) are universally true. This means that psychological egoism has more to do with the *underlying psychological mechanism* rather than the *motivational contents* of one's desires and actions.

Third, psychological egoism is not a doctrine of *achievement*. Although it claims that everybody is ultimately motivated by one's own exclusive self-interest, it does not claim that everybody (or even most people) *actually succeeds* in achieving their own self-interest. This is something that is not that hard to understand if one thinks about the difference between *attempting to achieve something* and *actually achieving that thing*.

Such failure to achieve one's self-interest usually occurs when one has one or more *false beliefs*. For instance, suppose that one is motivated to promote one's physical health (which is, intuitively, a major component of one's self-interest), and, thereby, regularly takes a herbal medicine, which is scientifically proven to be carcinogenic, by falsely believing that the herbal medicine possesses some mysterious powers that contributes to longevity. In such case, one is not really achieving one's best self-interest; quite the contrary. Nonetheless, this does not change the fact that one was primarily motivated by one's self-interest. So, such example is not a counter example to psychological egoism.

Fourth, psychological egoism is a *descriptive theory* of human psychology; not a *normative theory* of human psychology. It claims that it is *a matter of fact* that all human motivations are ultimately based on one's exclusive self-interest; not that people's motivations *should be* ultimately based on their exclusive self-interest. There is a standard name for the latter type of doctrine which is strictly normative: *ethical egoism*. Ethical egoism claims that, generally speaking, people *should* promote their own exclusive self-interest more than anything else.

The purpose of this section is to see whether Hobbes was a psychological egoist, not whether Hobbes was an ethical egoist. I want to deny that Hobbes was a psychological egoist, or at least, psychological egoism is not needed for his political philosophy. I take no

stance on whether Hobbes was an ethical egoist.

### 3.3 Was Hobbes a Psychological Egoist?

As we have seen, psychological egoism generally claims that all human actions are ultimately motivated by one's own self-interest. Different people have different conceptions of what a person's self-interest consists in. And this, in turn, results in slightly different versions of psychological egoism.

#### 3.3.1 Was Hobbes a Psychological Hedonist?

One version of psychological egoism interprets a person's good in purely *hedonistic terms*. This is a version that is usually known as *psychological hedonism*. Psychological hedonists define a person's good as the experience of pleasure and the absence of pain. In fact, they go slightly further than this; according to psychological hedonism, the experience of pleasure and the absence of pain *exhaust* a person's good or well-being.

So, self-interest, according to psychological hedonism, is none other than the experience of pleasure and absence of pain. From this, psychological hedonism can be characterized as follows:

**[PSYCHOLOGICAL HEDONISM]:** All human actions are motivated ultimately by a basic desire to experience pleasure and to avoid pain.

There are some people who have understood psychological egoism in this particular way when they attributed psychological egoism to Hobbes.<sup>4</sup> However, as long as one's theory of

---

<sup>4</sup>This seems to be the view of F. S. McNeilly in [McNeilly 1966], where he argues that Hobbes was not a psychological egoist after he had wrote Leviathan because, unlike Hobbes's earlier works, pleasure no longer

the good (or well-being) allows the possibility of things other than the experience of pleasure and the absence of pain to constitute or contribute to one's self-interest, one need not be committed to psychological hedonism in order to be committed to psychological egoism; one can think that all human actions are ultimately motivated by one's own self-interest even if one thinks that not all human actions are motivated by a basic desire to experience pleasure and avoid pain.

It is not very hard to see that Hobbes was not a psychological hedonist in our current interpretation. We have seen that, within Hobbes's moral and political philosophy, the achievement of long-term self-preservation is objectively each and everybody's greatest good. Note that actions that best secure one's prospects of self-preservation need not be *pleasurable*. For example, exercising regularly might be a good way to maintain physical health, but exercise is not always a pleasurable thing to do even when it is obvious that it would make the person healthier.

Furthermore, it is clear that Hobbes did not think that the fact that somebody is experiencing pleasure, by itself, guarantees that the person is achieving something that is truly good for him/her. According to Hobbes, there exists people who experience intense pleasure by conquering and having superior power over other people; namely, the glory-seekers.<sup>5</sup> However, Hobbes emphatically denies that these glory-seekers are people who are seeking what is truly advantageous for them<sup>6</sup>; which is their long-term self-preservation. In fact, he calls these type of people "mad."<sup>7</sup>

---

plays a central role in his philosophy. Jean Hampton also seems to be interpreting Hobbes as a psychological hedonist when she claims that Hobbes is committed to the view that all of our desires are ultimately produced by pleasure-producing and pain-avoiding physical mechanisms. Hampton thinks that Hobbes is *not* committed to the view that every human desire has *self-regarding motivational content*. However, her interpretation is still a version of psychological egoism (more specifically, psychological hedonism) according to our current framework. See [Hampton 1986, pp. 23-24]

<sup>5</sup>"Joy arising from imagination of a man's own power and ability is that exultation of the mind which is called GLORYING..." [Hobbes 1994: *Leviathan*, Chapter VI, Section 39]

<sup>6</sup>As we will later see, these people are the main culprits of the state of nature descending into a state of universal war.

<sup>7</sup>"The passion whose violence or continuance maketh *madness* is ... great vain-glory... [Hobbes 1994: *Leviathan*, Chapter VIII, Paragraph 18, emphasis on "madness" is mine]

So, not only did Hobbes not think that the experience of pleasure and the absence of pain exhausts a person's self-interest, but he also did not think that pursuing pleasure can always be regarded as a case of pursuing one's real self-interest. Therefore, we can at the very least say that Hobbes was not a psychological hedonist. However, this isn't yet sufficient to show that Hobbes was not a psychological egoist, since it might still be the case that Hobbes thought that all human actions are ultimately motivated by a basic desire to promote one's own self-interest, (and, thereby, was committed to psychological egoism), even though he thought that such actions may not always result in pleasurable (as well as less painful) experiences.

### **3.3.2 *Hobbes's Dictum and Tautological Egoism***

Now, many places within Hobbes's text seem to suggest that Hobbes was committed to the following doctrine:

**[HOBBS'S DICTUM]:** Everybody is motivated by their current desires or preferences - that is, everybody always aims to satisfy their current desires or preferences whenever they act.<sup>8</sup>

The doctrine, as it is stated, is not a version of psychological egoism. It is merely a theory of motivation. The reason why Hobbes's Dictum is not yet a version of psychological egoism is because it is not supplemented by a theory of self-interest or personal good. It claims that people in general are motivated by their current desires and preferences; however, it is silent on the issue of whether satisfying these desires and preferences will be actually good for the people in question.

---

<sup>8</sup>“Hobbes's Dictum” is a name that Nick Sturgeon has suggested for me to use.

So, there is an apparent way for Hobbes's Dictum to become a version of psychological egoism – namely, by being combined with the preference-satisfaction theory of the good, which claims that it is always good for a person to satisfy his/her current desires or preferences. That way, Hobbes's Dictum together with the preference-satisfaction theory of the good would virtually be claiming that everybody is motivated by his/her own self-interest – namely, to satisfy his/her current desires and preferences. This is a version of psychological egoism that is sometimes called “tautological egoism.”<sup>9</sup>

The adjective “tautological”, here, is purposely used to express the writer's reluctance to acknowledge tautological egoism as a genuine version of psychological egoism. This is understandable because tautological egoism does not in any way restrict the type of desires or preferences that somebody may have in order for him/her to count as an *egoist*. Suppose that you are a saint who greatly sacrificed your own personal well-being to advance some humanitarian cause. A tautological egoist will say that you are an *egoist*, a *selfish* person, rather than a *selfless* person; this is because you preferred to advance such humanitarian cause yourself, and by acting accordingly, you satisfied such preferences, which amounts to your own good. So, as long as people are motivated by their desires and preferences, we can clearly see that tautological egoism is virtually *non-falsifiable*.

One should note at this point that being supplied with any other theory of self-interest than the preference-satisfaction theory of the good will not make Hobbes's Dictum a version of psychological egoism. Suppose that one is a hedonist. Then, as long as one thinks that there can be certain preferences the satisfaction of which are not always pleasurable, one is not a psychological egoist even when one accepts Hobbes's Dictum.

So, was Hobbes committed to Hobbes's Dictum? It seems so. Consider how Hobbes defines voluntary action. According to Hobbes, “a voluntary act is that which proceedeth

---

<sup>9</sup>See Gert (1967) and “Introduction to Thomas Hobbes” contained in Hobbes (1991) and Kavka (1986, Chapter 2)

from the will”<sup>10</sup> and the will is simply “the last appetite in deliberating”<sup>11</sup>. In other words, what Hobbes is basically saying here is that every voluntary action proceeds from one’s final preferences that emerges after one’s practical deliberation process. This is basically Hobbes’s Dictum.

However, as we have just seen, the only way for Hobbes’s Dictum to imply psychological egoism is for it to be combined with the preference-satisfaction theory of the good. But, as we have seen previously, there are numerous pieces of textual evidence that suggest that Hobbes was not committed to the preference-satisfaction theory of the good. This means that the fact that Hobbes was committed to Hobbes’s Dictum does not make Hobbes a psychological egoist.

### **3.3.3 What Psychological Egoism is for Hobbes and Whether He Endorsed it**

Remember that psychological egoism generally claims that everybody ultimately seeks (either consciously or unconsciously) to achieve their own personal good. I have explained that different versions of psychological egoism can arise depending on what one thinks a self-interest consists in part. Hobbes thinks that a person’s *real self-interest (or real good)* consists in the achievement of the person’s own long-term self-preservation. This means that, for Hobbes, psychological egoism would amount to be claiming the following:

**[What “Psychological Egoism” claims for Hobbes]:** Everybody ultimately aims to achieve their *real self-interests (or real good)* whenever they act - that is, one’s actions are always motivated at bottom by a desire to achieve one’s own long-term self-preservation.

---

<sup>10</sup>Hobbes (1994, *Leviathan*: Chapter VI, Section 53)

<sup>11</sup>Hobbes (1994, *Leviathan*: Chapter VI, Section 53)

So, in order to see whether Hobbes was committed to psychological egoism, we would have to see whether Hobbes endorsed the above claim.

We can see that glory-seeking people, of which Hobbes acknowledges the existence, have the potential to falsify the claim that Hobbes had endorsed psychological egoism as it is formulated above. Glory-seekers are the type of people who pursue power and glory even at the very expense of their long-term self-preservation. So, obviously, these people are not achieving their real self-interest – namely, their long-term self-preservation – when they act. However, as we have seen, psychological egoism is not a doctrine of achievement. It is perfectly consistent with psychological egoism that people generally fail to achieve their real self-interest as long as they are ultimately motivated by it.

This means that we would have to distinguish between *two types* of glory-seekers. The first type of glory-seekers are the type of people who are, indeed, ultimately motivated by a basic desire to secure their own long-term self-preservation (that is, their *real self-interest*), but falsely believe that displaying typical glory-seeking behaviors is the best way to achieve their long-term self-preservation. The second type of glory-seekers are the type of people who display typical glory-seeking behaviors because they are motivated at bottom by a basic desire or passion for (*vain*) *glory*, which is, according to Hobbes, *not* their *real good* (*self-interest*.)

If I can show that Hobbes thought that there are at least some people who seek glory in the second type of way, then this suffices to show that Hobbes was not a psychological egoist in our current understanding of the term. Consider Hobbes's general discussion of glory and vain-glory:

*Joy arising from imagination of a man's own power and ability is that exultation of the mind which is called GLORYING...*" [Hobbes 1994: *Leviathan*, Chapter VI, Section 39]

Also, because there be some that taking pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their security requires, [Hobbes 1994: *Leviathan*, Chapter XIII, Section 4]

Vain-glorious men ... are inclined to rash engaging... [Hobbes 1994: *Leviathan*, Chapter XI, Paragraph 12]

The passion whose violence or continuance maketh *madness* is ... *great vain-glory*, ... [Hobbes 1994: *Leviathan*, Chapter VIII, Paragraph 18, emphasis on "madness" and "great vain-glory" are mine]

Here, we can clearly see that Hobbes is referring to people who are ultimately motivated by the passion or desire for glory itself; this is different from *falsely believing* that seeking glory is the best way to achieve one's long-term self-preservation. We can further see that Hobbes does not think of glory as something that constitutes one's *real self-interest* (or *real good*); he denounces the passion for glory as being a major cause of *madness*.

So, we can clearly see that Hobbes thought that these second type of glory-seekers are a real possibility. Actually, not only does Hobbes think that these second type of glory-seekers are a real possibility, but he also thinks that these second type of glory seekers actually exist, and as it will later turn out, the existence of such glory-seekers is essential to Hobbes's political philosophy.<sup>12</sup> So, according to Hobbes, not everybody is ultimately motivated by their real self-interest, which is their long-term self-preservation; some are ultimately motivated by a passion for glory which is Hobbes deems to be a representative case of madness. Therefore, Hobbes did not hold psychological egoism.

Then, how are we supposed to make sense of the typical passages that apparently seem to suggest that Hobbes was committed to psychological egoism? Let's go back to the major passages which many people think give support for thinking Hobbes as a psychological egoist.

---

<sup>12</sup>They are the main culprits for the state of nature dissolving into a universal state of war.

...of the voluntary acts of every man the object is some *good to himself*. [Hobbes 1994: *Leviathan*, Chapter XIV, Section 8]

For no man giveth but with intention of good to himself, because gift is voluntary, and of all voluntary acts the object is to every man his own good... [Hobbes 1994: *Leviathan*, Chapter XV, Section 16]

As we have seen before, Hobbes explicitly distinguishes between *real good* and *apparent good* (or what is merely *called good*.)<sup>13</sup> Therefore, the term “good” in these passages can be interpreted to mean either of two things: (1) an individual’s *real good* (i.e. real self-interest) or (2) an individual’s *apparent good*.

If the term “good” in these passages denote an individual’s *real good*, then we would have to say that Hobbes was, indeed, asserting psychological egoism in the above passages. Remember that a voluntary action, for Hobbes, simply denotes an action that proceeds from one’s final preferences after deliberation. The passages above would, then, basically be saying that whenever people act in their most preferred way, they always ultimately seek to achieve their real good (i.e. real self-interest), namely, their long-term self-preservation. This is psychological egoism for Hobbes.

However, as we have just seen, this conflicts with Hobbes’s acknowledging the existence of vain-glorious people who do not ultimately pursue their real self-interest or real good. If one pursues such interpretation, one has no choice but to say that Hobbes was just plainly inconsistent.

---

13

... good (like evil) is divided into *real* and *apparent*. [Hobbes 1991: *De Homine*, Chapter XI, Section 5]

Whence it happens that inexperienced men that do not look closely enough at the long-term consequences of things, accept what appears to be good, not seeing the evil annexed to it; afterwards they experience damage. And this is what is meant by those who distinguish good and evil as *real* and *apparent*. [Hobbes 1991: *De Homine*, Chapter XI, Section 5]

However, we have an alternate option. We could interpret the term “good” in the above passages as denoting an individual’s *apparent good*. An apparent good, for Hobbes, is something that the individual just happens to prefer; it is *apparently* good because the things that people just happen to prefer are not always *really* good or in the best interest of the person in question. Suppose that we interpret the term “good” in the above passages as denoting an individual’s apparent good. Then, this combined with Hobbes’s definition of voluntary act as an act “that which proceedeth from the will”<sup>14</sup> as well as his definition of the will as “the last appetite in deliberating”<sup>15</sup>, would imply that the sentence, “of all voluntary acts the object is to every man his own good”, would virtually be expressing the same proposition as the sentence, “For every human being, the basic aim of acting in one’s most preferred way is to obtain what one happens to prefer.” In other words, what Hobbes is claiming in the above passages is that everybody is motivated by his/her preferences. This is just Hobbes’s Dictum, and, as we have seen, Hobbes’s Dictum does not imply psychological egoism.

In order to show that this, indeed, is a better interpretative strategy, we would have to show that interpreting the term “good” as denoting an individual’s apparent good does not lead us to any similar inconsistency that interpreting the term as denoting an individual’s real good does. And, we can see that it, indeed, does not.

Let’s go back to behaviors of the second type of glory-seekers. When these glory-seekers acted voluntarily in typical glory-seeking ways, they were obviously not aiming to achieve their own real good. However, we can still say that the glorious outcomes that were sought by these glory-seekers when they acted voluntarily (i.e. when they acted in their most preferred way) were, despite its being not really good for them, something that was at least *what they most preferred*. As we have seen, for Hobbes, something that one happens to most prefer at a given moment is one’s apparent good. Therefore, we can say that when these second type of glory-seekers acted voluntarily by acting in their most preferred way, their basic aim

---

<sup>14</sup>Hobbes (1994, *Leviathan*: Chapter VI, Section 53)

<sup>15</sup>Hobbes (1994, *Leviathan*: Chapter VI, Section 53)

was to achieve their *apparent* or *seeming good*. So, we can say that, for Hobbes, everybody, including *even* these second type of glory-seekers, seek their (apparent) good when they act voluntarily.

This is exactly what Hobbes would be saying in the above passages if we interpret the term “good” as denoting an individual’s apparent good. And this seems to be the only way to make the above passages consistent with what Hobbes says about the existence of the second type of glory-seekers who obviously do not aim to achieve their long-term self-preservation, which, according to Hobbes, is their real self-interest.

In short, Hobbes did not endorse psychological egoism. The many passages that make it seem that he is advancing psychological egoism can be explained away by interpreting them as stating Hobbes’s Dictum. And, Hobbes’s Dictum does not imply psychological egoism unless it is combined with the preference-satisfaction theory of the good – a doctrine that Hobbes rejects.

### **3.3.4 Was Hobbes a *Subjective Egoist*?**

Now, we have just seen that Hobbes did not hold psychological egoism if we follow Hobbes’s own conception of real self-interest interpreted as the achievement of one’s long-term self-preservation. However, even though Hobbes was, indeed, not committed to psychological egoism understood in terms of his notion of real good (or real self-interest), one might still think that he is committed to a doctrine that many people would regard as a version of psychological egoism. And that is a doctrine that claims that every human action aims to achieve what each person *thinks* or *believes* to be his or her own real personal good. Let’s call this doctrine, “subjective egoism” and formulate it as follows:

**[SUBJECTIVE EGOISM]:** Everybody seeks to achieve what they *think* or *believe* to be their own personal good - that is, all human actions are ultimately motivated at bot-

tom by a basic desire to promote what *each person thinks or believes* to be his own real self-interest.

Note the difference between psychological egoism in its general form and subjective egoism. Psychological egoism generally claims that everybody is always motivated at bottom by their own self-interest. Here, what constitutes a person's self-interest is specified by what theory of personal good the evaluator adopts.

An important thing to remember is that one need not be *consciously aware* of the motivational contents (viz. that they are directed towards one's own self-interest) of one's basic desires in order for one's actions to qualify as instances of psychological egoism; it is sufficient for one's actions to be ultimately derived from an underlying psychological mechanism that always forces one to seek (either consciously or unconsciously) one's self-interest whenever one acts.

By contrast, for subjective egoism, the motivational contents of each person's actions matter. Subjective egoism claims that everybody is always motivated at bottom by a basic desire to achieve what each person *thinks or believes* to be his or her self-interest. So, in order for somebody's action to qualify as an instance of subjective egoism, that person would have to have been consciously aware that he or she was pursuing what he or she thinks or believes to be in his or her best self-interest.

Furthermore, what somebody *thinks or believes* may be *incorrect*. So, once we have fixed the contents of our working theory of self-interest, the fact that object *g* is in individual *x*'s best self-interest does not necessarily imply that individual *x* *would think or believe* that object *g* is in his or her best self-interest. Conversely, the fact that individual *x* *thinks or believes* that object *h* is in his or her best self-interest does not necessarily imply that object *h* is *in fact* in *x*'s best self-interest.<sup>16</sup>

---

<sup>16</sup>The two coincide only if we adopt a theory of self-interest that claims that whatever a person thinks or believes something to be good for him or her, that thing is, by that very fact, also really good for him or her.

Let me illustrate the distinction between psychological egoism and subjective egoism by a specific example. Suppose that, according to my theory of self-interest, maintaining good physical shape is a major component of a person's good. Suppose that somebody was aiming to achieve what is in fact very bad for that person according to this theory of self-interest; say, the person was ultimately motivated by a basic desire to intake as many saturated fat calories as possible.

Assuming my theory of self-interest, such a person is a counter example to psychological egoism; that is, the person was, in fact, not ultimately aiming for his own self-interest when he tried to consume all of the junk food that he could find. However, such person is not necessarily a counter example to subjective egoism. This is because even though the person did not actually aim for his real self-interest, the person could still have (mistakenly) *thought* or *believed* that intaking as many saturated fat calories as possible would be in his/her best self-interest.

The same thing can be applied to the second type of glory-seekers that we have discussed above. For Hobbes, the existence of this second type of glory-seekers is a counter-example to psychological egoism; these glory-seekers are the type of people who do not seek what is really good for their own self-interest by being ultimately motivated by a basic desire for (vain) glory, (which Hobbes sees as a representative case of a pathological desire.) However, it might still be the case that these glory-seekers think, albeit mistakenly, that the achievement of glory is much better than the achievement of long-term self-preservation, and, thereby, are generally aiming for what they *think to be (albeit mistakenly)* what is really good for them. In other words, even though the typical actions of glory-seeking people cannot be properly said to be ultimately based on a basic desire to promote what is *in fact* in their self-interest (which is their long-term self-preservation), it might still be argued that these actions are ultimately based on a basic desire to promote what glory-seeking people *think* to be in their

---

This means that the two coincide only when one adopts the simplistic version of the preference-satisfaction theory of good.

self-interest.

Showing that Hobbes was not a subjective egoist is more difficult than showing that he was not a psychological egoist. This is because even the second type of glory-seekers cannot count as counter-examples to subjective egoism. However, I believe that this can still be done.

Consider again the typical actions displayed by glory-seeking people. When somebody decides to pursue glory at the expense of one's own self-preservation, this might be so for three different reasons.

1. The person might (mistakenly) think that seeking glory is the best way to achieve his or her own self-preservation – which he or she correctly thinks to be the greatest good for him or herself. If this is the case, then such person is not a counter-example of subjective egoism (or even psychological egoism). This is because the typical glory-seeking behaviors that such person displays are in fact ultimately based on a basic desire to promote what is in fact his/her best self-interest – namely, his/her own long-term self-preservation. The fault of such person's behaviors lies in his/her faulty *instrumental reasoning* – (that is, he/she mistakenly think that glory-seeking behaviors are a good means to securing self-preservation) – not in a fault in the type of ultimate basic desire that the person had derived his or her preferences from. So, although such person is not acting in a way that is actually good for his/her own long-term interest, he/she is still acting in a way which he or she *thinks* or *believes* to be good for his or her long-term interest. Furthermore, we can say that such person's actions are still motivated at bottom by a basic desire to secure his or her self-preservation, which, according to Hobbes, is the person's greatest good. Therefore, the existence of such glory-seeking person does not show that Hobbes was not a subjective egoist (or even a psychological egoist).
2. The person might (mistakenly) think that the achievement of glory itself is a greater

good than the achievement of his/her own self-preservation, and, thereby, bases all of his/her actions ultimately on a basic desire for (vain) glory. If this is the case, then, although such person is a counter-example of psychological egoism, such person still does not count as a counter-example of subjective egoism. This is because although the actions of such person are not in fact ultimately based on a basic desire to promote his or her own personal good, we can say that the person's actions are still based ultimately on a basic desire to promote *what the person thinks or believes* to be his or her own personal good. So, the existence of such glory-seeking person only reaffirms that all human actions are ultimately motivated by what each person thinks or believes to be his or her own good, which is what subjective egoism claims.

(These are the two types of glory-seekers that we have encountered in the previous section. I, now, suggest a third type of glory-seeker. )

3. The person might correctly recognize that securing his/her own self-preservation is a much greater good than achieving glory and might also correctly judge that displaying such (glory-seeking) behaviors is generally not a very good way to secure his/her own self-preservation in the long run, but might have been swayed by his or her vehement passion for glory, and, thereby, acted in a typical glory-seeking way. If this is the case, then not only would the person's glory-seeking behaviors not be based on a basic desire to promote what is in fact conducive to the person's own good, but such behaviors would also not be based on a basic desire to promote what the person *personally thinks or believes* to be conducive to his/her own good. In such case, the person's behaviors will be ultimately based on a basic desire (i.e. a basic desire for glory or vain-glory) the satisfaction of which the person fully recognizes to be very likely to be detrimental to the achievement of his or her self-preservation, which the

person, again, fully recognizes to be his/her greatest good. The existence of this third type of glory-seeking people, unlike the first two types of glory-seeking people above, *does* serve as a counter-example to subjective egoism.

What remains to be shown is whether Hobbes himself acknowledges this third type of glory-seeking people as a genuine possibility. If he does, then I believe that we can reasonably conclude that Hobbes was not (even) a subjective egoist. The only thing that needs to be done is to show that Hobbes does not necessarily think that people always *think* that they are *acting in their own best self-interest* whenever they act.

Note that it is not very hard to observe this type of situation in our daily lives. I might prefer to play video games when I should really be studying for an important exam that is scheduled the next day. However, just because I preferred to play video games, and, then actually played video games instead of studying for the exam, does not necessarily imply that I was *thinking* or *judging* that playing video games was best for me at that very moment. I could have *correctly judged* that studying for the exam would be in my best interests at that very moment, but could have succumbed to my strong desire to play video games.

This is, I believe, an instance that can be explained by what Hobbes calls *the perturbations of the mind*. Let's go back to the part where Hobbes discusses about perturbations of the mind. There, Hobbes writes:

Emotions or perturbations of the mind are species of appetite and aversion . . . They are called perturbations because they frequently obstruct right reasoning. . . . although . . . *good must be sought in the long term, which is the job of reason, appetite seizeth upon a present good without foreseeing the greater evils that necessarily attach to it. Therefore appetite perturbs and impedes the operation of reason; whence it is rightly called a perturbation.* [Hobbes 1991: *De Homine*, Chapter XII, Section 1 emphasis added]

Here, we can see that Hobbes is distinguishing between the role of reason and the role of appetites and desires that eventually become perturbations. Reason instructs the agent to seek long term good, while appetites and desires, which eventually become perturbations, impel the agent to concentrate solely on the immediate good without considering the long term consequences that a given action would bring. The separation between the role of reason and the role of appetites and desires that eventually become perturbations leaves room for an agent to correctly judge what the best course of action would be by his or her reason, while being impelled to act in a way which he or she acknowledges to be contrary to his or her best interest by being overwhelmed by the influences of particular desires and aversions that are perturbations.

This type of phenomenon - failing to act in ways that one judges to be best - is usually known as *the problem of weakness of will*. One might think that Hobbes does not acknowledge that the problem of weakness of will is even possible. As we have seen, a voluntary act, for Hobbes, is an act that proceeds from one's will, where one's will is supposed to be one's final preference (in Hobbes's words, "last appetite in deliberating") to perform a specific action. The will, interpreted this way, corresponds to the notion of *revealed preference*<sup>17</sup> that is used in modern economics theory. In the theory of revealed preferences, one's actions or choices *reveal* what one has preferred. Likewise, for Hobbes, one's voluntary acts *reveal* what one has willed. For Hobbes, there is a necessary connection between a person's will and the person's voluntary actions; any desires that do not connect to a person's actions are, according to Hobbes, merely, *inclinations*.

And though we say in common discourse, a man had a will once to do a thing,  
that nevertheless he forbore to do, yet that is properly but an *inclination*, which

---

<sup>17</sup>The intuitive idea is something like this. Suppose that somebody chose to buy the consumption bundle *X*, when buying an alternative consumption bundle *Y* was perfectly within his/her budget constraints. Here, economists say that the person's action *reveals* that he/she preferred the consumption bundle *X* to the consumption bundle *Y*. This is because if he/she chose to buy *X* when he/she *could have* bought *Y* instead, then the reason for this, presumably, is because he/she preferred the consumption bundle *X* to the consumption bundle *Y*. See, [Varian, 2006, Chapter 7]

makes no action voluntary; because the action depends not of it, but of the last inclination or appetite. [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 53]

So, there is a sense in which the problem of weakness of the will does not exist within Hobbes's moral system. Since every voluntary action is, by definition, the result of one's will, and since one's will simply denotes one's final preference to perform a specific action after deliberation, there is no way that one could have acted contrary to one's final preferences; for Hobbes, one could not have preferred something but acted otherwise. This doesn't necessarily imply that Hobbes denied that one could act contrary to one's *best judgment*. As it is possible for one to *judge* that it would not be in one's best self-interest to play video games before the day of a very important exam, but, *prefers* to do so anyway; it is possible for somebody to *judge* that displaying typical glory-seeking behaviors would be detrimental to the achievement of his/her long-term self-preservation, which is one's real self-interest, while *preferring* to display such behaviors anyway. The passage above shows that this could be possible if one's deliberation process gets *perturbed by*, say, a vehement passion for glory.

So, we can see that Hobbes's discussions of perturbations make room for him to acknowledge the possibility for an agent to prefer to act in a way which the agent herself recognizes to be contrary to her own best interest. That is, we don't necessarily have to interpret Hobbes as being committed to the view that all human actions are ultimately motivated at bottom by a basic desire to promote what each person *thinks to be* his or her own exclusive personal good. For Hobbes, some preferred acts may be acts that the agent does not think or believe to be best. Therefore, Hobbes doesn't necessarily have to be seen as a subjective egoist.

I am aware that my arguments do not conclusively show that Hobbes was not committed to subjective egoism; it only alludes to a possible interpretation of him being not. However, from my own perspective, such is not that much of an important issue as I will now argue that, whatever is Hobbes's stance on psychological egoism, psychological egoism is not needed for his political philosophy.

### 3.4 Psychological Egoism is Not Needed for Hobbes's Political Philosophy

Many people think Hobbes has to assume some form of egoistic psychology in order to explain why he thinks that the state of nature, which is without any government authority or enforcement (and so is a state of anarchy) results in a state of war of all against all. This state of war of all against all is a miserable situation which everybody wants to escape. And the universal misery of the state of nature is, for Hobbes, what justifies the establishment of governments.

One might naturally ask why Hobbes thinks that the absence of government inevitably dissolves into a state of universal war. For instance, why can't people live harmoniously and peacefully without the enforcement of government power? What is usually regarded as the Hobbsian answer to this question is: people are *naturally and universally selfish*.

For example, Kavka lists six characteristics of general human psychology which he thinks are needed in order to properly explain why the state of nature, according to Hobbes, inevitably dissolves into a state of war, and, thereby, justifies Hobbes's argument against anarchy.

There are six such characteristics, described below, (...) All play a substantial role in Hobbes's arguments against anarchy, and several play a further role in other of his arguments.

1.*Egoism*. Individuals are primarily concerned with their own well-being, and act accordingly.

2.*Death-aversion*. Individuals are strongly averse to their own death, and act accordingly.

3.*Concern for Reputation*. Individuals care about their reputations, about what others think of them, and they act accordingly.

4.*Forwardlookingness*. Individuals care about their future, as well as present, well-being, and act accordingly.

5.*Conflicting Desires*. Satisfaction of one person's desires often interferes with, or precludes, satisfaction of another person's. (...)

6.*Rough Equality*. People are fairly equal in their intellectual and bodily powers. (...)

[Kavka [1986, pp. 33-34]]

Among these six characteristics, the first refers to psychological egoism<sup>18</sup> while the third refers to people's desires for glory. According to Kavka, these characteristics play a substantial role in Hobbes's argument against anarchy. In other words, according to Kavka, some version of a general egoistic psychology is needed in order for Hobbes's argument against anarchy, which is the foundation of Hobbes's entire political philosophy, to work.

As I have already explained in the beginning of this section, many contemporary Hobbes scholars, who have been influenced by modern game theory, have modeled the state of nature situation as a game of one-shot *Prisoner's Dilemma*.<sup>19</sup> It is true that this gets the work done; modeling the state of nature situation as a game of *Prisoner's Dilemma* does explain very well why, absent government enforcement, people in the state of nature will confront a constant state of war.

However, in order to model Hobbes's state of nature as a one-shot game of *Prisoner's Dilemma*, one has to assume that Hobbes thinks that people will generally *prefer* to perform unitary defection even when there is assurance<sup>20</sup> that the other player will cooperate. This

---

<sup>18</sup>Kavka later in the chapter argues that it is not, strictly speaking, psychological egoism, but only, what he calls, "predominant egoism" that is needed for Hobbes's argument against anarchy to work. According to predominant egoism, human beings in general are *predominantly* motivated by their own good or well-being. This leaves open that people can sometimes in certain situations act in altruistic and benevolent ways. I personally think that not even predominant egoism is needed for Hobbes's; what is needed is only that people are not self-less masochists.

<sup>19</sup>See Kavka [1986, pp. 109-112], Gauthier [1969, p. 79], Gauthier [1984, p. 170]

<sup>20</sup>Of course, if there *is* such assurance, then the game will no longer be a PD game. What I am emphasizing here is *the type of preferences* that each of the players in the PD has independent of the other player's preferences; the main point is that, for each player in the PD game, defection strictly dominates cooperation.

means that in order for the one-shot Prisoner's Dilemma game to be the correct model of Hobbes's state of nature, one has to think of Hobbes as being committed to some very restricted form of psychological egoism; that is, one would have to interpret Hobbes as thinking that people universally would not prefer to cooperate with another cooperator in the state of nature.

This is very implausible. Many people would think that they would actually prefer to cooperate with other cooperators rather than taking advantage of them in the state of nature. So, if the fact that the state of nature will inevitably deteriorate into state of constant war can only be explained by assuming that people are in general what Gauthier calls *straightforward maximizers*<sup>21</sup> (which are the type of people that a very restricted version of psychological egoism claims people in general to be), then this means that Hobbes's entire political philosophy is based on a very weak foundation.

So, if we could find a way to explain why Hobbes's state of nature will inevitably deteriorate into a state of constant war without relying on commonsensical psychological egoism, then I believe that it will significantly bolster the plausibility of Hobbes's political philosophy.

I believe that such task can be done. Consider Hobbes's own explanation for the cause of conflict in the state of nature.

In the state of nature there is in all men a will to do harm, but *not for the same reason* or with equal culpability. One man practices the equality of nature (...) this is the mark of *modest man* (...) Another, supposing himself superior to other, wants to be allowed everything (...) that is the sign of *an aggressive character*. In his case, the will to do harm derives from *vainglory*. [Hobbes 1997: *On the Citizen*, Chapter 1, Section 4 emphasis added]

---

<sup>21</sup>See [Gauthier, 1984, Chapter VI, Section 2.1]

Also, because there be *some* that taking pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their security requires, if *others* (*that otherwise would be glad to be at ease within modest bounds*) should not by invasion increase their power, they would not be able, long time, by standing only on their defence, to subsist. [Hobbes 1994: *Leviathan*, Chapter XIII, section 4 emphasis added]

Here we can see that Hobbes does not think that everybody in the state of nature will be primarily motivated by a desire for glory or vain-glory. It is true that Hobbes thinks that there would be a certain portion of the entire population that would indeed be aggressive and glory-hungry. However, Hobbes acknowledges that many people will be quite content with living within modest boundaries; these modest people will gladly prefer to cooperate with other people as long as they are assured that their counterparts will cooperate in return and not take advantage of them. However, Hobbes thinks that even these modest people will generally tend to attack other people. This is because,

And from this *diffidence* of one another, there is no way for any man to secure himself so reasonable as anticipation, that is, by force or wiles to master the persons of all men he can, so long till he see no other power great enough to endanger him. [Hobbes 1994: *Leviathan*, Chapter XIII, section 4 emphasis added]

In other words, the main reason why even the modest people tend to attack other people in the state of nature, according to Hobbes, is because they are unsure about whether their counterparts will actually return their cooperative behaviors with cooperation. That is, a typical modest person will be unsure about whether his or her counterpart is another modest person like him, or a glory-seeker who will take advantage of his good will.

If this is the case, then the main reason why the state of nature dissolves into a state of war of all against all is not because everybody naturally seeks glory or tends to maximize

his/her own immediate gain whenever they can, but rather because people in the state of nature are faced with *uncertainty*.

If uncertainty is the major cause of war in the state of nature, then I believe that the Hobbesian state of nature can be better modeled as a *Incomplete Information Bayesian Game* rather than a complete information Prisoner's Dilemma game.<sup>22</sup> If this is the case, then one does not need to assume that people in general would prefer to take advantage of other people's cooperatives behaviors. That is, one does not need to assume any restricted form of egoistic psychology in order to explain the conflict that arises in Hobbes's state of nature. The only assumption that is needed is that people in general are not masochists who would gladly prefer to be taken advantage of others' ill intentions. This, I believe is a very modest and plausible assumption. In short, regardless of whether Hobbes was actually committed to such view, we do not really need psychological egoism to make sense of Hobbes's political philosophy.

---

<sup>22</sup>One of the later chapters will be primarily concerned with providing such model.

## **Part II**

# **Reinvigorating The Leviathan**

## General Introduction for Part II

Our task, in part I, was to free Hobbes from three conventional interpretations with which he is usually associated; I believe that we have, thereby, *unleashed the Leviathan* from the traditional chains that obstructed its appreciation. In part II, we will try to restore and build-up the *Leviathan's* powers to its full potential via reconstruction.

The preliminary reconstruction process starts from chapter 4, titled, “Reconstructing Hobbes’s Theory of the Good as an Ideal Advisor Theory.” There, Hobbes’s theory of the good will first be summarized into an axiomatic format, and, then the theory will be reconstructed as a version of an ideal-advisor theory of the good. The purpose of reconstructing Hobbes’s theory of the good into an ideal-advisor theory is mainly for convenience. By identifying a person’s good as the satisfaction of the preferences that one’s idealized-self *would form on behalf of oneself*, we will be then in a position to utilize contemporary utility theory to give a precise formal representation of a person’s good from the perspective of Hobbes’s moral and political philosophy.

Chapter 5, titled, “A Contemporary Decision Theoretic Reconstruction of Hobbes’s Theory of the Good”, performs exactly the task that the title of the chapter suggests. Building up from the axiomatic summary that has been provided in chapter 4, chapter 5 fully develops the inchoate notions of quantity, summation, aggregation, and probability that were originally expressed in Hobbes’s works in the lights of contemporary utility theory. As a result, we will get a more sophisticated version of the theory of the good that Hobbes himself had implied in his works.

Building up from the formal apparatus that we have established in chapter 5, in chapter 6, titled, “A Bayesian Game-Theoretic Reconstruction of Hobbes’s State of Nature”, we will oppose to the traditional attempts to model Hobbes’s state of nature as a Prisoner’s Dilemma game, and provide what I believe is the best game-theoretic model of Hobbes’s state of nature to date. Not only does the Bayesian game-theoretic model provided here provide a more accurate representation of Hobbes’s state of nature than the traditional Prisoner’s Dilemma game by attributing the main cause of conflict to, just as it is described in Hobbes’s original text, uncertainty rather than selfishness, it frees Hobbes’s political philosophy from being based on a very contestable theory of human psychology which many people believe to be false: namely, psychological egoism.

## Chapter 4

# Reconstructing Hobbes's Theory of the Good as an Ideal-Advisor Theory of the Good

One of the main purposes of Hobbes was to erect a deductive system in which all normative conclusions about morality and political philosophy can be derived from non-normative facts. The project, initially, was intended to be reductive to the very bottom level in the most ambitious way. Normative concepts, such as moral or political obligation, moral goodness and rightness, were supposed to be grounded on and derived from facts about human psychology, such as what people, as a matter of fact, desire and prefer, and these facts about human psychology were, in turn, supposed to be ultimately grounded on and explained by physical motions. The major motive behind such grand project was to purge moral and political philosophy of any vestiges of Aristotelean metaphysics and to align moral and political philosophy perfectly with the then burgeoning scientific view of the universe.

However, as we have seen so far, Hobbes wasn't able to pursue this project in a completely consistent manner. This is because Hobbes's entire moral system completely rests on the fact that self-preservation is each and every individual's *greatest good*. As we have

seen, for Hobbes, it is not that self-preservation is the greatest good because everybody just happens to desire it as a matter of fact; rather it is because self-preservation is each and every individual's greatest good that each and every individual *should* desire it and aim to achieve it as best as he or she can. The fact that, within Hobbes's moral system, the goodness of self-preservation does not really depend on anybody's actually desiring it is reflected in Hobbes's discussion of madness. Madness, for Hobbes, is defined as being under the influence of any type of passion (e.g. a desire for glory) that makes one act in ways that are generally detrimental to the attainment of one's self-preservation. In essence, for Hobbes, madness consists in desiring or preferring to do something at the very expense of one's self-preservation. And as we have seen so far, the fact that something else is preferred to self-preservation does not make that something better than self-preservation. This, again, is because, for Hobbes, self-preservation is each and every individual's greatest good, and the reason why suffering madness is so bad is mainly because it directs the individual's attention away from its very achievement.

Keeping all of this in mind, let us now try to reconstruct Hobbes's theory of good in a slightly more systematic way.

## 4.1 Systematizing Hobbes's Theory of Personal Good

As I have just mentioned, the whole system of Hobbes's moral and political philosophy is based on the fact that *self-preservation* is *objectively* one's *greatest good* for each and every individual. Let's state this as our primary axiom for the task of reconstructing Hobbes's moral system.

**[AXIOM (A1)] (*Primary Axiom For Hobbes's Moral Philosophy*):** For every individual, self-preservation is (objectively) his/her greatest good.

We now define Hobbes's notion of general rationality as follows:

**[DEFINITION (D1)] (*General Rationality*):** Rationality is the ability to identify one's greatest good and direct one's actions towards its achievement.

First of all, I would like to mention that rationality, here, is construed in a very different manner than how it is usually construed in other social sciences, such as economics. In economics or standard rational choice theory, rationality is defined as a set of consistency requirements (e.g. [negative] transitivity, a[nti]symmetry, and so on) that one's preferences need to meet in order for those preferences to be properly represented by a suitable real-valued utility function. Here, rationality is defined as a set of *formal constraints* that are imposed on the *preference-relation itself*; it does not question nor evaluate the specific contents of those preferences.

This is not how Hobbes sees rationality. For Hobbes, reason has an *end of itself*; it aims at and seeks to achieve one's greatest good. And, as I have stated in the axiom 1, there is something that Hobbes deems to be (objectively) the greatest good of each and every individual: namely, his or her self-preservation. From this, we can derive the following proposition.

**[PROPOSITION (P1)] (*Definition of Substantial (Hobbesian) Rationality*):** Rationality is the ability to identify what is most conducive to the achievement of one's self-preservation and direct one's actions towards it.

This follows from axiom 1 and the definition of general rationality. Let us regard the contents of this proposition as our definition of "substantial Hobbesian rationality." As we have seen

from chapter 1, one's rationality is the operation and realization of one's combined rational faculties, which are one's prudence, reasoning, and scientific knowledge.

It should be noted that it is possible for somebody who is *irrational* in this substantial (Hobbesian) sense to be, nonetheless, *rational* in the standard decision theoretic sense that I have just explained.

An example of such person would be the "consistent glory-seeker" that we have seen previously. A consistent glory-seeker is a person who prefers outcomes that are glorious over outcomes that are conducive to his or her self-preservation. The fact that this person prefers glory over self-preservation indicates that this person is irrational in the substantial (Hobbesian) sense stated above.

However, this does not imply that the preferences of this consistent glory-seeker violates any of the standard axioms of decision theory. For example, this person's preferences might be perfectly transitive; that is, if this person sees outcome A as more glorious than outcome B and, thereby, strictly prefers outcome A to outcome B, then whenever this person encounters a third outcome C which the person perceives to be far less glorious than outcome B, which makes the person strictly prefer outcome B to outcome C, the person, then, strictly prefers outcome A to outcome C as well. If this is the case, then there is nothing irrational about this person's preferences in the standard decision theoretic sense. It would be useful to have a separate notion of this minimal or formal sense of rationality in our tool box.

**[DEFINITION (D2)] (*Formal (Minimal) Rationality*):** One's preferences are formally (or minimally) rational if and only if they satisfy the axioms of decision theory.

Here, one should note that this definition of "formal (or minimal) rationality is context-specific. There are many different decision theoretic representation models that rely on different sets of axioms. So, when I say that one's preferences are formally or minimally

rational in the decision theoretic sense, I will be referring to the specific decision theoretic model that is being applied in that particular context.<sup>1</sup> Again, the fact that one's preferences are formally (or minimally) rational in this sense merely indicates that one's preferences are mutually consistent enough for there to be a real-valued utility function representing them. We will later use this notion of formal (or minimal) rationality to derive the utility functions of glory-seekers.

Now, if we look at the definition of substantial Hobbesian rationality (i.e. P1) carefully, we can see that it requires one's rational abilities to perform two distinctive tasks: one is to *identify* what would best achieve one's self-preservation, and the other is to *direct* (by effectively motivating) one's actions towards achieving what has been so identified. Whenever one's combined rational abilities perform these two tasks successfully, one will always prefer to act in ways that are actually conducive to one's own self-preservation. This means that if one happens to prefer to act in ways that are not only in-conducive but are actually detrimental to the achievement of one's self-preservation, then this means that one's rational capabilities are failing in either (or both) of these two ways. The resulting preferences, from the perspective of the substantial Hobbesian sense, will, thereby, be irrational. We have just derived the following notion of irrational preferences:

**[PROPOSITION (P2)] (*Irrational Preferences*):** One's preferences are irrational in the substantial Hobbesian sense if and only if one prefers to act in ways that are detrimental to the achievement of one's self-preservation

As we have just seen, there are two distinct ways for one's preferences to be (substantially) irrational. One is for one to fail to correctly identify what best achieves one's self-preservation.

---

<sup>1</sup>Asymmetry and negative transitivity might be the only set of axioms that I am referring to if the context is requiring only an ordinal representation, whereas the so called independence and archimedean axioms might be needed if the context is requiring an expected utility representation.

The other is for one to be not strongly motivated to act according to what has been identified as the best way to achieve self-preservation. As we have seen from the previous chapters, according to Hobbes, the former happens when one is insufficiently informed about the relevant facts, while the latter happens when one is swayed by the wrong type of basic passion.<sup>2</sup> I have called preferences that are irrational in the first way, "unconsidered", and preferences that are irrational in the second way "unbalanced." (As we have seen, when one's preferences are unbalanced, it is usually the result of what Hobbes calls "madness.")

We now derive the corollary of P2.

**[COROLLARY OF P2 (C1)] (*Irrational Preferences*):** One's preferences are irrational in the substantial Hobbesian sense if and only if one's preferences are (a) unconsidered or (b) unbalanced.

As we have seen, Hobbes thinks that the satisfaction of such substantially irrational preferences are only *apparently* (as opposed to *really*) good for the agent. The following notion of rational preferences is logically equivalent to the corollary.

**[Corollary of P2 (C1)] (*Rational Preferences*):** One's preferences are (substantially) rational if and only if they are (a) well-considered and (b) well-balanced.

When an agent's preferences are well-considered and well-balanced - that is, when the agent's preferences are substantially rational - the agent will always prefer to act in ways

---

<sup>2</sup>Here, the wrong kind of basic passion, in most cases, will denote one's basic desire for (vain-) glory; however, any other basic passion (such as "severe dejection" as mentioned by Hobbes) that obstructs the proper operation of one's basic desire for self-preservation in one's deliberation process would be equally deemed as "the wrong kind of basic passion."

that are, in fact, most conducive to his or her self-preservation. According to Axiom 1, self-preservation is each and every individual's greatest good. Something that is the greatest good implies that it is *really good*. From this, we now arrive at the main theorem of Hobbes's theory of individual good.

**[MAIN THEOREM (T1)] (*Real Good*):** An object *O* is *really good* for a given individual *X* if *O* satisfies *X*'s *rational preferences* - preferences which are both well-considered and well-balanced.

This can be derived from A1, P2 and its corollaries. For Hobbes, the satisfaction of just any kind of preferences is *apparently good*. However, when the preferences are substantially irrational, their satisfaction would be *only* apparently good and *not* really good. The satisfaction of preferences that are substantially rational are *not only* apparently good *but also* really good.

There is one thing that we should be careful about here. Stated in this particular way, it seems as though the fact that one has preferred something, given that such preferences are rational, is *what actually made* that something to be *really good* for that particular person. In other words, it seems that one's rational preferences are (what may be called as) *the good-makers*.

However, this is not entirely correct. Within Hobbes's moral system, what is really good for any given individual is already fixed and given; it is, namely, the person's own self-preservation which amounts to the person's greatest good. The basic role that rationality plays in this picture is simply to *identify* what course of action would best achieve this pre-determined greatest good - namely, the individual's self-preservation, and *direct* the individual's non-cognitive passions so that the individual actually manages to prefer acting in this pre-identified way.

When a given individual's preferences are (substantially) rational, this simply means that these preferences have *detected* and, thereby *reflect*, the best course of action that would lead to the individual's self-preservation. It is true that what is really good for a given individual (namely, the achievement of the individual's long-term self-preservation) is *extensionally equivalent* to the satisfaction of the individual's rational preferences. However, this does not imply that the status of self-preservation being the greatest good for a given individual was *causally determined* by its being preferred by some rational agent.

## **4.2 Reconstructing Hobbes's Theory of Personal Good as an Ideal-Advisor Theory**

We have just seen that, within Hobbes's moral system, achieving long-term self-preservation is objectively the greatest good for each and every individual. Practically speaking, in order for one to properly achieve this greatest good, one's preferences would have to be substantially rational in the Hobbesian sense. This means that one would have to be sufficiently informed about the relevant facts (such as the nature of the consequences that are expected to ensue by performing each available course of action as well as their respective probabilities of occurring) as well as be primarily concerned about achieving one's long-term self-preservation when making a decision on how one should act.

It is not very hard to expect that, in reality, people are not always in a position to form substantially rational preferences in this way; in some situations people might not have enough information and in some situations people might be swayed by the wrong kind of passion and wrongfully value something that is inimical to the achievement of his or her long-term self-preservation.

In such cases, it might be quite true that somebody else who does have a sufficient amount of information and is, more than anything else, concerned about the agent's long-term self-

preservation could be in a *better position* to make decisions on how to act *on behalf of the agent*. This is a possibility that Hobbes explicitly acknowledges.

... so that he who hath by *experience* or *reason* the greatest and surest prospect of consequences deliberates best himself, and is able, when he will, to *give the best counsel unto others*. [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 57]

The person who is able to deliberate best for himself would be, of course, a person who is capable of forming substantially rational preferences. However, Hobbes states above that a person who deliberates best for himself (namely, a person who is sufficiently informed and is motivated in the right way) is not only in a position to form substantially rational preferences for him or herself, but such person would also be in a position to provide the best available *advice* concerning what would be substantially rational to prefer *for other people* as well.

Of course, such person, in order to give proper advice concerning the formation of substantially rational preferences, would need to be primarily concerned about the *advisee's* long-term self-preservation rather than his or her own when he or she is giving advise. Furthermore, among the set of sufficient information that the advisor possesses, complete information concerning the advisee's bio-medical condition as well as the advisee's psychological dispositions (such as the set of desires and aversions the advisee has) should be included. For instance, if the advisee happens to have diabetes, then certain types of foods that are normally recommended to people who do not have diabetes might not be recommendable to the advisee. If the advisee happens to have a very weak heart (in the physical sense), then certain types of exercises might not be recommendable. Also, if the advisee is psychologically disposed such that he or she feels certain types of situations abnormally stressful or unpleasant, then this fact should be taken into account when providing advice to the advisee.

Now, among the many entities (such as one's parents, close friends, significant other, etc.) that might be able to perform such advisory role, the best candidate would seem to be what I would call one's "*idealized-self*". A person's idealized-self is an imaginary entity that

possesses full information concerning the nature as well as the likelihoods of the chain of consequences that will ensue by performing a specific course of action available to the person at a given time. A person's idealized-self also possesses complete first-hand knowledge of the actual person's bio-medical condition, personal history as well as his or her psychological dispositions.

In chapter 1, I have argued that Hobbes was not committed to any version of the preference-satisfaction theory of the good. Not only did I argue that Hobbes wasn't committed to the most simplistic form of the preference-satisfaction theory of the good, but, I have also, specifically, argued that there is no way for us to interpret Hobbes as being committed to, what may be called, "the idealized preference-satisfaction theory of the good" (i.e. the theory that claims that what is really good for somebody is the satisfaction of the type of preferences that he/she would have formed if he/she were deliberating in idealized circumstances) in a *non question-begging way*. So, a reader might naturally wonder why I am heading towards a direction which I have argued against previously.

However, the reader should note my acknowledgment at the end of chapter 2 where I claimed that it may be possible for us to interpret Hobbes as advancing an idealized preference-satisfaction theory of the good *given that* we are willing to relax just one requirement in Hobbes's ambitious project of trying to propose a moral system that is fully reducible to non-normative facts; that is, to grant that long-term self-preservation is *objectively* each and every individual's greatest good.

By granting this one single relaxation, we are now able to respect the rest of Hobbes's reductive project by interpreting him as proposing a version of an idealized preference-satisfaction theory of the good that claims: "An object *O* is really good for somebody if and only if (a) it is consistent with the achievement of the person's long-term self-preservation and, (b) it satisfies the type of preferences that the person would have formed if he/she were deliberating in idealized circumstances."

Such formulation is undoubtedly normative from the very start; it tries to define what is

really good for somebody by already assuming that the person's long-term self-preservation is objectively good for him/her. However, analyzing normative terms in such non-reductive ways is not an uncommon feature of similar types of *dispositional theories of value* that have been proposed in contemporary ethical theory. Consider the dispositional theory of value that Michael Smith proposes:

According to the dispositional theorist we can analyze rightness in terms of a disposition to desire under suitable conditions. But what are these 'suitable conditions'? The dispositional theorist should say that an agent's  $\phi$ -ing is right just in case he would desire to  $\phi$  if he were to deliberate in accordance with the principles of reason corresponding to *moral principles, principles that permit us to derive evaluative truths from truths about our circumstances, principles like the principle of limited altruism*. [Smith et al. 1989, p. 110, emphasis added]

In other words, according to Smith, doing  $X$  is morally right if and only if one would desire to do  $X$  if one were to deliberate in accordance with certain *moral principles* which we already accept as *valid principles of reason*. Such analysis tries to ground moral truths on the basis of desires and preferences that we would have formed under idealized circumstances; yet, it never tries to be fully reductive. So, a dispositional theory of value, (which idealized preference-satisfaction theories of the good could be seen as a specific type of), need not be fully reductive to be meaningful.

Given that we relax the requirement of full reductivity, we can see that Smith's dispositional analysis of value (if one remembers) is very similar to the general spirit of Hobbes's analysis of real good; as we have seen in chapter 1, Hobbes thinks that what is really good for a given individual is the satisfaction of the type of desires or preferences that he/she would form after he/she went through a deliberation process that had fully incorporated *the laws of nature*, which, according to Hobbes, are "precept[s] or general rule[s], found out by reason, by which a man is forbidden to do that which is destructive of his life or taketh

away the means of preserving the same, and to omit that by which he thinketh it may be best preserved.” [Hobbes 1994: *Leviathan*, Chapter XIV, Paragraph 3] Here, as one can see, the achievement of long-term self-preservation is already assumed to be objectively each and every individual’s greatest good. But, this (as I have just argued) does not entirely prevent us from reconstructing Hobbes’s theory of real good as some sort of dispositional theory.

So, let us build into our notion of a person’s *idealized-self* that he/she knows that long-term self-preservation is objectively the greatest good for his/her actual-self, and cares, more than anything else, to help achieve this aim for his/her actual-self. We might think of a person’s idealized-self as a private angel, who is fully knowledgeable about the person’s circumstances, and who, more than anything else, wants the person’s life to go as well-off as possible.

As it may be readily expected, an actual person, due to numerous practical constraints, might not always be in a position to form substantially rational preferences. However, the preferences that are formed *on behalf of* a person *by* the person’s *idealized-self* in any given situation is guaranteed to be substantially rational. Let us state this fact as an extension of corollary C1.

**[EXTENSION OF C1 (E1)] (*Alternate Characterization of Rational Preference*):** One’s preferences are substantially rational (in the Hobbesian sense) if and only if it is the type of preference that *would be* formed by one’s *idealized-self on behalf of one’s actual-self*.

The fact the preferences formed by one’s idealized-self is substantially rational (in the Hobbesian sense) means that the satisfaction of the preferences given by one’s idealized-self is guaranteed to be *really good* for one’s actual-self in any given situation; note that this was not true of the satisfaction of one’s actual preferences. From this, we are now able to derive an alternate formulation of real goodness for a given individual.

[**ALTERNATE THEOREM (T2)**] (*Real Good*): An object *O* is *really good* for individual *X* if and only if *O* satisfies *the preferences formed by O's idealized-self on behalf of O's actual-self*.

This is derived from T1 along with the fact that the preferences formed by one's idealized-self on behalf of one's actual-self are guaranteed to be both well-considered and well-balanced, and, thereby, substantially rational (i.e. E1).

As noted before, one should be cautious not to think that it was the fact that the preferences were given by one's idealized-self that *made* its satisfaction really good for the individual in question. The primary role that one's idealized-self plays in this picture is simply to correctly detect and suggest the best way to achieve what has already been predetermined to be objectively one's greatest good - namely, one's long-term self-preservation.

Viewed in this way, we can see that our reconstruction of Hobbes's theory of non-moral good can be seen as a version of, what might be called, "The the Ideal-Advisor Theory of Good", which has been proposed by Peter Railton.<sup>3</sup> According to Railton's account, what is objectively non-morally good for a given individual is what one's *idealized-self* – who has an "unqualified cognitive and imaginative powers, and full factual and nomological information about his physical and psychological constitution, capacities, circumstances, history and so on. [Railton, 1986a, contained in 2003, see pp. 10-11]" would *advise one's actual-self to do*. Let's see how Railton describes his theory himself.

...let us introduce the notion of an *objectified subjective interest* for an individual

*A*, as follows. Give to an actual individual *A* unqualified cognitive and imagi-

---

<sup>3</sup>See [Railton, 1986b,a]. Other people who have presented a similar approach regarding (not a person's individual good) but regarding moral goodness or moral right are Roderick Firth and Richard Brandt. See [Firth, 1952, 1955, Brandt, 1955, 1998]. These latter views (that concern moral goodness and moral rightness) are usually known as "the ideal *observer* theory". People who try to distinguish the sort of approach taken by Railton (which concerns a person's own individual good) with these latter views sometimes call the former view "the ideal *advisor* theory". (Note that Railton does extend his basic approach to questions concerning moral goodness and moral rightness in the latter parts of the papers.)

native powers, and full factual and nomological information about this physical and psychological constitution, capacities, circumstances, history, and so on. A will have become A+, who has complete and vivid knowledge of himself and his environment, and whose instrumental rationality is in no way defective. We now ask A+ to tell us not what he currently wants, but what he would want his nonidealized self A to want - or, more generally, to seek - were he to find himself in the actual condition and circumstances of A. [Railton 1986a, contained in 2003, p. 11]

This, as we have just seen, is remarkably similar to our current reconstruction of Hobbes's theory of (non-moral) good. Our reconstruction of Hobbes's theory of (non-moral) good further resembles Railton's ideal-advisor theory in the sense that, just as on our theory, Railton thinks that it is not, properly speaking, what one's idealized-self's wants one's non-idealized-self to want that *makes* something to be objectively in one's subjective interest. Rather, it is, what Railton calls, *the reduction basis*, that objectively determines and grounds what is non-morally good for a given individual.

Let us say that his reduction basis is the constellation of primary qualities that make it be the case that [A] has a certain *objective interest*. That is, we will say that Lonnie has an objective interest in drinking clear liquids in virtue of this complex, relational, dispositional set of facts. *Put another way, we can say that the reduction basis, not the fact that [A+] would have certain wants, is the truth-maker for the claim that this is an objective interest of [A's.]* The objective interest thus explains why there is a certain objectified interest, not the other way around. [Railton 1986a, contained in 2003, pp. 11-12, emphasis added]

The reduction basis in our reconstruction of Hobbes's theory of (non-moral) good would be, by using Railton's terminology, the complex, relational, dispositional set of facts that makes the achievement of long-term self-preservation objectively good for each and every

individual. So, again, our current reconstruction of Hobbes's theory of (non-moral) good is remarkably similar to that proposed by Railton.

Now, the question to ask concerning our current reconstruction is this: to what extent can we reasonably say that our current reconstruction of Hobbes's theory of non-moral good is really *Hobbesian*?

First, it is obviously not Hobbesian in the conventional sense since the conventional interpretation sees Hobbes as advocating one of the most simplest version of the preference-satisfaction theory of the good. However, I believe that I have, in the previous chapters, provided sufficient textual evidence that demonstrates that Hobbes was not committed to such theory. So, we can say that what is commonly thought as *Hobbesian theories* in contemporary moral philosophy do not accurately represent Hobbes's real views. So, the fact that my reconstruction is rather *un-Hobbesian* in this conventional sense does not really show that my theory is un-Hobbesian properly sort of speak.

However, it might be argued that, even if we concede that Hobbes was not committed to the most simplistic version of the preference-satisfaction theory of the good, reconstructing Hobbes's theory of non-moral good as a version of the ideal-advisor theory is going too far. For one thing, there is nowhere where Hobbes explicitly mentions the notion of one's idealized-self giving advise to one's non-idealized-actual-self. The only textual evidence that justifies my reconstruction of Hobbes's theory of non-moral good as a version of the ideal advisor theory comes from the quoted passage above which claims that the best deliberators are the type of people who could give the best counsel to other people as well.

However, Hobbes does think that one's deliberation process can be significantly improved by being sufficiently influenced by one's combined rational faculties, and that the preferences derived from such deliberation process would be *substantially more rational* than the preferences derived from deliberation processes that were devoid of such rational process. Furthermore, we have seen that Hobbes explicitly distinguishes between what is *apparently good* and what is *really good* for a given individual, and strictly thinks that it

is only the satisfaction of one's substantially rational preferences that count as really good for the individual.<sup>4</sup> This already implies that the preference-ordering that a given individual would form if he or she were substantially rational, in very many cases, would be *quite different* from the individual's actual preference-ordering.

In other words, we can say that there are basically *two* preference-orderings that are associated with any given individual; (a) the individual's *actual* preference-ordering, and (b) the individual's *substantially rational* preference-ordering. The fact that this view is actually *Hobbesian* is supported, as we have seen, by various textual evidence found in Hobbes's work.

Now, my suggestion is that we just go a little further and think of an individual's substantially rational preference-ordering, (which is distinguished from the individual's actual preference-ordering), as the preference-ordering given by what we have been calling the individual's "idealized-self." I believe that such a stretch is innocuous since it is simply interpreting an individual's substantially rational preferences, which Hobbes already acknowledges to be possibly distinct from the individual's actual preferences, to be the type of preferences that would be given by the individual's idealized-self on behalf of the individual's non-idealized-actual-self.

As we can see from the quoted passage above, Hobbes clearly thinks that people who deliberate best for themselves are also the type of people who can provide the best counsel to *others*. If it is the case that Hobbes thinks that the best deliberators (who are the best deliberators in virtue of being substantially rational) can give the best advice to *other people*, I don't see any reason why Hobbes would deny that these people are the type of people who

---

<sup>4</sup>So, it is slightly ironic when Railton characterized Hobbes as an advocate of the preference-satisfaction theory of good as we have seen from Chapter 1, and then argued that:

Yet this theory is deeply unsatisfactory, since it seems incapable of capturing important elements of the critical and self-critical character of value judgments. [Railton, "Fact and Value"(1986b) contained in 2003, p.49]

The full-information ideal advisor theory of good that Railton proposes is intended to solve this very problem of capturing the critical and self-critical character of value judgments. What Railton did not notice was the fact that Hobbes had already proposed a theory of good that was very similar to the one that he had proposed!

could also give the best advice (not only to other people, but also) to *themselves* in non-idealized circumstances. So, I believe that adding the notion of preferences given by one's idealized-self is, on the whole, consistent with a broadly Hobbsian spirit.

Now, the reason for adding the notion of an ideal-advisor into our current reconstruction of Hobbes's theory of non-moral good is only for convenience. Once we have the notion of two distinct preference-orderings (one attributed to one's actual non-idealized-self and the other attributed to one's idealized-self), we are now in a position to utilize contemporary utility theory to give a numerical representation of these two distinct preference-orderings.

Contemporary utility theory has been developed in order to analyze and give mathematical representations of people's preferences. Contemporary utility theory attempts to do this by constructing "utility functions" that represent a person's preference-relation. A utility function is a mathematical function that takes alternatives (or consequences or outcomes or propositions) as arguments and generates a real-number as its value such that: for a given individual, alternative  $a$  is preferred to alternative  $b$  if and only if  $U(a) > U(b)$ .

In this section, I have tried to identify a person's real good with the preferences formed by the person's idealized-self. The major strength of this approach is that, by this identification, the utility function that is supposed to represent the *preferences* of the person's idealized-self can now be further interpreted as representing what is *really good* for the person's non-idealized actual self. In other words, the utility function of a person's idealized-self can be seen as representing, not only the person's idealized-self's preference-ordering, but also, what John Broome calls, the actual non-idealized person's *betterness relation*.<sup>5</sup> I believe that such approach could give a more precise understanding of some of Hobbes's key texts that we have encountered previously.

---

<sup>5</sup>See [Broome, 1991, pp. 121-122]

## **Chapter 5**

### **A Contemporary Decision Theoretic**

### **Reconstruction of Hobbes's Theory of the Good**

The purpose of this chapter is to give a formal representation of Hobbes's theory of real good. In order for one to appreciate the significance of what I am trying to do in this section, one would have to have a general understanding of contemporary decision theory, which is known as *utility theory*. I have tried to write the earlier parts of each subsection to serve as a rather non-technical introduction to modern utility theory for those who are uninitiated to the field. I hope that this section will turn out to be helpful for the uninitiated. Furthermore, I feel that there is a general tendency to misunderstand the major purpose of modern utility theory as well as its major claims in the philosophical community. I will try to correct these misunderstandings as best as I can as I move on.

## 5.1 Measurement Theory and the Assignment of Numbers

### 5.1.1 A General Introduction

Modern utility theory can be seen as a sub-field of what is now known as *measurement theory*.<sup>1</sup> In our everyday lives, we try to measure various things by assigning numbers to those things. We assign numbers and measure people's height to know which person is taller (or shorter) than another person. We assign numbers and measure temperature to know which object is hotter (or colder) than another object. The numbers that are assigned are meant to represent a certain relation that holds among the things that are being measured. So, height (which is a number) represents the relation "... is taller than ..." in the sense that  $a$  is taller than  $b$  if and only if the height of  $a$  is greater than the height of  $b$ . Temperature represents the relation "... is hotter than ..." in a similar way.

Measurement theory is concerned with the foundational issues of such measuring practice; it focuses on when and how a given type of measurement can take place for a given set of objects, and what kind of manipulations of the assigned numbers can be justifiably performed without rendering a statement totally meaningless. For example, if a toddler is 100 cm tall and Joe is 180 cm tall, then we can justifiably say that Joe is 1.8 times taller than the toddler. However, if location A is 10°C and location B is 18°C, then we cannot justifiably say that location B is 1.8 times hotter than location A.

As we have seen, a given relation can be represented by more than one scales. Length can be represented in millimeters, centimeters, meters, inches, foot, etc. Temperature can generally be represented by the Fahrenheit scale (°F) or the Celsius (centigrade) scale (°C). The reason why it is meaningful to say that Joe (who is 180 cm tall) is 1.8 times taller than the toddler (who is 100 cm tall) is because the fact that the number assigned to Joe within a given scale of length is 1.8 times greater than the number assigned to the toddler remains

---

<sup>1</sup>See Krantz et al. [1971], Roberts [1979] Everything written in this subsection is based on these two books and Professor Blume's lecture notes on "Ordinal Representations."

unchanged regardless of the particular scale we use to measure Joe and the toddler's height. For example, if we use the meter scale, Joe's height (which is 1.8 M tall) is still 1.8 times greater than the toddler's height (which is 1 M), and if we use the inch scale, Joe's height (which is roughly 70.9 inches) is still 1.8 greater than the toddler's height (which is roughly 39.4 inches.) We can easily see that this does not hold in the case of temperature. Specifically, the number 64.4, (which represents the same temperature as 18 °C in the Fahrenheit scale), is not 1.8 times greater than the number 50 (which represents the same temperature as 10°C in the Fahrenheit scale.) Therefore, saying that the temperature of a particular location is  $x$  times greater than the temperature of another location is just plainly meaningless. This shows that we have to be extremely careful in interpreting what the numbers signify when we are dealing with a particular area of measurement.

As I have briefly stated above, measurement, in general, is concerned with assigning numbers to objects in a way that *preserves, corresponds, and represents* the relations that the objects bear to one another. In mathematical terms, measurement is about finding *homomorphisms* from a given empirical relational system that we wish to investigate to a certain numerical relational system. A relational system is a  $n + 1$  tuple of the form  $(X, R_1, \dots, R_n)$  where  $X$  denotes a set of objects and  $R_i$  denotes a relation that bears on  $X$ . For instance, when we are dealing with measuring temperature of different locations, the empirical relational system that we are interested in would be the 2-tuple  $(L (= \text{the set of locations}), H (= \text{the binary relation "...is hotter than..."}))$  An example of a numerical relational system would be:  $(\mathbb{R} (= \text{the set of real numbers}), >, \geq, +, \times)$

Formally, A homomorphism of one relational system to another relational system is a mapping  $f$  from one relational system *into* another relational system which preserves all the relations of the former.

Suppose  $\alpha$  and  $\beta$  are two relational systems such that  $\alpha = (A, R_1, R_2, \dots, R_n)$  and  $\beta = (B, R_1^\#, R_2^\#, \dots, R_n^\#)$ . Then, a function  $f : A \rightarrow B$  is a homomorphism from  $\alpha$  *into*  $\beta$ , if and

only if, for all  $a_1, a_2, \dots, a_{r_i} \in A$ ,

$$R_i(a_1, a_2, \dots, a_{r_i}) \iff R_i^\# [f(a_1), f(a_2), \dots, f(a_{r_i})], i = 1, 2, \dots, n$$

2

Here, we can see that the mapping  $f$  preserves the relational structure of the former set in the sense that any elements that bore a certain relation to one another in the former set will be mapped to elements in the latter set which bear a corresponding relation to one another in a similar way. When the relational system  $\beta$  is a numerical relational system, then we might roughly think of  $f$  as a *scale* that measures the non-numerical objects in the former set. The function  $f$  does not need to be *onto* or *one-to-one* in order for it to qualify as a homomorphism. However, when a homomorphism is *one-to-one* is, then it is called a *isomorphism*.

For instance, in the case of measuring temperature of different locations, we can say that we are seeking a homomorphism  $t$  (temperature) from the empirical relational system ( $L$  (= the set of locations),  $H$  (= the binary relation “...is hotter than...”)) into the numerical relational system ( $\mathbb{R}$  (= the set of real numbers),  $>$ ). So, the function  $t$  (temperature) represents the empirical relation  $H$  “...is hotter than...” in the sense that, for all  $a, b \in L$ ,  $aHb$  ( $a$  is hotter than  $b$ ), if and only if,  $t(a) > t(b)$ .

A major part of measurement theory is to specify sufficient (or, more ideally, necessary and sufficient) conditions for such homomorphism from a given empirical, non-numerical relational system to a given numerical relational system to exist. (In most cases, the numerical relational system into which we map the empirical and non-numerical relational system will be the real number system.) The problem of finding sufficient conditions for there to be a homomorphism is called *the representation problem*.

The major theorem that states that a certain set of conditions (i.e. axioms) are sufficient

---

<sup>2</sup>See Roberts [1979, p. 52]

for there to be such homomorphism is called a *representation theorem*. When we are dealing with the problem of measuring people's preferences, the representation theorem is usually stated in terms of the existence of a *utility function*. We will get back to this in the following subsections.

Now, the main reason why we assign numbers to nonnumerical objects in the first place is to grasp the qualitative relational structure of the objects in a more convenient and parsimonious way. Suppose that  $L$  is the set that consists of five different locations:  $a, b, c, d$ , and  $e$ . Suppose that we want to know which is the second hottest location among the five. Without using a numerical scale, such a temperature, we might have to compare every possible pair among the five to determine how each location bears the relation "... is hotter than..." to another. This is very inefficient. However, suppose we are given each location's temperature, say,  $t(a) = 30(^{\circ}C)$ ,  $t(b) = 13(^{\circ}C)$ ,  $t(c) = 15(^{\circ}C)$ ,  $t(d) = 7(^{\circ}C)$ , and  $t(e) = 25(^{\circ}C)$ . Then, by comparing these numbers, we can see right away that location  $e$  is the second hottest location among the five.

The reason why assigning numbers to represent a certain relational system is so convenient is due to the fact that we are very much familiar with using and manipulating numbers by many years of habit. By many years of habit, we know intuitively, almost by second nature, that the set of real numbers (i.e.  $\mathbb{R}$ ) is ordered by the binary relation  $>$  (i.e. "...is greater than..."), which is both asymmetric (i.e.  $x > y$  imply  $y \not> x$ ) and negatively transitive (i.e.  $x \not> y$  and  $y \not> z$  imply  $x \not> z$ .) Even if one has never heard of the terms "asymmetry" and "negative transitivity" before, one is still very much aware of the implications of  $>$  and how it behaves; for instance, one knows that  $30 > 25$  imply  $25 \not> 30$  almost instantly. So, when we see the temperatures of the five different locations, we can list them in the order of  $t(a) = 30(^{\circ}C) > t(e) = 25(^{\circ}C) > t(c) = 15(^{\circ}C) > t(b) = 13(^{\circ}C) > t(d) = 7(^{\circ}C)$  and from this we are able to see that  $e$  is the second hottest location among the five.

However, this familiarity of numbers is a double-edged sword. Since we are so familiar with certain kinds of mathematical operations such as addition (+) and multiplication ( $\times$ ), it

is very easy for us to misapply these operations to the numerical representations even when there is no corresponding operation in the empirical, nonnumerical relational system that we are seeking to measure. In other words, it is very easy for us to read into the numbers too much and mistakenly think that certain properties that only hold for the real number system also hold for the empirical nonnumerical relational system which are merely represented numerically.

For instance, temperature in our current example is a homomorphism from the nonnumerical relational system  $(L, H)$  into the numerical relational system  $(\mathbb{R}, >)$ . Here, we can see that the original nonnumerical relational system does not have any concatenation operations that behave similarly to the mathematical operation of addition or multiplication. Therefore, we should not say such statements as “the sum of locations  $b, c, d, e$  is twice as hot as location  $a$ .” Although it is true that “ $13 + 15 + 7 + 25 = 2 \times 30$ ”, such mathematical operation does not reveal any significant empirical structure of the nonnumerical system  $(L, H)$ , since there are simply no nonnumerical operations of this nonnumerical relational system that corresponds and behaves like the mathematical operation of addition and multiplication. We can state this fact alternatively by saying that although there exists homomorphisms (namely, different scales of temperature) from  $(L, H)$  into  $(\mathbb{R}, >)$ , there exists no homomorphisms from  $(\mathbb{R}, >, +, \times)$  into  $(L, H)$ .

As we have seen, in many cases, a given nonnumerical relational system can be represented by more than one numerical scale. We all know that temperature can be measured either from the Fahrenheit or the Celsius scale.<sup>3</sup> Both the Fahrenheit scale and the Celsius scale are perfectly adequate representations of the “...is hotter than...” relation of nonnumerical objects. And it is interesting to see how these two different scales are related to each other. As it is well known, we can convert any given temperature written in the Celsius scale (C) to a temperature written in the Fahrenheit scale (F) by the following formula:

---

<sup>3</sup>Note that absolute temperature is measured by the Kelvin scale, which behaves very differently from the Fahrenheit and Celsius scale.

$F = \frac{9}{5}C + 32$ . We can see that this is a specific instance of a mathematical transformation of the form:  $\phi(x) = \alpha x + \beta$  where  $\alpha > 0, \beta \in \mathbb{R}$ . Such transformation  $\Phi$  is generally known as a “positive affine (or linear) transformation.”

In the case of temperature, any scale that is a positive affine transformation of some other temperature scale is another legitimate temperature scale. The newly obtained scale is *legitimate* in the sense that any statement that was true (or false) of the former scale remains true (or false) of the new scale. When this holds, we say that the scale is *unique up to positive affine transformation*.

This basically means that performing a positive affine transformation of one legitimate scale will result in another legitimate scale which represents the relational features of the nonnumerical relational system just as well as the original scale. In other words, the *family of scales* that equally represent the nonnumerical relational system under investigation is picked out by performing a positive affine transformation to one scale to another.

Not all scales are unique up to positive affine transformation. Some scales are only unique up to (strictly) increasing (ordinal) transformations (i.e.  $x \geq y$  iff  $\phi(x) \geq \phi(y)$ ) and other scales are unique up to similarity transformations (i.e.  $\phi(x) = \alpha x, \alpha > 0$ ), and so on. Each type of transformation picks out a particular family of scales.

From this, it is possible to determine the *scale type* of a particular scale by looking at what kind of *admissible transformations* can be legitimately performed in a way that preserves the relational structure of the nonnumerical relational system in question.

The following table summarizes some common scale types and their admissible transformations.

Table 5.1: Some Common Scale Types

4

---

<sup>4</sup>This table is taken from Roberts [1979, p. 64] with only minor modifications.

<i>Admissible Transformations</i>	<i>Scale Type</i>	<i>Example</i>
$\phi(x) = x$ (Identity)	Absolute Scale	Counting
$\phi(x) = \alpha x, \alpha > 0$ (Similarity Transformation)	Ratio Scale	Mass, Temperature on Kelvin scale, Length, etc.
$\phi(x) = \alpha x + \beta, \alpha > 0, \beta \in \mathbb{R}$ (Positive Affine (Linear) Transformation)	Interval Scale	Temperature (Fahrenheit, Celsius), etc.
$x \geq y$ iff $\phi(x) \geq \phi(y)$ (Strictly ) Monotone Transformation (=Increasing transformation, Ordinal transformation)	Ordinal Scale	Preferences, Hardness, etc.
Any One-to-One Transformation	Nominal Scale	Number Uniforms, etc.

As we can see from the table, *temperature* is unique up to *positive affine transformation*, and, is, therefore, an *interval scale*. While *length* is unique up to what is known as *similarity transformation*, and, is, therefore, an *ratio scale*. A given scale type determines what kind of statements can be meaningfully asserted about the non-numerical relational system in question. A statement is meaningful if and only if its truth-value remains the same in all admissible transformations of a given scale type.

In a ratio scale, the *ratio* between two magnitudes remain constant among all admissible transformations. So, it is meaningful to say that the length of a certain object is  $n$  times greater than that of another object. In an interval scale, *the ratio between differences* remain constant among all admissible transformations.<sup>5</sup> So, in the case of temperature, it is meaningful to say that the degree to which  $a$  is hotter than  $b$  is  $n$  times as great as the degree to which  $c$  is hotter than  $d$ . Some scales are only preserve the *order* of the objects. We might think of the number of stars that a particular movie receives from any of one's favorite movie review

---

<sup>5</sup>Here is a proof. Suppose that  $a, b, c$ , and  $d$  are four values in an interval scale and  $|a - b| / |c - d| = n$ . Now, suppose that we perform a positive affine transformation  $\phi(x) = \alpha x + \beta, \alpha > 0, \beta \in \mathbb{R}$  for these four values. Then,  $\phi(a) = \alpha a + \beta, \phi(b) = \alpha b + \beta, \phi(c) = \alpha c + \beta, \phi(d) = \alpha d + \beta$ .  $|\phi(a) - \phi(b)| / |\phi(c) - \phi(d)| = |(\alpha a + \beta) - (\alpha b + \beta)| / |(\alpha c + \beta) - (\alpha d + \beta)| = \alpha |a - b| / \alpha |c - d| = |a - b| / |c - d| = n$ . Therefore, the ratio ( $n$ ) between the differences ( $(a - b)$  and  $(c - d)$ ) are preserved in an interval scale.

site is an ordinal scale. A movie that receives 4 out of 5 stars might be better than a movie that receives 1 out of 5 stars. However, it is meaningless to say that a 4 star movies is 4 times greater than a 1 star movie or that the degree to which a 3 star movie is greater than a 1 star movie is twice as the degree to which a 5 star movies is greater than a 4 star movie. Suppose that we perform a strictly increasing ordinal transformation  $\phi(x) = x^3$  to all of the values (i.e. stars) that each movies receives. Then, the order of the movies remain unchained. However, we can see that none of the statements concerning the ratios or the ratio of the differences between the movies remain true after such transformation.

A scale that is unique up to (strictly) increasing transformation is called an *ordinal scale*, while interval, ratio, and absolute scales are instances of what are known as *cardinal scales*.

### **5.1.2 Why Bother with Utility Theory in Interpreting Hobbes?**

So, this is the basics of measurement theory. What does all of this have to do with Hobbes? I have said that the purpose of this section is to provide a formal representation of Hobbes's theory of real good. Why provide such a formal representation?

To begin with, I would like to point out that it was Hobbes *himself* who already had a rather quantified, albeit under-developed, notion of preferences and the good. Consider how Hobbes explains the process of deliberation:

When in the mind of man appetites and aversions, hopes and fears, concerning one and the same thing arise alternately, . . . , so that sometimes we have an appetite to it, sometimes an aversion from it, sometimes hope to be able to do it, sometimes despair or fear to attempt it, *the whole sum* of desires, aversions, hopes and fears continued till the—6884 be either done or thought impossible, is that we call DELIBERATION. [*Hobbes 1994: Leviathan*, Chapter VI, Paragraph 49, emphasis added]

In deliberation, the last appetite or aversion immediately adhering to action, or

to the omission thereof, is that we call the WILL, the act (not the faculty) of *willing*. ... *Will* therefore is the *last appetite in deliberating*. [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 53]

Here, we can see that Hobbes thinks that one of the essential features of deliberation is to somehow *sum up* the various desires and aversions that one feels towards particular consequences to determine on the whole what one most prefers to act. The fact that Hobbes thinks that desires and aversions can generally be summed up to determine (what he calls) one's "last appetite" indicates that Hobbes thought that it is, in principle, possible to *measure* the strength and degrees of various desires and aversions, and *aggregate* these varying degrees of desires and aversions into a whole. The notion of aggregation is more apparent in the following passage:

And because in deliberation the appetites and aversions are raised by foresight of the good and evil consequences and sequels of the action whereof we deliberate, the good or evil effect thereof dependeth on the foresight of a long chain of consequences, of which very seldom any man is able to see to the end. But for so far as a man seeth, if the good in those consequences be greater than the evil, the whole chain is that which writers call *apparent* or *seeming good*. And contrarily, when the evil exceedeth the good, the whole is *apparent* or *seeming evil*; so that he who hath by *experience* or *reason* the greatest and surest prospect of consequences *deliberates best* himself, and is able, when he will, to give the best counsel unto others. [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 57 emphasis added]

Remember that one's *apparent good* was interpreted as the satisfaction of what one just happens to prefer at a given moment; as we have seen, getting what one happens to prefer at a given moment may not always be *really good* for one.

In any case, we can see that Hobbes was implicitly utilizing the notion of *expectation*

in order to explain people's preferences. That is, according to Hobbes, one prefers act  $a$  to act  $b$  if and only if the expectation of act  $a$  is greater than the expectation of act  $b$ , where the expectation of an act is calculated by aggregating all of the good and bad consequences that the act is expected to generate and weighing them up according to each consequence's respective probabilities. This already assumes that the goodness and badness of different types of consequences is *measurable*, and that the overall goodness of the whole is somehow determined by the goodness of the individual parts that constitute it.

Furthermore, the fact that Hobbes thinks that one's estimations of likelihoods can be improved by more experience and better reasoning powers indicate that his notion of probabilities were somewhat *objective*. That is, he assumes that one's estimations of probabilities can get closer and closer to the *truth*.

The general intuition behind all of this, I believe, is remarkably similar to the (objective) expected utility theory developed by John von-Neumann and Oskar Morgenstern in [Von Neumann and Morgenstern, 1944]. And I believe that it would be quite interesting to see how Hobbes's general intuition concerning the quantification of desires, aversions, and goodness can be formalized in the light of contemporary utility theory.

## **5.2 An Ordinal Representation of Hobbes's Theory of Real Good<sup>6</sup>**

### **5.2.1 An Ordinal Representation for $Bob_i$**

We are aiming to represent Hobbes's theory of real good. According to T2 (from section 4.2), something is really good for a given individual if and only if that something satisfies the preferences formed by the individual's idealized-self on behalf of the individual's actual-self.

---

<sup>6</sup>The general flow of this subsection follows that of Kreps [1988, chapters 2, 3]

Let us take an arbitrary individual: Bob. Now, we would need to distinguish three different types of Bob: (a) The *idealized-Bob* (i.e.  $Bob_+$ ) who is substantially rational in the sense defined in P1 and whose preferences for Bob are always fully-considered and fully-balanced, (b) The *actual non-idealized Bob* (i.e.  $Bob_S$ ) who is primarily under the influence of a *basic passion for self-preservation*, and (c) The *actual non-idealized Bob* (i.e.  $Bob_G$ ) who is primarily under the influence of a *basic passion for glory*.

For simplicity, let the set of consequences be:  $C = \{\text{Death, Mortified Life, Moderate Life, Glorified Life}\}$  All three types of Bobs have strict preferences (denoted by the binary relation  $\succ_i$ ,  $i = Bob_+, Bob_S, Bob_G$ ) on  $C$ .  $a \succ_{Bob_S} b$  will be interpreted as expressing the proposition that the actual non-idealized Bob, who is primarily under the influence of a basic passion for self-preservation, *strictly prefers* option  $a$  to option  $b$ ; using Hobbes's own terminology,  $a \succ_{Bob_S} b$  will mean that when given a choice between  $a$  and  $b$ ,  $Bob_S$  *wills*  $a$ . This interpretation applies to the other two types of Bobs as well.

Without giving any interpretation,  $\succ$  is simply a binary relation. Now, we should ask: what sort of properties should the binary relation  $\succ$  satisfy in order for it to be properly regarded as a strict preference relation? A rather obvious thing that comes to mind is that  $\succ$  should be *asymmetric*. A binary relation  $R$  is asymmetric if and only if  $xRy$  imply *not*  $yRx$ . This is a reasonable property that one would expect to hold for a strict preference relation  $\succ$ . Suppose that any given type of Bob strictly prefers Moderate Life to Death. Then, it seems very natural to think that this implies that that type of Bob does *not* strictly prefer Death to Moderate Life.

This, by itself, does not say that there is something intrinsically wrong about preferring Death to Moderate Life in itself. (Although Hobbes does think that there is something intrinsically wrong about not desiring self-preservation.) What it says is that given that one strictly prefers Moderate Life to Death it would be inconsistent to strictly prefer Death to Moderate Life as well. If Bob claimed that he strictly prefers Moderate Life to Death and also claimed that he strictly prefers Death to Moderate Life, we would seriously be suspicious of

his sincerity. So, here is our starting axiom for  $\succ$ .

**[AXIOM(A-2-1)]:** The binary relations  $\succ_{Bob+}$ ,  $\succ_{Bob_S}$ , and  $\succ_{Bob_G}$  are *asymmetric*.

Another property that we might require of the binary relation  $\succ_i$  ( $i = Bob+, Bob_S, Bob_G$ ), in order for it to be properly regarded as a strict preference relation, is *negative transitivity*. If the binary relation  $\succ$  is negatively transitive, then  $a \not\succeq b$ <sup>7</sup> and  $b \not\succeq c$  imply  $a \not\succeq c$ .

Empirically speaking, there are some situations where negative transitivity does not seem to hold for strict preferences. This is, in many cases, due to the fact that our perceptual abilities are not fine enough to detect small differences, but are able to discriminate when these small differences accumulate. For example, one might not have a strict preference between one spoon of sugar to two spoons of sugar in one's cup of coffee. Furthermore, one might not have a strict preference between two spoons of sugar to three spoons of sugar. However, it might very well be true that one does have a strict preference between one spoon of sugar and three spoons of sugar in one's cup of coffee. If this is the case, one's preferences do not obey negative transitivity.

Negative transitivity can also fail when the objects have *multiple attributes* that influence one's preferences separately. Suppose that the two things that one considers when choosing an automobile are performance and fuel economy. Suppose that there are three cars: A=(moderate performance, high fuel economy), B=(high performance, low fuel economy), and C=(moderate performance, moderate fuel economy). Then, one might not prefer A to B and one might not prefer B to C. But, one might very well prefer A to C. Again, this violates negative transitivity.

However, despite all of its problems, assuming that the strict preference relation  $\succ$  is negatively transitive seems to be unproblematic at least in our current framework. There are only four elements in the set of consequences C: namely, Death, Mortified Life, Moderate Life, and Glorious Life. And there does not seem to be any issues of epistemic indiscrimina-

---

<sup>7</sup>Read as: it is *not* the case that  $a$  is strictly preferred to  $b$ .

cies that might add up to cause problems for negative transitivity, nor does there seem to be any issues concerning multiple attributes that might cause incomparability issues that violate negative transitivity. So, I believe that it is safe to state negative transitivity as a property that our strict preference relations satisfy as one of our axioms.

**[AXIOM(A-2-2)]:** The binary relations  $\succ_{Bob+}$ ,  $\succ_{Bob_S}$ , and  $\succ_{Bob_G}$  are *negatively transitive*.

The fact that our strict preference relations are both asymmetric and negatively transitive implies that they are also *transitive* (i.e.  $a \succ b$  and  $b \succ c$  imply  $a \succ c$ ).

**[LEMMA(L-2-1)]:** The binary relations  $\succ_{Bob+}$ ,  $\succ_{Bob_S}$ , and  $\succ_{Bob_G}$  are *transitive*.<sup>8</sup>

We now define the following two relations in terms of  $\succ_i$ .

**[WEAK PREFERENCE( $\succsim_i$ ) (D-2-1)]:**  $a \succsim_i b \equiv_{df} b \not\succeq_i a$

and

**[INDIFFERENCE( $\sim_i$ ) (D-2-2)]:**  $a \sim_i b \equiv_{df} a \not\succeq_i b$  and  $b \not\succeq_i a$

from this, we can prove the following properties of these two induced relations.

**[LEMMA(L-2-2)]:** For  $i = Bob+, Bob_S, Bob_G$ ,

- (a)  $\succsim_i$  is *complete* (i.e. for all  $x, y$ , either  $x \succsim_i y$  or  $y \succsim_i x$ ) and *transitive*,
- (b)  $\sim_i$  is an *equivalence relation*; that is,  $\sim_i$  is *reflexive* (i.e. for all  $x$ ,  $x \sim_i x$ ), *symmetric* (i.e. for all  $x, y$ ,  $x \sim_i y$  implies  $y \sim_i x$ ), and *transitive*.

---

<sup>8</sup>**Proof.** Suppose that  $a \succ_i b$  and  $b \succ_i c$  ( $i = Bob+, Bob_S, Bob_G$ ). Since  $a \succ_i b$ , negative transitivity implies that either  $a \succ_i c$  or  $c \succ_i b$ . (According to negative transitivity,  $(x \not\succeq y \text{ and } y \not\succeq z) \Rightarrow x \not\succeq z$ . This is logically equivalent to  $x \succ z \Rightarrow \text{either } x \succ y \text{ or } y \succ z$ .) However,  $c \succ_i b$  cannot be the case since we've already assumed  $b \succ_i c$  and, by asymmetry, this implies  $c \not\succeq_i b$ . Therefore,  $a \succ_i c$  ( $i = Bob+, Bob_S, Bob_G$ ). This completes the proof.  $\square$

$$(c) \quad [(x \succ_i y) \wedge (y \sim_i z) \Rightarrow x \succ_i z] \text{ and } [(x \sim_i y) \wedge (y \succ_i z) \Rightarrow x \succ_i z]^9$$

Note that we have interpreted the absence of strict preference in either direction as *indifference*. However, such interpretation might be problematic in some cases when the absence of strict preference in either direction does not imply real indifference but rather *incomparability*.

This again can be illustrated by the automobile example that we have seen when we were discussing about some possible problems for negative transitivity. Take again the three cars: A=(moderate performance, high fuel economy), B=(high performance, low fuel economy), and C=(moderate performance, moderate fuel economy). Suppose that one strictly prefers one car over another if and only if one of the cars is at least as good as the other car in both performance and fuel economy and that one of the cars is strictly better than the other car in at least one aspect. If so, then we can see that the two cars A and B are incomparable to each other, and the two cars B and C are incomparable to each other. So,  $A \not\succeq_i B$  and  $B \not\succeq_i A$ , which implies  $A \sim_i B$  according to our definition, and  $B \not\succeq_i C$  and  $C \not\succeq_i B$ , which implies  $B \sim_i C$  according to our definition. However, it turns out  $A \succ_i C$ , which violates transitivity of  $\sim_i$ .

What this shows is that, unlike indifference, incomparability is not transitive. So, if there is a possibility that the absence of strict preference in either direction would imply, not indifference, but incomparability, our definition of  $\sim_i$ , as well as the fact that it is an

---

<sup>9</sup>**Proof.**

- (a) By asymmetry, for all  $x$  and  $y$ , either  $x \not\succeq_i y$  or  $y \not\succeq_i x$ . By definition of  $\succsim_i$ , this means that for all  $x$  and  $y$ , either  $x \succsim_i y$  or  $y \succsim_i x$ . Therefore,  $\succsim_i$  is complete. By negative transitivity, for all  $x, y$ , and  $z$ ,  $x \not\succeq_i y$  and  $y \not\succeq_i z$  imply  $x \not\succeq_i z$ . By definition of  $\succsim_i$ , this means that  $x \succsim_i y$  and  $y \succsim_i z$  imply  $x \succsim_i z$ . Therefore,  $\succsim_i$  is transitive.  $\square$
- (b) By asymmetry of  $\succ_i$ , for all  $x$ ,  $x \not\succeq_i x$ . And this implies  $x \sim_i x$ . Therefore,  $\sim_i$  is reflexive. Now, suppose  $x \sim_i y$ . Then,  $x \not\succeq_i y$  and  $y \not\succeq_i x$ . By definition of  $\sim_i$ , this implies  $y \sim_i x$ . Therefore,  $\sim_i$  is symmetric. Now, suppose  $x \sim_i y$  and  $y \sim_i z$ . Then, by definition of  $\sim_i$ , this implies  $x \not\succeq_i y$  and  $y \not\succeq_i x$  and  $y \not\succeq_i z$  and  $z \not\succeq_i y$ . By negative transitivity of  $\succ_i$ , this implies  $x \not\succeq_i z$  and  $z \not\succeq_i x$ , which means  $x \sim_i z$ . Therefore,  $\sim_i$  is transitive. Since,  $\sim_i$  is reflexive, symmetric, and transitive, it is an *equivalence relation*.  $\square$
- (c) By asymmetry of  $\succ_i$ , for all  $x, z$ , there are only three possibilities:  $x \succ_i z$  or  $z \succ_i x$  or  $x \sim_i z$  (i.e.  $x \not\succeq_i z$  and  $z \not\succeq_i x$ ). According to the definition of  $\sim_i$ ,  $y \sim_i z$  if and only if  $y \not\succeq_i z$  and  $z \not\succeq_i y$ . By asymmetry,  $x \succ_i y$  implies  $y \not\succeq_i x$ . Therefore, by negative transitivity,  $z \not\succeq_i x$ . So, either  $x \succ_i z$  or  $x \sim_i z$  has to be the case. Suppose  $x \sim_i z$ . Then, by symmetry and transitivity of  $\sim_i$ ,  $x \sim_i y$ , which contradicts our assumption that  $x \succ_i y$ . Therefore,  $x \succ_i z$ . The other part can be proved in a similar way.  $\square$

equivalence relation (which implies transitivity) might not be justifiable.

However, again, the issue does not arise in our current model where the set of consequences  $C$  has only four elements which are distinct enough to reasonably expect that there would be no cases of incomparability issues. The only possible problem that might occur is when we try to determine  $Bob_G$ 's preferences over the two lives – Mortified Life and Death.

Between the two lives, which would  $Bob_G$ , who is primarily under the influence of the basic passion for glory, prefer? This depends on how much  $Bob_G$  dislikes a life without power, and this, again depends on how strongly  $Bob_G$  is being influenced by the basic passion for glory. If  $Bob_G$  happens to just utterly abhor a life without power, then  $Bob_G$  might strictly prefer Death to Mortified Life; if not, he would prefer Mortified Life to Death, but to a much lesser extent to which  $Bob_+$  and  $Bob_S$  would prefer Mortified Life to Death. The main point is that regardless of the direction of preference, the two lives, Mortified Life and Death, would be comparable for  $Bob_G$ . So, the problem of incomparability does not arise in our current model.

This is enough for us to state our first representation theorem for there to exist an ordinal representation.

**[THEOREM(T-2-1)]:**  $\succ_i$  ( $i = Bob_+, Bob_S, Bob_G$ ) is a strict preference relation if and only if there exists a function  $U_i : C \rightarrow \mathbb{R}$  such that for all  $x, y \in C$ ,  $x \succ_i y$  if and only if  $U_i(x) > U_i(y)$ . Furthermore,  $u_i$  is unique up to strictly increasing transformation.

The general strategy for proving a representation theorem is to provide an example of a function that does the job; by doing so, we have shown that at least one such function exists.

Here is one such example. Define  $U_i(x) \equiv_{df} \#\{y \mid x \succ_i y\}$  That is, let  $U_i(x)$  be the *cardinality* (i.e. the number of elements) of the set of elements that  $x$  is strictly preferred to. Note that the function is well-defined - that is, for every value  $x$ ,  $U_i(x)$  exists. The function works not only in our current model, but also when  $C$  is an arbitrarily *finite* set as well.<sup>10</sup>

---

<sup>10</sup>The theorem holds true for sets that are denumerable (i.e. countably infinite) as well. For uncountably

Before we move on, I think that it would be helpful to understand the general intuition behind the entire process. When the strict preference relation (i.e.  $\succ$ ) defined on the set of consequences is asymmetric and negatively transitive, we can see that it behaves in the same way as the “...is the greater than...” relation (i.e.  $>$ ) defined on the set of real numbers. Moreover, we can see that the weak preference relation (i.e.  $\succeq$ ), which is complete and transitive, behaves in the same way as the “...is the greater than or equal to...” relation (i.e.  $\geq$ ), and the indifference relation (i.e.  $\sim$ ), which is an equivalence relation, behaves in the same ways as the “... is equal to...” relation (i.e.  $=$ )

This means that we can find a homomorphism from the nonnumerical relational system  $(C, \succ, \succeq, \sim)$  to the numerical relational system  $(\mathbb{R}, >, \geq, =)$  by mapping each consequence to a real number, the relation  $\succ$  to  $>$ , the relation  $\succeq$  to  $\geq$ , and the relation  $\sim$  to  $=$ . And this is exactly what the utility function  $U_i$  is in effect doing.

I believe that this can be more intuitively understood graphically. Consider the following diagram that summarizes the representation part of Theorem T-2-1.

---

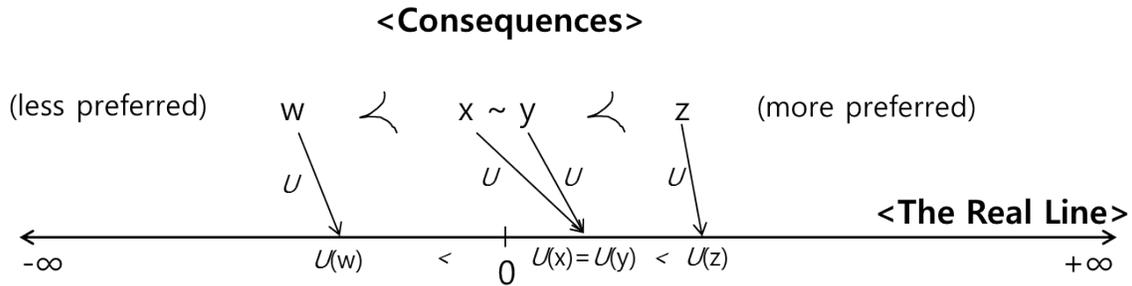
infinite sets, we would have to require that the set of consequences to have a countable order-dense subset in addition to  $\succ$  being asymmetric and negatively transitive in order for there to be an ordinal representation to exist.

<sup>11</sup>**Proof.** (The proof is based on the proof contained in Larry Blume’s lecture notes “Ordinal Representations.”) Suppose  $x \succ_i y$ . Then, by the transitivity of  $\succ_i$  (Lemma L-2-1), for all  $z$  such that  $y \succ_i z$ ,  $x \succ_i z$ . Therefore,  $z \in \{w \mid x \succ_i w\}$  and  $z \in \{w \mid y \succ_i w\}$  which implies  $\#\{w \mid x \succ_i w\} \geq \#\{w \mid y \succ_i w\}$ . Now, there is at least one element that is in  $\{w \mid x \succ_i w\}$  which is *not* in  $\{w \mid y \succ_i w\}$ ; namely,  $x$ . Therefore,  $\#\{w \mid x \succ_i w\} > \#\{w \mid y \succ_i w\}$  which implies  $U_i(x) > U_i(y)$ .

Now, suppose  $U_i(x) > U_i(y)$ . By asymmetry of  $\succ_i$ , for all  $x, y$ , there are only three possibilities:  $x \succ_i y$  or  $y \succ_i x$  or  $x \sim_i y$  (i.e.  $x \not\succ_i y$  and  $y \not\succ_i x$ ).  $y \succ_i x$  cannot be the case, since  $y \succ_i x$  implies  $U_i(y) > U_i(x)$  which contradicts our assumption. Suppose  $x \sim_i y$ . Then, by Lemma L-2-2c (and the symmetry of  $\sim_i$ ),  $y \succ_i z$  if and only if  $x \succ_i z$ , which implies that  $z \in \{w \mid y \succ_i w\}$  if and only if  $z \in \{w \mid x \succ_i w\}$ . Therefore,  $\#\{w \mid x \succ_i w\} = \#\{w \mid y \succ_i w\}$ , which implies  $U_i(x) = U_i(y)$ . This contradicts our assumption that  $U_i(x) > U_i(y)$ . Therefore, if  $U_i(x) > U_i(y)$ , then  $x \succ_i y$ .

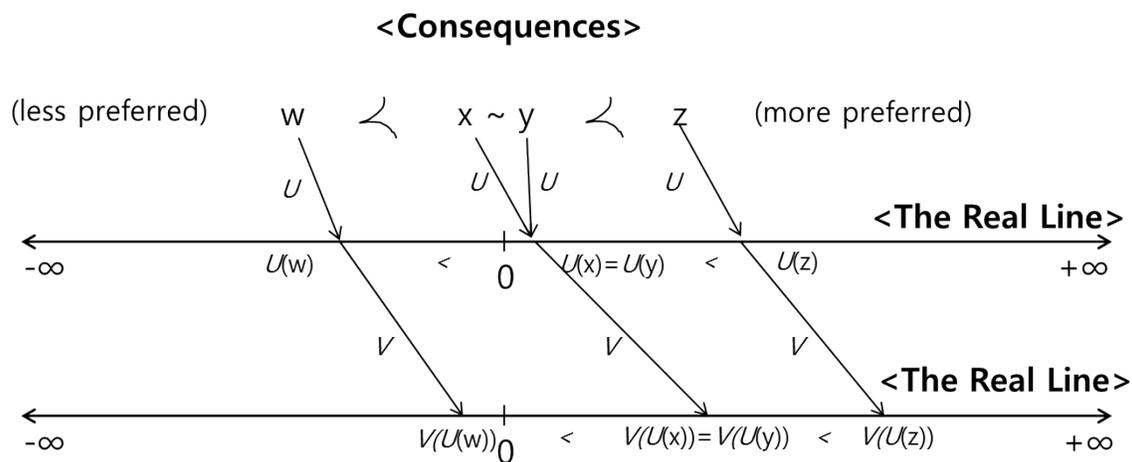
Now, for the uniqueness part of the theorem. Suppose that  $F : \mathbb{R} \rightarrow \mathbb{R}$  is a strictly increasing function. Then,  $F(U_i(x)) > F(U_i(y))$  if and only if  $U_i(x) > U_i(y)$  if and only if  $x \succ_i y$ . Therefore, if  $U_i$  is a function that represents  $\succ_i$  on  $C$ , then any strictly increasing function  $F$  of  $U_i$  also represents  $\succ_i$  on  $C$ . That is,  $U_i$  is unique up to strictly increasing transformation.  $\square$

**FIGURE 5.1: HOW FUNCTION  $U$  REPRESENTS ONE'S PREFERENCE-ORDERING ON THE REAL LINE**



We can think of the utility function  $U_i$  as a “thermometer” (in this case, an *ordinal* thermometer) that measures each type of Bob’s preferences, just like an ordinary thermometer measures temperature. And just as there are more than one thermometers that can be used to measure temperature, we can expect that there will be more than one utility functions that can be used to measure each type of Bob’s preferences. It happens to be that the difference utility functions that can be used to measure each Bob’s preferences, in our current framework, are picked out by performing a (strictly) increasing transformation of any given utility function of Bob<sub>*i*</sub>. This is what is claimed by the uniqueness part of Theorem T2-1, which claims that the utility function  $U_i$  is unique up to strictly increasing transformation. This, again, can be represented graphically as below.

**FIGURE 5.2: FUNCTION  $V$  AS A STRICTLY INCREASING TRANSFORMATION OF FUNCTION  $U$**



Here,  $V$  is a strictly increasing transformation of  $U_i$ . And we can see that  $V \circ U_i$  is another utility function that represents  $Bob_i$ 's preferences in the sense that, for  $Bob_i$ ,  $x \succ_i y$  if and only if  $V(U_i(x)) > V(U_i(y))$ . The way that  $V \circ U_i$  serves as another utility function is by retaining the order of the values of  $U_i$ , which, in turn, retains the order of the consequences ordered by  $Bob_i$ 's preferences.

We have just shown the most general result that whenever  $Bob_+$ ,  $Bob_S$ , and  $Bob_G$ 's strict preference relations are asymmetric and negatively transitive, (both of which, as we have seen, are quite plausible assumptions to make of a strict preference relation), there exist utility functions  $U_i : C \rightarrow \mathbb{R}$  that represent each  $Bob_i$ 's preferences over the consequences. This does not specify what the actual order of the consequences that represent each type of  $Bob_i$ 's preferences would be; it only claims that as long as each type of  $Bob_i$ 's preferences are asymmetric and negatively transitive there are utility functions to represent them. However, for us, it is possible to infer from Hobbes's text what may be considered to be the actual preference-ordering of each  $Bob_i$ .

I think that it is quite apparent what each type of  $Bob_i$ 's preference concerning the two

consequences, Death and Moderate Life, would be: For all types of  $Bob_i$ ,  $Death \prec_{Bob_i} Moderate Life$ . It also seems obvious that  $Death \prec_{Bob_S, Bob+} Mortified Life \prec_{Bob_S, Bob+} Moderate Life$  and  $Death \prec_{Bob_G} Moderate Life \prec_{Bob_G} Glorified Life$ . This follows from our stipulation that  $Bob_S$  and  $Bob+$  are the two types of Bob who are primarily concerned with securing the actual-Bob's long-term self-preservation (which would make them think that even a Mortified life is better than Death), and that  $Bob_G$  is the type of Bob who is primarily concerned with obtaining glory and honor (which would make him think that a Glorified Life is definitely better than a Moderate Life.)

What needs to be figured out is: (a) where Glorified Life would fit into the preference-ordering of  $Bob_S$  and  $Bob+$ , and (b) where Mortified Life would fit into the preference-ordering of  $Bob_G$ .

To answer (a), I claim that, for both  $Bob_S$  and  $Bob+$ , their preferences on  $C$  would be:  $Death \prec_{Bob_S, Bob+} Mortified Life \prec_{Bob_S, Bob+} Moderate Life \prec_{Bob_S, Bob+} Glorified Life$ . To this, one might argue: why would anybody like  $Bob_S$  or  $Bob+$ , who are assumed to be primarily under the influence of a basic desire for self-preservation, care anything about whether their lives are glorious?

To this, I believe that it would be helpful to recall what Hobbes means by "glory." According to Hobbes, glory is the pleasure that one experiences by the self-recognition of one's own superior power over other people.<sup>12</sup> This means that a Glorified Life is a life that has everything that a Moderate Life has *plus power*. Now, for Hobbes, power is defined as one's present means to satisfy one's yet unmet preferences.<sup>13</sup> Therefore, everything being equal, not only a person who is primarily motivated by a desire for glory, but also a person who is primarily motivated by a desire for self-preservation, would prefer a life that has more power to a life that has less; for having more power would imply that one has better means to secure

<sup>12</sup>"Joy arising from imagination of a man's own power and ability is that exultation of the mind which is called GLORYING..." [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 39]

<sup>13</sup>"The power of a *man* (to take it universally) is his present means to obtain some future apparent good. . ." [Hobbes 1991: *Leviathan*, Chapter X, Paragraph 1]

one's own self-preservation. Hobbes explains the reason why even people who are primarily motivated by his or her own self-preservation would prefer a Glorified Life (i.e. a life with power) to a Moderate Life in the following way:

So that in the first place, I put for a general inclination of all man-kind, 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, but because he cannot assure the power and means to live well, which he hath present, without acquisition of more. [Hobbes 1994: *Leviathan*, Chapter XI, Paragraph 2]

This means that all three types of *Bob<sub>i</sub>s* would prefer a Glorified Life to a Moderate Life other things being equal.

Now, how should we answer question (b)? That is, how would *Bob<sub>G</sub>*, who is primarily under the influence of a basic passion for glory, feel about a Mortified Life. The major issue is whether *Bob<sub>G</sub>* would find a Mortified Life a life that is worth living or whether he would find such a life even worse than Death? I believe that the following passage from *Leviathan* is indicative of the correct answer to this question.

Vain-glorious men ... are inclined to rash engaging, and in the approach of danger or difficulty, to retire if they can; because not seeing the way of safety, they will rather hazard their honour, which may be salved with an excuse, than their lives, for which no salve is sufficient." [Hobbes 1994: *Leviathan*, Chapter XI, Paragraph 12]

What this passage is suggesting is that *Bob<sub>G</sub>* would, in general, rashly engage in risky activities in an attempt to obtain glory and honor which *Bob<sub>G</sub>* values very highly. However, when the prospects for gaining glory becomes significantly dim, and when the prospects for getting killed becomes significantly high, even a person like *Bob<sub>G</sub>*, who is primarily under

the influence of a basic passion for glory, would opt for dishonor which would, at least, keep him alive. This suggests that  $Bob_G$  would actually prefer a Mortified Life to Death.

From this, we can summarize the preferences of the three type of  $Bob_i$ s as follows:

< $Bob_i$ 's Preferences>

- [ $Bob'_+$ 's Preferences] : Death  $\prec_{Bob_+}$  Mortified Life  $\prec_{Bob_+}$  Moderate Life  $\prec_{Bob_+}$  Glorified Life
- [ $Bob'_S$ 's Preferences] : Death  $\prec_{Bob_S}$  Mortified Life  $\prec_{Bob_S}$  Moderate Life  $\prec_{Bob_S}$  Glorified Life
- [ $Bob'_G$ 's Preferences] : Death  $\prec_{Bob_G}$  Mortified Life  $\prec_{Bob_G}$  Moderate Life  $\prec_{Bob_G}$  Glorified Life

We can see here that, *ordinally speaking*, the three types of  $Bob_i$ s have the same preferences on  $C$ . This is because what distinguishes between the preference of a person who is primarily motivated by glory, (such as  $Bob_G$ ), and the preference of a person who is primarily motivated by self-preservation, (such as  $Bob_S$  and  $Bob_+$ ), is *not* the *order* of the respective consequences itself, but rather, *the relative distance* of each of the consequences to one another.

For instance, suppose that we fix the distance from Death to Glorified Life for all three types of  $Bob_i$ . Then, the relative distance between Death and Mortified Life will be much shorter for  $Bob_G$  than it is for  $Bob_S$  and  $Bob_+$ , and the relative distance between Moderate Life and Glorified Life will be much longer for  $Bob_G$  than it is for  $Bob_S$  and  $Bob_+$ . What this implies is that, unlike either of the two  $Bob_S$  or  $Bob_+$ ,  $Bob_G$  would be much likelier to accept a risky gamble between a Glorified Life and Death rather than choosing an option that would firmly secure a Moderate Life. What this means will become more apparent in the next section. What is important to understand at this point is that, when there are no aspects of risk or uncertainty involved, the preferences of all three types of  $Bob_i$ s are indistinguishable.

Combining the results of Theorem T-2-1 and each type of  $Bob_i$ 's preferences, we are now able to derive the (ordinal) utilities of each consequence (i.e. type of life) for each type of  $Bob_i$ 's.

<Bob<sub>i</sub>'s (Ordinal) Utilities>

- [Bob<sub>+</sub>'s Utilities]:  $U_{Bob_+}(\text{Death}) < U_{Bob_+}(\text{Mortified Life}) < U_{Bob_+}(\text{Moderate Life}) < U_{Bob_+}(\text{Glorified Life})$
- [Bob<sub>S</sub>'s Utilities]:  $U_{Bob_S}(\text{Death}) < U_{Bob_S}(\text{Mortified Life}) < U_{Bob_S}(\text{Moderate Life}) < U_{Bob_S}(\text{Glorified Life})$
- [Bob<sub>G</sub>'s Utilities]:  $U_{Bob_G}(\text{Death}) < U_{Bob_G}(\text{Mortified Life}) < U_{Bob_G}(\text{Moderate Life}) < U_{Bob_G}(\text{Glorified Life})$

Note that at this point, there are no specific numbers assigned. Any set of real numbers that satisfy the above inequalities will be able to be deemed as the utilities of each type of life for the specific Bob<sub>i</sub> in question. And, once values are assigned, any strictly increasing transformation of the values will qualify as another legitimate assignment of utilities for each type of life.

We now state our main theorem for this subsection.

**[THEOREM (T3)] (Ordinal Representation of Hobbes's Theory of Real Good):** For all  $x, y \in C$ ,  $x$  is *really better* (or *substantially better*) than  $y$  for Bob<sub>i</sub> ( $i = S, G$ ) if and only if  $U_{Bob_+}(x) > U_{Bob_+}(y)$ .<sup>14</sup>

We are also able to derive some results about the rationality of Bob<sub>S</sub>'s and Bob<sub>G</sub>'s preferences as well. Theorem T1 states that what is really good for somebody is to satisfy the person's substantially rational preferences, where, here, "(substantially) rational preferences" mean, according to the corollary of P2, preferences that are both well-considered and well-balanced. We have seen that preferences formed by one's idealized-self on behalf of one's actualized-self is guaranteed to be substantially rational in this way. From this, we derive the following proposition.

<sup>14</sup>**Proof.** In section 7.2, we have established that something is *really good* for a given person if and only if that thing satisfies the type of preferences that is formed by the person's idealized-self on behalf of the person's actualized-self. (This is Theorem T2.) Theorem T-2-1 claims that, given that one's strict preferences are asymmetric and negatively transitive, there exists a utility function such that one prefers  $x$  to  $y$  if and only if the utility of  $x$  (i.e.  $U(x)$ ) is greater than the utility of  $y$  (i.e.  $U(y)$ ). Bob<sub>+</sub> is assumed to be the idealized-self of Bob<sub>S</sub> and Bob<sub>G</sub>. Furthermore, by Axiom A-2-1, Bob<sub>+</sub>'s strict preferences (i.e. the binary relation  $\succ_{Bob_+}$ ) are both asymmetric and negatively transitive. Therefore, there is a utility function  $U_{Bob_+}$  such that  $x \succ_{Bob_+} y$  if and only if  $U_{Bob_+}(x) > U_{Bob_+}(y)$ . Therefore, by T2, T-2-1, A-2-1, and by our assumption that Bob<sub>+</sub> is the idealized-self of Bob<sub>i</sub> ( $i = S, G$ ), for all  $x, y \in C$ ,  $x$  is *really better* than  $y$  for Bob<sub>i</sub> ( $i = S, G$ ) [if and only if  $U_{Bob_+}(x) > U_{Bob_+}(y)$ ]. □

**[PROPOSITION (P3)]:** It is substantially rational for  $Bob_i (i = S, G)$  to strictly prefer  $x$  to  $y$  if and only if  $x \succ_{Bob_+} y$  if and only if  $U_{Bob_+}(x) > U_{Bob_+}(y)$ .

From  $\langle Bob_i$ 's Preferences  $\rangle$  and  $\langle Bob_i$ 's Utilities  $\rangle$  above, we can see that all three types of  $Bob_i$ 's preferences are substantially rational for *all* pair-wise comparisons of the elements in  $C$ .

Furthermore, D2 claims that one's preferences are formally (or minimally rational) if and only if they satisfy the axioms of decision theory. I have said that the axioms of decision theory are context-specific in the sense that they depend on what sort of representation theorem we are looking for. In our current framework of ordinal representation, it was both necessary and sufficient for one's strict preference relation to be asymmetric and negatively transitive for there to be a utility function representing those preferences. Axioms A-2-1 and A-2-2 state that all three types of  $Bob_i (i = +, S, G)$ 's preferences are asymmetric and negatively transitive. Therefore, we can say that all three types of  $Bob_i (i = +, S, G)$ 's preferences are all formally (or minimally) rational as well.

### 5.2.2 Clarifying the Meaning of Utility Functions

Before I move on to the next step of our formalization, I would like to clarify one thing about utility functions. It is quite unfortunate that the term "utility" has been used, historically, in so many different ways in both the economics as well as the philosophy community alike.<sup>15</sup>

The term "utility" started with classical utilitarianism. At first, the term meant a *disposition* or a *tendency* to produce good consequences. Consider how Jeremy Bentham, the forefather of classical utilitarianism, defines "utility":

By *the principle of utility* is meant that principle which approves or disapproves of every action whatsoever, according to the tendency which it appears to have

---

<sup>15</sup>See John Broome's "Utility" contained in [Broome, 1999] for a very informative discussion about this topic.

to augment or diminish the happiness of the party whose interest is in question ... By *utility* is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered... [Bentham, “An Introduction to the Principles of Morals and Legislation” contained in [Mill and Bentham, 1987, pp. 65-66 emphasis added]]

However, throughout the history of moral philosophy, the dispositional aspect of the term “utility” gradually dropped, and now, the term is simply used to denote a type (or types) of good consequences (e.g. pleasure, happiness, etc.) that the utilitarian urges one to maximize. Utilitarianism, in contemporary moral philosophy, is understood as a form of consequentialism, which, in its most general form, claims that what is morally right is to maximize the realization of good consequences.<sup>16</sup> Consider how J.J.C. Smart defines *act-utilitarianism*:

Act-utilitarianism is the view that rightness or wrongness of an action is to be judged by *the consequences, good or bad*, of the action itself. ... According to the act-utilitarian, then, the rational way to decide what to do is to decide to perform that one of those alternative actions open to us (including the null-action, the doing of nothing) which is likely to *maximize the probable happiness or well-being* of humanity as a whole, or more accurately, or all sentient beings. [Smart, “An outline of a system of utilitarian ethics” contained in Smart and Williams 1973, p. 9, p. 42, emphasis added]

So, when a utilitarian philosopher claims that people should *maximize utility*, what he/she is claiming is that people should *maximize good consequences*. I believe that this is the

---

<sup>16</sup>Some writers restrict utilitarianism to a narrower category, which claims that what is morally right is to maximize utility (where this is supposed to denote either pleasure or happiness), and distinguish utilitarianism from a more broader category, which utilitarianism is one part of, called “consequentialism” which recognizes that there could be good consequences other than utility (i.e. pleasure or happiness.)

definition of “utility” to which philosophers in general are most accustomed; “Utility”, for philosophers, is understood in the lights of utilitarianism, and, understood in this way, the term simply denotes a kind of externally measurable good consequence which may be related to, but are largely independent of, people’s preferences.

This is not how the term “utility” is used in contemporary economic theory or formal decision theory. In contemporary economic theory or formal decision theory, utility is simply a numerical value that represents a person’s preferences in the sense that, relative to a utility function  $U$ , a person prefers  $x$  to  $y$  if and only if the utility of  $x$  is greater than the utility of  $y$  (i.e.  $U(x) > U(y)$ )”

However, when a philosopher, (who understand “utility” in the lights of utilitarianism), hears the statement of the form, “a person prefers  $x$  to  $y$  if and only if the utility of  $x$  is greater than the utility of  $y$  (i.e.  $U(x) > U(y)$ )”, uttered by a decision theorist, it is very easy for him/her to interpret the statement as saying that the reason why the person either should or would prefer  $x$  to  $y$  is *because  $x$  generates greater amounts of good consequences (in many cases, pleasure) than  $y$* . Described in this way, this makes it sound that it would have been perfectly possible for the person to prefer  $y$  to  $x$  even when the utility of  $y$  is lesser than the utility of  $x$  if the person had happened to not care at all about maximizing good consequences (especially, pleasure.)

Similarly, game theory, which is an extension of decision theory, assumes that every player is a utility-maximizer. To a moral philosopher, who understands the term “utility” in terms of utilitarianism, such assumption can be very easily read as saying that, in game theory, everybody is assumed to care only about good consequences and their maximization. Such assumption seems very problematic, since, to the moral philosopher, it seems quite evident that utility (understood as meaning good consequences) is not the only thing that people in general care about. The natural corollary of this type of reasoning is that the purpose of both ethics and political philosophy would be better served without knowing any formal decision theory or game theory. I believe that this represents part of the reasoning

processes of many moral philosophers who tend to think that formal decision theory and game theory can offer very little, if not zero, insights to questions of ethics and political philosophy.

Such objection makes sense only when “utility” is construed in the lights of utilitarianism as denoting pleasure or some narrow category of good consequences. However, when utility is construed in that way, then the objection is no longer directed towards formal decision theory or game theory.

As I have explained, in contemporary economics and decision theory, a utility function is merely a scale (i.e. a thermometer) that measures people’s preferences. When the decision theorist says that, relative to a utility function  $U$ , one prefers  $x$  to  $y$  if and only if the utility of  $x$  is greater than the utility of  $y$ , what he/she was saying was something that is akin to the statement:  $x$  is hotter than  $y$  if and only if the temperature of  $x$  is higher than the temperature of  $y$ . So, when a moral philosopher argues that it is possible for one to prefer  $y$  to  $x$  even when the utility of  $y$  is lesser than the utility of  $x$ , this, for the decision theorist, would sound something similar to:  $y$  can be hotter than  $x$  even if the temperature of  $y$  is lower than the temperature of  $x$ .

As we already know, this is impossible unless the thermometer behaves in very weird ways, which, at that point, would make the thing useless *as* a thermometer. The same thing holds for preferences. It is just simply impossible for somebody to prefer  $y$  to  $x$  when  $U(x) > U(y)$ ; if that happened, we can say that the function  $U$  is not a utility function. In decision theory, a utility function is *simply defined as* a measure of a person’s preferences; the amount of good consequences is irrelevant to the assignment of utilities. If somebody happens to prefer  $y$  to  $x$  even when the amount of good consequences of  $x$  far exceeds that of  $y$ , then this, by definition, implies  $U(y) > U(x)$ .<sup>17</sup>

Such utility measure does not always exist; and this is why proving a representation

---

<sup>17</sup>It would also imply that the maximization of good consequences is not the only thing that the person cares about.

theorem is important. But when it does exist (and if  $U$  is any such function), it is almost a tautological truth that one prefers  $x$  to  $y$  if and only if  $U(x) > U(y)$ .<sup>18</sup>

So, here is where the confusion lies. Utilitarianism, as it is understood in contemporary moral philosophy, claims that what is morally right, and, thereby, the most rational way to act is to *maximize utility* – namely, to maximize good consequences (such as, pleasure, happiness, and so on.) Decision theory, which is supposed to tell us something about rational decision-making, claims that, as long as people are rational<sup>19</sup>, they are utility maximizers – meaning that there is a utility function (representing their preferences) of which people can be seen as maximizing the numeric value when they act. Conflating the two meanings together makes it sound that decision theorists are claiming that everybody is a perfect act-utilitarian who always succeeds in maximizing good consequences without a mistake, whereas what the decision theorist was merely saying was that people can be seen as acting according to their well-behaved preferences.

In short, the term “utility” is used in completely different ways in both contemporary economic theory/formal decision theory and in contemporary moral philosophy. Unnecessary confusion inevitably arise when one inadvertently conflates the two meanings together. So, one needs to be extremely careful either when one encounters the term “utility” in other people’s writing or when one decides to use the term in one’s writing oneself.

There is one last thing that I would like to clarify about utility functions before moving on to the next subsection. I have explained above that what distinguishes between the preferences of  $Bob_G$  and the preferences of  $Bob_S$  and  $Bob_+$  are the relative distances that each consequence in  $C$  bears to one another in each type of  $Bob_i$ s preference-ordering. However,

---

<sup>18</sup>A related, but a slightly different objection might be: “Applying decision theory or game theory to ethics and political philosophy is problematic, since it assumes that there can always be utility functions that can be used to properly represent people’s preferences.” This is a better objection since it is not based on confusing the meaning of *utility*; what the objection is basically claiming is that people’s preferences seldom meet the necessary and sufficient conditions for there to exist a utility function representing them. This is an important issue for those who would like to use rational choice theory as a positive theory of human behavior and social explanation. I will say more on this on the following chapter.

<sup>19</sup>which means that their preferences meet a minimum set of conditions.

at this point, the utility function that we have derived is only an *ordinal* utility function. And with an ordinal utility function, it is impossible to measure the relative distances between two consequences in each type of *Bob<sub>i</sub>*'s preference-ordering.

With an ordinal utility function, *only the order* of the numbers matter. For instance, suppose  $U_{Bob_G}(\text{Death}) = 0$ ,  $U_{Bob_G}(\text{Mortified Life}) = 1$ ,  $U_{Bob_G}(\text{Moderate Life}) = 10$ , and  $U_{Bob_G}(\text{Glorified Life}) = 100$ . This is a legitimate assignment of utility numbers, since it correctly represents *Bob<sub>G</sub>*'s preferences *ordinally*. However, with such assignment of utility numbers, it is just simply meaningless to say such things as, "For *Bob<sub>G</sub>*, Moderate Life is 10 times better than Mortified Life" or "*Bob<sub>G</sub>* would prefer going up from Moderate Life to Glorified Life 90 times as much as he would prefer going up from Death to Mortified Life." This is because  $U_{Bob_G}(\text{Death}) = 3$ ,  $U_{Bob_G}(\text{Mortified Life}) = 4$ ,  $U_{Bob_G}(\text{Moderate Life}) = 5$ , and  $U_{Bob_G}(\text{Glorified Life}) = 6$  is another legitimate assignment of utility numbers and we can see that none of these statements hold true in the new assignment of utility numbers.

As it will turn out, we will only be able to estimate the relative distances of each consequence in each type of *Bob<sub>i</sub>*'s preference-ordering only when each type of *Bob<sub>i</sub>* faces risk or uncertainty. And this is when each type of *Bob<sub>i</sub>*'s preferences will start to become distinguishable.

### **5.3 An Expected Utility Representation of Hobbes's Theory of Real Good**

In the previous subsection when we were giving an ordinal representation of Hobbes's theory of real good, the consequences that each type of *Bob<sub>i</sub>* was expressing his preference towards were sure-outcomes that involved no elements of uncertainty. This is not what usually happens in real life. In real life, when somebody performs a certain action, it is seldom the case that a specific outcome occurs for sure. Whether or not a specific outcome realizes depends

on what the state of the world actually turns out to be. And what state of the world actually turns out to be comes in various degrees of likelihoods (or probabilities.)

### **5.3.1 Clarifying the Meaning of Expected Utility Theory**

To model this correctly, we would have to use what is known as “expected utility theory.” And expected utility theory is another field, which I believe, is very easy to misunderstand. Confusion is likely to occur by, again, borrowing the meaning of “utility” from classical utilitarianism (as denoting good consequences) and interpreting the phrase “expected utility theory” in terms of it.

When the term “utility” is understood in terms of utilitarianism, as denoting good consequences, it is very easy to think that expected utility theory is some normative ethical theory that claims that the (morally) right thing to do is to maximize the *expectation* of good consequences.

Expected utility theory, understood as a normative ethical theory in this way (call it “expected utilitarianism”), has its appeal when the action that one had performed, despite one’s intention to maximize good consequences, turns out to be a complete disaster. According to utilitarianism, in such cases, what one had done was, *as a matter of fact, morally wrong*. However, even somebody who has very strong utilitarian inclinations might still want to leave some room to say that such action was at least *morally admissible* or even *rational*. That is, for all that one may have known, the probability of such disaster from happening might have been extremely low, while the probability of there being maximum amount of good consequences might have been very high. If one pursues this line of thought a little bit further, one might end up advocating expected utilitarianism that claims that what is morally right (or rational) to do is, not to maximize good consequences *per se*, but to maximize *the expectation of good consequence*.

Such is the appeal of expected utilitarianism. But, as a normative ethical theory, expected

utilitarianism is very implausible. To illustrate this, consider the following example. Suppose that you are chosen randomly by an all-powerful being who gives you the following two options to choose from: (a) everybody in the universe lives a very modest but independent life, and (b) the all-powerful being flips a fair coin; if the coin lands heads, everybody becomes a slave and lives a very miserable life; if the coin lands tails, everybody lives an extremely luxurious and independent life. Suppose that the expected life quality of (b) is greater than that of option (a). What would you choose?

It is obvious what you *should* choose if you adopted expected utilitarianism; you should choose option (b) since it has a higher expectation of good consequences.

It is not obvious that such a prescription is right; I believe that many people would think that option (a) is better. And the reason why many people might think this way is because they think that *risk* (or *variance*) is another aspect that we should take into account when making important decisions.

Option (a) might have a lower expected life quality than option (b); however, option (a) is completely risk-free. Option (b) might have a higher expected life quality than option (a); however, option (b) is very risky (i.e. it has high variance.) So, when expected utilitarianism requires you to choose option (a) over option (b) solely by the fact that option (a) has a higher expectation of good consequences, what it is, in effect, requiring is for you to be completely *risk-neutral* about the consequences in question. In other words, it requires you to be care-free about risk.

This is not how we normally think. When we try to decide what to do, we usually take *both* the *expectation* as well as the *risk* (i.e. variance) into consideration. If a normative ethical theory requires one to be risk-neutral, it, at least, owes one an explanation as to why. I believe that there is no such explanation that expected utilitarianism offers. One could very well be risk-neutral about external consequences; but, this does not mean that one *has to be*.

The implausibility of expected utilitarianism can be also illustrated by the following example that involves insurance. Suppose that you buy a new car that costs \$20,000. You are a

safe driver; but, suppose that there is a 1% chance of your car getting completely destroyed by some external cause that is out of your control. An insurance company offers you a full coverage of your car at \$401. Should you buy the insurance or not?

If you buy the insurance you retain \$19,599 worth of cash and merchandise for sure. If you do not buy the insurance, then your expected pay-off becomes:  $\frac{99}{100}(\$20,000) + \frac{1}{100}(-\$20,000) = \$19,600$ . Other things being equal, expected utilitarianism would require you *not* to buy the insurance.

When we think that most insurance policies generally work in this way, (that is, they offer you a certainty equivalent that is lower than the expectation), expected utilitarianism would virtually imply that it is *almost always morally wrong (or, at the very least, irrational) to buy insurance*.

This is not how we normally think. And, this is why expected utilitarianism does not seem to provide the right prescriptive guidelines when there is a lot of risk involved. So, as a normative ethical theory, the plausibility of expected utilitarianism is not something that is intuitively apparent.

Whatever the plausibility of “expected utilitarianism”, it is *not* what “expected utility theory” in decision theory claims. So, one should be cautious not to read any connotations of expected utilitarianism into expected utility theory.

As we have seen, in contemporary utility theory, the term “utility” is exclusively used to denote a numerical value that represents a person’s *preference*. A “utility function” is a *scale* that is designed to measure a person’s preference. The same thing holds for expected utility theory as well.

In contemporary decision theory, expected utility theory consists in a set of axioms and a main representation theorem (derived from the axioms) that claims, “ if a decision maker’s preferences abide by the set of axioms, there *exists* a utility function such that the decision maker strictly prefers  $x$  to  $y$  if and only if the *expected utility* of  $x$  is greater than the *expected utility* of  $y$ .” What this is basically saying is that as long as a decision maker’s preferences

(which are revealed by his or her choice behaviors) meet the set of axioms, we can *find* a utility function (i.e. a scale) which the decision maker is acting *as if* he or she was maximizing the expectation of.

The “as if” clause here is very important. Expected utility theory does not require the decision maker to understand anything about utility functions or probabilities or require the decision maker to be consciously aware that he or she is acting in a way the expected utility theory describes. What expected utility theory is saying is that as long as the decision maker’s preferences (which are revealed by his or her choice behaviors) abide by the set of axioms, there exists a utility function that enables us to interpret the decision maker’s choice behaviors as if he or she was maximizing the expectation of this utility function.

Now, if one understands the significance of what this is saying, one would realize that the claim is something that is very bold that requires theoretical defense. We have just seen that maximizing the expectation of something can be the sole consideration in one’s practical deliberation only when one is (or is required to be by some normative ethical theory) *risk-neutral* about the values or consequences the expectation of which is being sought to be maximized. We have seen that when one is either risk-averse or risk-loving, the fact that a certain action maximizes the expected consequences is not a decisive reason to act in that way.

However, expected utility theory claims that whenever somebody’s preferences meet a certain set of axioms (all of which, as we will soon see, seem pretty reasonable), we can always find a utility function the expectation of which the person acts *as if* he or she was seeking to maximize.

What’s remarkable is that this holds *regardless of one’s attitude towards risk*. That is, whether one is generally risk-averse or whether one is generally risk-loving, one can always be interpreted as a *expected utility maximizer* as long as one’s preferences satisfy a certain set of formal conditions. (Of course, the respective utility functions of a risk-averse person and a risk-loving person will turn out to be different.)

What is also remarkable of expected utility theory is that the number of axioms that jointly purport to render one to be an expected utility maximizer is quite few, and, they all seem to be pretty reasonable assumptions; assumptions that one would expect that any reasonable person's preferences would satisfy.

In reconstructing Hobbes's theory of real good, I plan to combine the strategies of both Von Neumann and Morgenstern's *objective* expected utility theory<sup>20</sup> and Leonard Savage's *subjective* expected utility theory.<sup>21</sup> I will not delve too much into the technical details and try to simplify the exposition as best as I can.

### 5.3.2 *An Expected Utility Representation (Von-Neumann and Morgenstern's Framework)*

We first describe our current framework. As before, the set of consequences is  $C = \{\text{Death, Mortified Life, Moderate Life, Glorified Life}\}$ . We now think of the set  $P$  of all probability distributions on  $C$ . A probability distribution is a function  $p : C \rightarrow [0, 1]$  such that  $\sum_{x \in C} p(x) = 1$ . It might be convenient to think of  $p \in P$  as a lottery ticket that gives you prize  $x \in C$  with probability  $p(x)$ .

Remember that  $P$  is the set of all possible probability distributions on  $C$ . Since  $C$  has more than one members,  $P$  is, in effect, an *uncountably infinite* set. Furthermore, we also assume that any convex combination of any two members of  $P$  is also a member of  $P$ . That is, if  $p \in P$  and  $q \in P$ , then for  $a \in [0, 1]$ ,  $ap + (1 - a)q \in P$ . It might be convenient to think of  $ap + (1 - a)q$  as a *compound lottery* which gives the lottery  $p$  (with probability  $a$ ) and the lottery  $q$  (with probability  $1 - a$ ) as its respective prizes.

---

<sup>20</sup>John Von Neumann and Oskar Morgenstern's expected utility theory has been first appeared in [Von Neumann and Morgenstern, 1944]. The theory has been reproduced in many books (see [Luce and Raiffa, 1957, Fishburn, 1970, Harsanyi, 1977]) afterwards, and has been explained in introductory text books in decision theory (see [Resnik, 1987, Kreps, 1988, Binmore, 2009].)

<sup>21</sup>The major contribution of Leonard Savage is to combine the theory of subjective probability with Von Neumann and Morgenstern's expected utility theory. The ground breaking work is [Savage, 1972]. The theory is reproduced and explained in [Fishburn, 1970, Kreps, 1988].

In this framework, each type of  $Bob_i$  ( $i = +, S, G$ ) is offered to express his preferences towards lotteries (i.e. probability distributions) in  $P$ . If we are willing to identify the lottery  $p \in P$  that gives  $x \in C$  as its prize with probability  $p(x) = 1$  with the sure consequence  $x \in C$ , then  $C$  can be seen as a proper subset of  $P$ , and, therefore, each type of  $Bob_i$  can also be seen as expressing his preference on  $C$  as well.

Lotteries in  $P$  would be options such as: “one-half chance of achieving a Glorified Life and a one-half chance of Death”, “.1 chance of Death, .4 chance of Mortified Life, .3 chance of Moderate Life, .2 chance of Glorified Life”, etc.

We can see that when each type of  $Bob_i$  is required to express his preferences towards lotteries, the *objective probabilities* of the lotteries are already predetermined and given to him. So, the choice that each type of  $Bob_i$  is making is different from betting on a horse race or a sports event in which it is standardly assumed for there to be no such objective probabilities that can be given.

We now state the axioms.

**[AXIOM(A-3-1)]:**  $\succ_{Bob_i}$  ( $i = +, S, G.$ ) on  $P$  and  $C$  is *asymmetric* and *negatively transitive*.

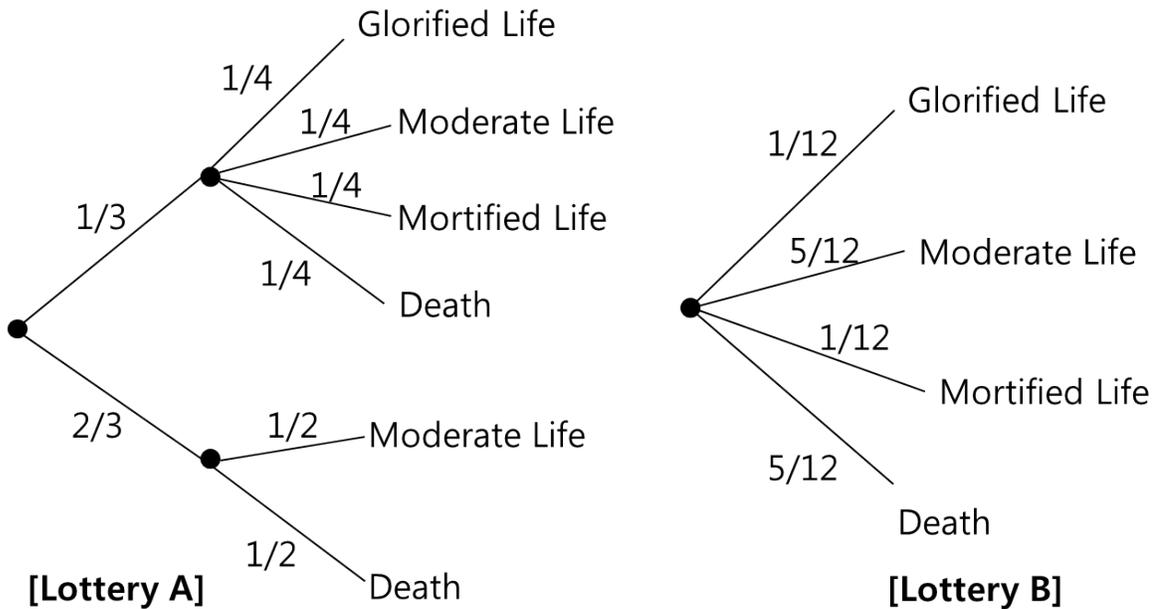
As we have seen, asymmetry and negative transitivity are two properties that fit very well with our intuitive understanding of what a preference-relation should be. The other relations,  $\succeq_{Bob_i}$  (weak preference) and  $\sim_{Bob_i}$  (indifference) are defined in a similar way as in D-2-1 and D-2-2.

**[AXIOM(A-3-2)]:** (*Reduction of Compound Lotteries*) Let  $a, b, c \in [0, 1]$ . Then, for all

$$p, q \in P, c(ap + (1 - a)q) + (1 - c)(bp + (1 - b)q) \sim_{Bob_i} (ac + b - bc)p + (1 - (ac + b - bc))q$$

I believe that what this axiom is claiming can be more easily understood by a concrete example. Consider the following two lotteries.

Figure 5.3: Reduction of Compound Lotteries



Lottery B is a simple lottery that gives the prizes Glorified Life with probability  $1/12$ , Moderate Life with probability  $5/12$ , Mortified Life with probability  $1/12$ , and Death with probability  $5/12$ . Whereas, lottery A is a compound lottery that gives the lottery (which gives the prizes Glorified Life, Moderate Life, Mortified Life, and Death, each with probability  $1/4$ ) as a prize with probability  $1/3$ , and the lottery (which gives the prizes Moderate Life and Death each with probability  $1/2$ ) as a prize with probability  $2/3$ . The two lotteries are different. However, if we calculate the probabilities of each lottery, we can confirm that both lottery A and lottery B provide each of the prizes (i.e. Glorified Life, Moderate Life, Mortified Life) with exactly the same probabilities.

Axiom A-3-2 claims that, in such cases, each type of  $Bob_i$  is *indifferent* between the two lotteries. That is, according to axiom A-3-2, the only things that each type of  $Bob_i$  considers in determining his preference between two options are the specific consequences and their respective final probabilities of occurring; and not the particular way the consequences are randomized.

Such assumption might not be true for ordinary people in many cases. For example, I believe that, for many people, it is not obvious, at first sight, that lottery A provides the prize Moderate Life with probability  $5/12$ . To know that lottery A provides Moderate Life with probability  $5/12$  requires computation. The computation will become more complicated as the compound lottery becomes more complex. People might not like computation. If so, then they might actually prefer lottery B to lottery A even though each lottery generates exactly the same probabilities for each of the consequences. So, what axiom A-3-2 is basically assuming is that *computation comes for free*.

This might be problematic for ordinary people. However, I believe that this will not be a problem for at least  $Bob_+$  (i.e. the actual Bob's *idealized-self*) who is assumed to be fully-rational in the Hobbesian substantial sense. I believe that assuming that such idealized being has the requisite computational abilities to be indifferent between two different compound lotteries that assign exactly the same final probabilities for each of the consequences would not be such a problematic assumption.

The assumption might be contestable for  $Bob_S$  and  $Bob_G$ . However, I believe that the axiom can be applied to even non-idealized people if they are patient enough to go through the computations of probabilities when required.

For now, let's assume that both  $Bob_S$  and  $Bob_G$  are such patient individuals. In the end, both  $Bob_S$  and  $Bob_G$  will not be required to understand anything about probabilities in Savage's framework which will be dealt in the next subsection. Until then, let us assume that they do have some elementary knowledge of arithmetic and probability theory which enable them to compute probabilities in compound lotteries.

**[AXIOM(A-3-3)]:** (*Independence*) For all  $p, q, r \in P$  and  $a \in (0, 1]$ , if  $p \succ_{Bob_i} q$  then  $ap + (1 - a)r \succ_{Bob_i} aq + (1 - a)r$

Suppose a given type of  $Bob_i$  is considering the following two options: (a) an arbiter tosses a fair coin, and if the coin lands heads  $Bob_i$  experiences Moderate Life, and if the coin lands

tails he suffers Death, and (b) an arbiter tosses a fair coin, and if the coin lands heads  $Bob_i$  experiences Mortified Life, and if the coin lands tails he suffers Death.

Here, we can see that what happens when the coin lands tails is exactly *the same* for both options (a) and (b). What axiom A-3-3 claims is that, in such cases,  $Bob_i$  preference for the two options (a) and (b) should be determined by what happens when the coin lands heads.

As it turns out, all types of  $Bob_i$  strictly prefers a Moderate Life to a Mortified Life which are the respective consequences for (a) and (b) when the coin lands heads ; therefore, all types of  $Bob_i$  strictly prefers (a) to (b). The basic logic behind this reasoning is this: if one is going to die no matter what one chooses if the coin lands tails, then why not choose an option that at least gives a better result when the coins lands heads?

This axiom is usually known as the *independence* or the *substitution axiom*. The thought behind this is that as long as what happens at a particular state of the world is the same for any two options, the decision maker’s preference between the two options is not influenced by what particular consequence is realized in that particular state of the world.

This means that, for all types of  $Bob_i$ ,  $[\frac{1}{2}\text{Moderate Live} + \frac{1}{2}X] \succ_{Bob_i} [\frac{1}{2}\text{Mortified Life} + \frac{1}{2}X]$  for all  $X \in C$ . I believe that this is a reasonable assumption in our current context although, historically, there have been objections that have been raised against it.<sup>22</sup>

**[Axiom(A-3-4): (Archimedean)]** For all  $p, q, r \in P$ , if  $p \succ_{Bob_i} q \succ_{Bob_i} r$ , then there exist

---

22

The famous objection is what is now known as the *Allais Paradox*. Suppose that there are one hundred balls in a bag that are numbered from 1 to 100.

Table 5.2: Allais Paradox

Exactly one ball is drawn, and you receive a prize according to the lottery you choose. You are offered a choice between the two lotteries  $p_1$  and  $p_2$  which are described in the table below.

Lottery $p_1$				Lottery $p_2$			
<b>Ball #</b>	1	2 to 11	12 to 100	<b>Ball #</b>	1	2 to 11	12 to 100
<b>Prize</b>	\$100	\$100	\$100	<b>Prize</b>	\$0	\$500	\$100

Which lottery do you prefer? Many subjects in the experiment answered that they prefer lottery  $p_1$  to  $p_2$ . Suppose that you are now asked to choose between the following two lotteries  $p_3$  and  $p_4$ .

$$a, b \in (0, 1) \text{ such that } ap + (1 - a)r \succ_{Bob_i} q \succ_{Bob_i} bp + (1 - b)r.$$

This, I believe, is the axiom that does the most heavy duty in deriving the representation theorem. What it is basically claiming is that if there are three options that one strictly prefers the first to the second and one strictly prefers the second to the third, we can always find a lottery that takes the first and third options as prizes which one would strictly prefer to the second option, and another lottery for which one would strictly prefer the second option to it. What this virtually amounts to saying is that there are no options that are *infinitely good* or *infinitely bad*. Let me explain.

Suppose that one strictly prefers having a Ferrari to having a Honda which one strictly prefers to having no car at all. Now consider the lottery  $[a \cdot \text{Ferrari} + (1 - a) \cdot \text{NoCar}]$  which gives you a Ferrari with probability  $a$  ( $0 \leq a \leq 1$ ) and No Car with probability  $(1 - a)$ . Presumably, with a high enough  $a$ , we could expect that one would strictly prefer the lottery to getting a Honda for sure. Similarly, with a low enough  $a$ , we could expect that one would strictly prefer getting a Honda to the lottery. And this is exactly what the Archimedean axiom

Lottery $p_3$				Lottery $p_4$			
Ball #	1	2 to 11	12 to 100	Ball #	1	2 to 11	12 to 100
Prize	\$100	\$100	\$0	Prize	\$0	\$500	\$0

Which lottery do you prefer? Many subjects in the experiment answered that, this time, they prefer lottery  $p_4$  to  $p_3$ . It can be easily seen that this violates the independence axiom.

**Proof.** Suppose that  $a$  denotes the lottery that gives the outcome \$1,000,000 for sure,  $b$  denotes the lottery that gives \$0 with probability  $\frac{1}{11}$  and \$5,000,000 with probability  $\frac{10}{11}$ ,  $d$  the lottery that give \$0 for sure. Then, we can write each of lotteries as follows:  $p_1 = \frac{11}{100}a + \frac{89}{100}a$ ,  $p_2 = \frac{11}{100}b + \frac{89}{100}a$ ,  $p_3 = \frac{11}{100}a + \frac{89}{100}d$ ,  $p_4 = \frac{11}{100}b + \frac{89}{100}d$ . According to the independence axiom, this implies,  $p_1 \succ p_2$  iff  $\frac{11}{100}a + \frac{89}{100}a \succ \frac{11}{100}b + \frac{89}{100}a$  iff  $\frac{11}{100}a + \frac{89}{100}d \succ \frac{11}{100}b + \frac{89}{100}d$  iff  $p_3 \succ p_4$ .  $\therefore$  Therefore, if the agent reports  $p_1 \succ p_2$  and  $p_4 \succ p_3$ , then this violates the independence axiom.  $\square$

There are many ways to respond to this objection. One is to think that this example invalidates the plausibility of the independence axiom. Another is to think that the subjects in the experiment were acting irrationally; that once they are informed of their mistakes there are likely to change their preference according to the independence axiom. (It is known that this is the stance that Leonard Savage took after he realized that he failed to be consistent with the independence axiom in Allais's experiment himself.) Lastly, one might try to argue that the results of the experiment is not really inconsistent with the independence axiom by individuating outcomes more finely. For example, one might argue that getting \$0 in lottery  $p_2$  is different from getting \$0 in lottery  $p_4$ ; one is more likely to feel regret in the former than the latter. See [Broome, 1991, Chapter 5] for an account of such solution.

is claiming. Of course, the specific values of  $a$  that would reverse one's preferences would depend on how much one values a Ferrari as well as how much one dislikes having no car.

However, now, consider a more drastic example. Presumably, many people would strictly prefer receiving \$100 to receiving no money at all, which they would strictly prefer to death. The Archimedean axiom claims that there is a number  $a \in (0, 1)$  such that one would strictly prefer the lottery  $[a \cdot \$100 + (1 - a) \cdot \text{Death}]$  to receiving no money. To many people, this would sound outrageous. To them, having even the slightest probability of death would render the lottery not worth playing. Many people think that death is something that is very close to being *infinitely bad*, of which the Archimedean axiom denies the existence.

The usual defense of the Archimedean axiom to this kind of objection goes as follows.<sup>23</sup> Suppose that a trustworthy friend tells you that, across the street, there is an envelop that has a a hundred dollar bill inside in his mailbox which he would like to give me as a gift. You can either refuse the gift or accept the gift and cross the street to retrieve it. I believe that many people would gladly choose to accept the gift and cross the street even though crossing the street increases ever so slightly the chances of getting hit by a car and dying. This shows that death, unlike what many people commonly think, is not something that is infinitely bad.

I think that the same thing can be reasonably argued for each type of  $Bob_i$ .

First of all, it is obvious that  $Bob_G$  will not have an infinite aversion against death. If he did, then he would never choose a gamble that involved a good chance of glory and a slight chance of death over securing a moderate life, which defies the basic spirit of Hobbes's main text.<sup>24</sup>

Next, I believe that the Archimedean axiom can even be defended for  $Bob_+$  and  $Bob_S$  as well. It is true that both  $Bob_+$  and  $Bob_S$  are the type of Bobs that are primarily influenced by a basic desire for self-preservation. And because of that, both  $Bob_S$  and  $Bob_+$

---

<sup>23</sup>See [Fishburn, 1970, p.110] and [Kreps, 1988, p. 45]

<sup>24</sup>“Vain-glorious men ... are inclined to rash engaging...” [Hobbes 1994: *Leviathan*, Chapter XI, Paragraph 12] “Another, supposing himself superior to other, wants to be allowed everything (...) that is the sign of an *aggressive character*. In his case, the will to do harm derives from *vainglory*.” [Hobbes 1997: *On the Citizen*, Chapter 1, Section 4 emphasis added]

would, more than anything else, be concerned with securing his (i.e. the actual-Bob's) long-term self-preservation. This implies that  $Bob_+$  would have a very strong aversion against death; however, this does not mean that they would have *infinite aversion* towards death. For instance, I believe that it is not unreasonable to assume that both  $Bob_+$  and  $Bob_S$  would strictly prefer the lottery (which gives 99.9999999999% chance of living a Moderate Life and 0.0000000001% chance of Death) to living a Mortified Life for sure.

So, all of the axioms from A-3-1 to A-3-4, at least, seem to be defensible assumptions for each type of  $Bob_i$ . The remarkable thing is that this is all that we need to derive the *expected utility representation theorem*, which is stated below.

**[THEOREM(T-3-1)] (*Expected Utility Representation*):** The binary relation  $\succ_{Bob_i}$  ( $i = +, S, G$ ) (i.e.  $Bob_i$ 's strict preference) on  $P$  satisfies axioms A-3-1 to A-3-4 if and only if there exists functions  $U_{Bob_i} : P \rightarrow \mathbb{R}$  and  $u_{Bob_i} : C \rightarrow \mathbb{R}$  such that: For all  $p, q \in P$ ,  $p \succ_{Bob_i} q$  if and only if  $U_{Bob_i}(p) > U_{Bob_i}(q)$  if and only if  $\sum_{x \in C} p(x)u_{Bob_i}(x) > \sum_{x \in C} q(x)u_{Bob_i}(x)$ . Furthermore, functions  $U_{Bob_i}$  and  $u_{Bob_i}$  are unique up to positive affine transformation.

Some writers distinguish the big utility function (i.e.  $U_{Bob_i}$ ) with the small utility function (i.e.  $u_{Bob_i}$ ) by calling the latter a "pay-off function." The difference is that the domain of  $U_{Bob_i}$  is  $P$  while the domain of  $u_{Bob_i}$  is  $C$ . However, we can think of them as the same thing if we are willing to think that a consequence  $x$  in  $C$  is the same thing as a lottery (i.e. a probability distribution) in  $P$  that gives  $x$  as its prize with probability 1. And, this is what I will be doing from now on; that is, from what I write afterwards, I will not distinguish between  $U_{Bob_i}$  and  $u_{Bob_i}$ .

Granting this, what the theorem is basically claiming is that as long as each type of  $Bob_i$ 's preferences meet the four axioms from A-3-1 to A-3-4, we can find a utility function such that: (a) it represents a given  $Bob_i$ 's preferences, and (b) the utility of a lottery *equals its*

*expected utility*.<sup>25</sup>

Remember that we say that a utility function represents a person's preference if the person strictly prefers option  $p$  to option  $q$  if and only if the utility of  $p$  is greater than the utility of  $q$ . So, what the theorem is basically claiming is that as long as each type of  $Bob_i$ 's preferences meet the four axioms, any given type of  $Bob_i$  strictly prefers lottery  $p$  to lottery  $q$  if and only if *the expected utility* of  $p$  is greater than *the expected utility* of  $q$ .

That is, as long as a given type of  $Bob_i$ 's preferences meet the four axioms, there is a utility function that makes it look *as if* that given type of  $Bob_i$  was *a expected utility maximizer*. Furthermore, if we find one such utility function, then we can find another utility function by performing a positive affine transformation on the former.

I hope that everybody who is reading this finds this remarkable. As we have seen, people do not generally act in order to maximize the expectation of some external value. This is because, as we have seen, maximizing the expectation of some external value, in effect, requires one to be *risk-neutral* about the value in question. And, as we have seen, people can be *risk-averse* or *risk-neutral* about many things. However, theorem T-3-1 claims that as long as a decision maker's preference meet the four axioms, it is possible for us to *see* the decision maker as an expected utility maximizer.

As one can see, there is absolutely no axiom above that requires one to be risk-neutral about anything. This means that, as long as their preference meet the axioms, even risk-averse or risk-loving people can be seen as expected utility maximizers. To put this in another way, we can say that, as long as one's preference satisfy the four axioms, there is a utility function whose value towards which one is *risk-neutral* regardless of one's attitude towards risk! (I hope that everybody who is reading this finds this quite remarkable.)

I will not go into the details of the proof. Anybody who is interested should consult any standard textbook in modern decision theory.<sup>26</sup> What is important for our current purpose is

---

<sup>25</sup>That is,  $U_{Bob_i}(px + (1-p)y) = pU_{Bob_i}(x) + (1-p)U_{Bob_i}(y)$ . When a utility function has this property, it is said that the utility function is *linear* or *expectational*.

<sup>26</sup>The original proof is in [Von Neumann and Morgenstern, 1944]. Many people have reproduced the proof

to understand the general strategy that one uses to prove the main theorem.

The first major step is to establish the following lemma from the axioms:

**[LEMMA(L-3-1)]:** (*Continuity*) Suppose  $p, q, r \in P$  and  $p \succ_{Bob_i} q \succ_{Bob_i} r$ . Then, there exists a unique  $a \in [0, 1]$  such that  $ap + (1 - a)r \sim_{Bob_i} q$ .

Some authors even state this lemma as a separate axiom.<sup>27</sup> What this lemma is saying is that if a given type of  $Bob_i$  strictly prefers option  $p$  to option  $q$  to option  $r$ , then there is a single lottery (whose prizes consists in option  $p$  and option  $r$ ) to which  $Bob_i$  will be indifferent with getting option  $q$  for sure.

The basic intuition is something like this. Let's go back to the car example where one's preferences are  $Ferrari \succ Honda \succ NoCar$ . Now, compare the lottery  $[a \cdot Ferrari + (1 - a) \cdot NoCar]$  (where  $a \in [0, 1]$ ) with the option of getting a Honda for sure. The Archimedean axiom claims that when  $a$  is sufficiently close to 1, one would prefer the lottery to the Honda, and when  $a$  is sufficiently close to 0, one would prefer a Honda to the lottery.

Now, pick a  $a$  where one strictly prefers the lottery to getting a Honda for sure. Continuously decrease the value of  $a$  by making it less and less. Eventually, one is going to reach a certain point where one feels *indifferent* between the lottery and a Honda (i.e. where decreasing the value of  $a$  anymore will result in making one strictly prefer a Honda to the lottery.) This is the point that lemma L-3-1 claims to exist.

At this point, the decision maker will be indifferent to getting the lottery to getting a Honda for sure. And since a utility function represents the decision maker's preference, this implies that the utility of the lottery will be equal to the utility of a Honda.

Now, let's apply this fact to our current framework. What we have to do is to use continuity (i.e. Lemma L-3-1) to *calibrate* the utilities of each type of  $Bob_i$  for the consequences in  $C \subseteq P$ . What we are aiming to do is to construct a utility function that represents a given type

---

of the main theorem by using slightly different sets of axioms. Different versions of the proof can be found in [Luce and Raiffa, 1957, Jensen, 1967a, Fishburn, 1970, Harsanyi, 1977, Resnik, 1987, Kreps, 1988, Binmore, 2009], etc.

<sup>27</sup>For instance, see[Luce and Raiffa, 1957, Harsanyi, 1977, Resnik, 1987]

of  $Bob_i$ 's preferences for lotteries in  $P$  such that *the utility of a lottery equals the expected utility of the lottery*.

To do so, we first arbitrarily assign two different numbers to represent the utilities of two distinct consequences in  $C$  (or two lotteries in  $P$  that assigns probability 1 to two different consequences.) Assigning any two numbers for any two distinct consequences will be fine as long as one assigns a greater number to the preferred consequence between the two.

However, it is convenient to assign utility 0 to the least preferred consequence and utility 1 to the most preferred consequence in  $C$ . This process is called *normalization* - which means that we are assigning the *zero* and *unit* for the utility scale under consideration. This means that, in our current framework, we assign  $U_{Bob_i}(Death) = 0$  and  $U_{Bob_i}(GlorifiedLife) = 1$  for all  $i$ .

We now try to find a lottery between Death (which is the worst outcome) and Glorified Life (which is the best outcome) (i.e.  $[p \cdot \text{Glorified Life} + (1 - p) \cdot \text{Death}]$ ) to which each type of  $Bob_i$  would feel indifferent to securing a Moderate Life for sure.

Let's start with  $Bob_+$ . Presumably, since  $Bob_+$  is defined as the idealized-Bob who is, more than anything else, concerned with achieving the actual-Bob's long-term self-preservation, the value of  $p \in [0, 1]$  would be quite high (i.e. very close to 1.) In any case, we know from Lemma L-3-1, (which can be derived from Axioms A-3-1 to A-3-4), that such  $p$  is guaranteed to exist.

Let  $p^+$  be the value of  $p$  that makes  $Bob_+$  feel indifferent between the lottery  $[p \cdot \text{Glorified Life} + (1 - p) \cdot \text{Death}]$  and securing a Moderate Life for sure. For now, let's do not specify the value that  $p^+$  takes. Just remember that it has to be quite close to 1. Now, assign the utility of Moderate Life as  $p^+$ . So,  $U_{Bob_+}(ModerateLife) = p^+$ .

We do the same thing for the consequence, Mortified Life. That is, let  $q^+ \in [0, 1]$  be the value of  $q$  that would make  $Bob_+$  indifferent to playing the lottery  $[q \cdot \text{Glorified Life} + (1 - q) \cdot \text{Death}]$  and securing a Mortified Life for sure. Obviously the value of  $q^+$  should be smaller than that of  $p^+$  (i.e.  $p^+ < q^+$ .) Again, such a value is guaranteed to exist by Lemma

L-3-1. Now, assign  $q^+$  as the utility of Mortified Life. So,  $U_{Bob_+}(MortifiedLife) = q^+$ .

We have just constructed an *expected utility function* for  $Bob_+$ . That is:

- $U_{Bob_+}(GlorifiedLife) = 1$ ,
- $U_{Bob_+}(ModerateLife) = p^+$
- $U_{Bob_+}(MortifiedLife) = q^+$
- $U_{Bob_+}(Death) = 0$

We can easily verify that our current utility function  $U_{Bob_+}$  has the *expected utility property* (i.e. it is *linear*); *the utility of any lottery equals to its expected utility of the sure outcomes*.

For instance, we know that  $Bob_+$  is indifferent between the lottery [ $p^+$ ·Glorified Life +  $(1 - p^+)$ ·Death] and the sure outcome Moderate Life. This means that the utility of the lottery [ $p^+$ ·Glorified Life +  $(1 - p^+)$ ·Death] and the utility of Mortified Life should be the *same*; that is,  $U_{Bob_+}(ModerateLife) = U_{Bob_+}(p^+ \cdot GlorifiedLife + (1 - p^+) \cdot Death)$ .

Now, take the expected utility of the lottery [ $p^+$ ·Glorified Life +  $(1 - p^+)$ ·Death]:  
 $p^+U_{Bob_+}(GlorifiedLife) + (1 - p^+)U_{Bob_+}(Death) = p^+ \cdot 1 + (1 - p^+) \cdot 0 = p^+$ , which, as it turns out, is equal to  $U_{Bob_+}(ModerateLife)$ .

Combining all of this we get:

$$\begin{aligned}
 &U_{Bob_+}(ModerateLife) \\
 &= U_{Bob_+}(p^+ \cdot GlorifiedLife + (1 - p^+) \cdot Death) \\
 &= p^+U_{Bob_+}(GlorifiedLife) + (1 - p^+)U_{Bob_+}(Death) \\
 &= p^+
 \end{aligned}$$

We can see here that the utility of the lottery [ $p^+$ ·Glorified Life +  $(1 - p^+)$ ·Death] equals to taking the expectation of the utilities of its sure consequences, Glorified Life

and Death. Furthermore, if we think of the sure consequence of a Moderate Life as a lottery that gives Moderate Life with probability 1, then we can also confirm that the utility of the lottery  $[1 \cdot \text{Moderate Life}]$  equals the expected utility of its sure consequence:  $U_{Bob_+}(1 \cdot \text{Moderate Life}) = 1 \cdot U_{Bob_+}(\text{Moderate Life}) = p^+$ . The same thing applies for the utility of Mortified Life. That is,  $U_{Bob_+}$  is an expected utility function where the utility of a lottery is equal to the expected utility of the lottery.

Remember that a utility function is a scale that measures a person's preference in such a way that the person strictly prefers option  $a$  to option  $b$  if and only if the utility of option  $a$  is greater than the utility of option  $b$ . So, for any lotteries (i.e. probability distributions)  $p, q \in P$ ,  $Bob_+$  will strictly prefer the lottery  $p$  to lottery  $q$  if and only if the utility of lottery  $p$  is greater than the utility of lottery  $q$ . However, we have just seen that, for  $U_{Bob_+}$ , the utility of a lottery *equals* the expected utility of its sure outcomes.

This means that,  $Bob_+$  will strictly prefer lottery  $p$  to lottery  $q$  if and only if the expected utility of lottery  $p$  is greater than the expected utility of lottery  $q$ . In short,  $Bob_+$  is an *expected utility maximizer* who acts *as if* he was trying to maximize the expectation of  $U_{Bob_+}$ .

Furthermore, the uniqueness part of theorem T-3-1 claims that the expected utility function  $U_{Bob_+}$  is unique up to *positive affine transformation*. This means that if  $U_{Bob_+}$  is a utility function that represents  $Bob_+$ 's preferences and has the expected utility property explained above, then  $V = aU_{Bob_+} + b$  ( $a > 0, b \in \mathbb{R}$ ) is another utility function that represents  $Bob_+$ 's preferences and has the expected utility property as well.<sup>28</sup>

---

<sup>28</sup>Here is a general proof where  $C$  is assumed to be an arbitrary finite set. Suppose  $x_1, \dots, x_n \in C$ . Since  $Bob_+$ 's preferences on  $P$  have an expected utility representation, this implies that  $p \succ_{Bob_+} q$  iff  $U_{Bob_+}(p) > U_{Bob_+}(q)$  iff  $\sum_{x_i \in X} p(x_i)u_{Bob_+}(x_i) > \sum_{x_i \in X} q(x_i)u_{Bob_+}(x_i)$  iff  $p(x_1)u_{Bob_+}(x_1) + \dots + p(x_n)u_{Bob_+}(x_n) > q(x_1)u_{Bob_+}(x_1) + \dots + q(x_n)u_{Bob_+}(x_n)$  iff  $a\{p(x_1)u_{Bob_+}(x_1) + \dots + p(x_n)u_{Bob_+}(x_n)\} > a\{q(x_1)u_{Bob_+}(x_1) + \dots + q(x_n)u_{Bob_+}(x_n)\}$  (for  $a > 0$ ) iff  $p(x_1)au_{Bob_+}(x_1) + \dots + p(x_n)au_{Bob_+}(x_n) > q(x_1)au_{Bob_+}(x_1) + \dots + q(x_n)au_{Bob_+}(x_n)$  iff (since both  $\{p(x_1) + \dots + p(x_n)\} = 1$  and  $\{q(x_1) + \dots + q(x_n)\} = 1$ )  $p(x_1)au_{Bob_+}(x_1) + \dots + p(x_n)au_{Bob_+}(x_n) + \{p(x_1) + \dots + p(x_n)\}b > q(x_1)au_{Bob_+}(x_1) + \dots + q(x_n)au_{Bob_+}(x_n) + \{q(x_1) + \dots + q(x_n)\}b$  (for any real number  $b$ ) iff  $p(x_1)\{au_{Bob_+}(x_1) + b\} + \dots + p(x_n)\{au_{Bob_+}(x_n) + b\} > q(x_1)\{au_{Bob_+}(x_1) + b\} + \dots + q(x_n)\{au_{Bob_+}(x_n) + b\}$  Therefore, if  $Bob_+$ 's preferences on  $P$  have an expected utility representation with the pay-off function on prizes  $u_{Bob_+} : X \rightarrow \mathbb{R}$ , then the pay-off function  $v = au_{Bob_+} + b$  (for real numbers  $a > 0$  and  $b$ ) can also be used in the expected utility representation to represent the same preferences. That is,  $u_{Bob_+}$

I said that it is convenient to assign utility 1 to the most preferred outcome and utility 0 to the least preferred outcome. The reason why such assignment is convenient is because, that way, the utility of any sure-outcome will be equal to the probability of the most preferred outcome in a lottery [ $p \cdot \text{Most Preferred Outcome} + (1 - p) \cdot \text{Least Preferred Outcome}$ ] to which the individual will be indifferent with the sure-outcome in question.

In section 7.1, I have explained that the type of admissible transformations determines the *scale type* of a particular scale. As we have seen, what type of transformations are admissible is stated in the uniqueness part of theorem T-3-1. There, it was stated that  $U_{Bob_+}$  is unique up to positive affine transformation. According to table 7.1, a scale that is unique up to positive affine transformation is an *interval scale*. An interval scale is one kind of *cardinal scale*. And in an interval scale, the *ratio between differences* remain constant for all admissible transformation.

This means that, when  $|U_{Bob_+}(x) - U_{Bob_+}(y)| = k \cdot |U_{Bob_+}(w) - U_{Bob_+}(z)|$ , it is not completely meaningless to say that the distance between option  $x$  and option  $y$  in  $Bob_+$ 's preference-ordering is  $k$  times as great as the distance between option  $w$  and option  $z$  in  $Bob_+$ 's preference-ordering.<sup>29</sup>

From this, we are able to state the relative distances between the consequences in  $C$  in  $Bob_+$ 's preference-ordering without rendering such statements meaningless. There is no textual evidence in Hobbes that would enable us to know the relative distances among consequences in  $Bob_+$ 's preference-ordering in a precise way. But, we are able to make reasonable estimations.

First, remember that  $Bob_+$  is the idealized-Bob who is, more than anything else, concerned with securing the actual-Bob's long-term self-preservation. This implies that, within  $Bob_+$ 's preference-ordering, moving up from Death to Mortified Life would worth more than moving up from Mortified Life to Moderate Life *or* moving up from Moderate Life to Glori-

---

is unique up to positive affine transformation.□

<sup>29</sup>See footnote 5 for a proof that shows that a positive affine transformation retains the ratio of differences.

fied Life. Since,  $U_{Bob_+}$  is an (*interval*) *scale* that measures  $Bob_+$ 's preferences (*cardinally*), this implies that  $|U_{Bob_+}(Mortified\ Life) - U_{Bob_+}(Death)| = q^+ - 0 = q^+$

$> |U_{Bob_+}(Moderate\ Life) - U_{Bob_+}(Mortified\ Life)| = p - q$  and

$|U_{Bob_+}(Mortified\ Life) - U_{Bob_+}(Death)| = q^+ - 0 = q^+$

$> |U_{Bob_+}(Glorified\ Life) - U_{Bob_+}(Moderate\ Life)| = 1 - p$ .

Furthermore,  $Bob_+$  (unlike  $Bob_G$ ) would not attach much value to glory (i.e. power) *per se*;  $Bob_+$ 's concern for power would be limited to his interests in securing the actual-Bob's self-preservation. Therefore, for  $Bob_+$ , moving up from Moderate Life to Glorified Life would not be worth more than moving up from Mortified Life to Moderate Life (which means moving up from an abject life to a decent life.) This implies that  $|U_{Bob_+}(Moderate\ Life) - U_{Bob_+}(Mortified\ Life)| = p^+ - q^+$

$> |U_{Bob_+}(Glorified\ Life) - U_{Bob_+}(Moderate\ Life)| = 1 - p$ .

Summarizing all of this we get:

(a)  $p^+ > q^+$  (i.e. a Moderate Life is strictly preferred to a Mortified Life)

(b)  $2q^+ > p^+$

(c)  $p^+ + q^+ > 1$

(d)  $2p^+ > 1 + q^+$

Of course, there will be more than one set of values for  $p^+$  and  $q^+$  that satisfy the inequalities (a), (b), (c), (d). However, as one can easily confirm by simple algebra, there are *lower bounds* for both  $p^+$  and  $q^+$ , which are:

- $q^+ > \frac{1}{3}$  - <sup>30</sup>

- $p^+ > \frac{2}{3}$  - <sup>31</sup>

---

<sup>30</sup>Multiplying 2 to each side of (b) and connecting it with (d), we get:  $4q^+ > 2p^+ > 1 + q^+$ . From this, we get:  $q^+ > \frac{1}{3}$ .

<sup>31</sup>From the fact that  $q^+ > \frac{1}{3}$  and (d), we get:  $2p^+ > 1 + q^+ > 1 + \frac{1}{3} = \frac{4}{3}$ . From this, we get:  $p^+ > \frac{2}{3}$ .

Summarizing this, we get:

- $U_{Bob_+}(GlorifiedLife) = 1,$
- $U_{Bob_+}(ModerateLife) = p^+ > \frac{2}{3}$
- $U_{Bob_+}(MortifiedLife) = q^+ > \frac{1}{3}$
- $U_{Bob_+}(Death) = 0$

We now go over the same process for  $Bob_S$  and  $Bob_G$ . Remember that in our current stage of reconstruction (which utilizes Von-Neumann and Morgenstern's objective expected utility theory) all three types of Bobs are expressing his preferences towards lotteries with known objective probabilities. Therefore, if any of their preferences (along with their respective utility functions that represent those preferences) diverge, it will not be because any given type of Bob was insufficiently informed about the relevant probabilities, but rather because they were primarily being influenced by different kinds of basic passions.

Both  $Bob_+$  and  $Bob_S$  are assumed to be primarily influenced by the basic passion for self-preservation. Therefore,  $Bob_S$ 's utilities for each of the consequences in  $C$  will be identical to that of  $Bob_+$ . Let  $p^S$  and  $q^S$  respectively be the values of  $p$  in the lottery [ $p$ ·Glorified Life +  $(1 - p)$ ·Death] to which  $Bob_S$  would be indifferent to securing a Moderate Life and a Mortified Life. Then, we are able to summarize  $Bob_S$ 's (cardinal) utilities for the sure consequences as follows:

- $U_{Bob_S}(GlorifiedLife) = 1,$
- $U_{Bob_S}(ModerateLife) = p^S = p^+ > \frac{2}{3}$
- $U_{Bob_S}(MortifiedLife) = q^S = p^+ > \frac{1}{3}$
- $U_{Bob_S}(Death) = 0$

Now, let's move on to the preferences of  $Bob_G$ . Obviously,  $Bob_G$ , who is primarily influenced by the basic passion for glory, will have different preferences towards the various lotteries in  $P$ . And based on  $Bob_G$ 's preferences between the lottery  $[p \cdot \text{Glorified Life} + (1 - p) \cdot \text{Death}]$  and the sure consequences Moderate Life and Mortified Life, we are able to calibrate the relative distances of these sure consequences in  $Bob_G$ 's preference-ordering as we did before in the case of  $Bob_+$ .

Again,  $Bob_G$  is the type of Bob who is assumed to be primarily under the influence of a basic passion for glory. It is not hard to expect that such type of Bob would value a glorified life extremely highly. So, for  $Bob_G$ , moving up from Moderate Life to Glorified Life would definitely be worth more than moving up from Mortified Life to Moderate Life or moving up from Death to Mortified Life in  $Bob_G$ 's.

Furthermore, the fact that  $Bob_G$  strongly desires glory and honor implies that he would have a strong aversion against dishonor and mortification. Of course, as we have seen,  $Bob_G$ 's aversion against dishonor would not be so strong to the extent that he would rather prefer Death to a Mortified Life: but his preference for a Mortified Life would not be that strong. This implies that, for  $Bob_G$ , moving up from Death to Mortified Life would be worth less than moving up from Mortified Life to Moderate Life.

Let  $p^G$  and  $q^G$  respectively be the values of  $p$  in the lottery  $[p \cdot \text{Glorified Life} + (1 - p) \cdot \text{Death}]$  to which  $Bob_G$  would be indifferent to securing a Moderate Life and a Mortified Life. By summarizing all of the facts above, we get the following four inequalities:

- (a)  $p^G > q^G$  (i.e. a Moderate Life is strictly preferred to a Mortified Life)
- (b)  $1 - p^G > p^G - q^G$
- (c)  $1 - p^G > q^G$
- (d)  $p^G - q^G > q^G$

Again, there will be more than one set of values of  $p^G$  and  $q^G$  that would satisfy these

four inequalities. However, we are able to derive the *lower bounds* of  $p^G$  and  $q^G$  from these inequalities, which can be summarized below:

- $q^G < \frac{1}{3}$  - <sup>32</sup>

- $p^G < \frac{2}{3}$  - <sup>33</sup>

From this, we are able to summarize  $Bob_G$ 's (cardinal) utilities for the sure consequences as follows:

- $U_{Bob_S}(Glorified\ Life) = 1,$

- $U_{Bob_S}(Moderate\ Life) = p^G < \frac{2}{3}$

- $U_{Bob_S}(Mortified\ Life) = q^G < \frac{1}{3}$

- $U_{Bob_S}(Death) = 0$

It will be convenient to see what we have established so far graphically.

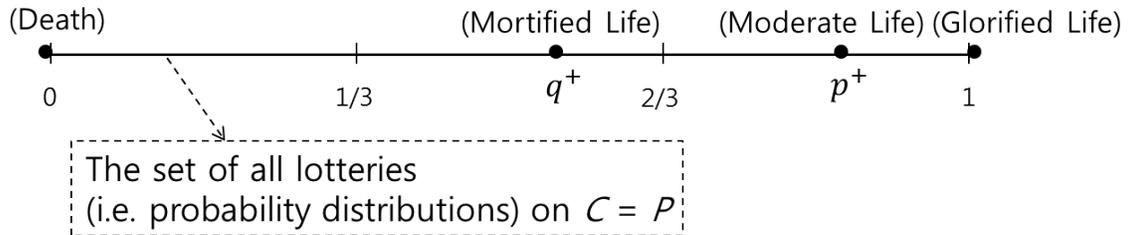
---

<sup>32</sup>Multiplying 2 to each side of (d) and connecting it with (b), we get:  $1 + q^G > 2p^G > 4q^G$ . From this, we get:  $q^G < \frac{1}{3}$ .

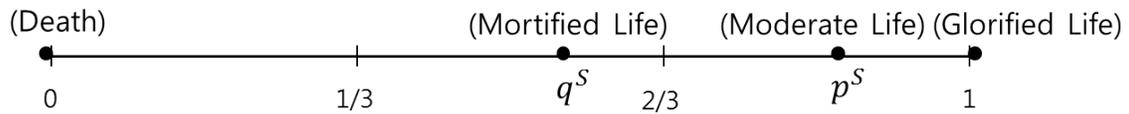
<sup>33</sup>From the fact that  $q^G < \frac{1}{3}$  and (b), we get:  $2p^G < 1 + q^G < \frac{4}{3}$ . From this, we get:  $p^G < \frac{2}{3}$ .

Figure 5.4: Cardinal Preferences of Each Type of Bob

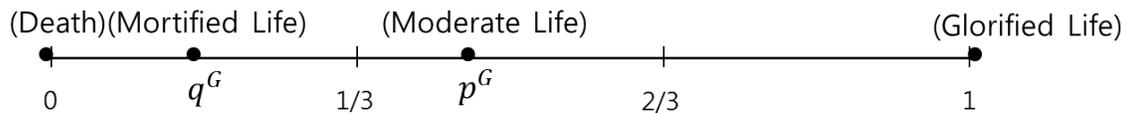
**[ $Bob_+$ 's (Cardinal) Preferences]**



**[ $Bob_S$ 's (Cardinal) Preferences]**



**[ $Bob_G$ 's (Cardinal) Preferences]**



We can see that for  $Bob_+$  and  $Bob_S$ , among the four consequences in  $C$ , the distance between Mortified Life and Death is the greatest, the distance between Moderate Life and Mortified Life is next, and the distance between Glorified Life and Moderate Life is the shortest within their respective preference-ordering. For  $Bob_G$ , the distance between Glorified Life and Moderate Life is the greatest, and the distance between Mortified Life and Death is the shortest within his particular preference-ordering. The relative distances between two consequences correspond to each type of Bob's utilities, which, in turn, indicates how much moving up from one consequence to another is subjectively worth for that particular type of Bob.

We now state our main theorem for this section.

**[THEOREM (T4)] (Expected Utility Representation of Hobbes's Theory of Real Good):**

For all  $p, q \in P$ ,  $p$  is *really better* (or *substantially better*) than  $q$  for  $Bob_i$  ( $i = S, G$ ) if and only if  $U_{Bob_+}(p) > U_{Bob_+}(q)$  if and only if  $\sum_{x \in C} p(x)U_{Bob_+}(x) > \sum_{x \in C} q(x)U_{Bob_+}(x)$ . That is, lottery  $p$  is really better (or substantially better) than another lottery  $q$  for  $Bob_i$  ( $i = S, G$ ) if and only if the expected utility of option  $p$  for  $Bob_+$  is greater than the expected utility of option  $q$  for  $Bob_+$ .

34

We are also able to derive some interesting results about the rationality of  $Bob_S$ 's and  $Bob_G$ 's preferences as well. I will not repeat the arguments, but we can give a similar account (as we did at the end of section 7.2.1) of what it means for the preferences of  $Bob_S$  and  $Bob_G$  to be substantially rational in our current expected utility representation framework.

**[PROPOSITION (P4)]:** It is substantially rational for  $Bob_i$  ( $i = S, G$ ) to strictly prefer  $p$  to  $q$  if and only if  $p \succ_{Bob_+} q$  if and only if  $\sum_{x \in C} p(x)U_{Bob_+}(x) > \sum_{x \in C} q(x)U_{Bob_+}(x)$ .

That is, it is substantially rational to prefer  $p$  to  $q$  if and only if the expected utility for the idealized-Bob of  $p$  is greater than the expected utility for the idealized-Bob of  $q$ .

In the framework of ordinal representation, all three types of  $Bob_i$ s were substantially rational. We can easily see that this is not the case in our current expected utility representation framework.

Consider the lottery  $[a \cdot \text{Glorified Life} + (1 - a) \cdot \text{Death}]$  where  $p^G < a < \frac{2}{3}$ . The lottery is basically a lottery that gives you less than 2/3 chance of obtaining a Glorified Life and more than 1/3 chance of Death. As one can see, the lottery is a pretty risky gamble to take.

---

<sup>34</sup>**Proof.** Again, this can be proved by T2 (which states that what is really good for a given individual is the satisfaction of the preferences formed by the individual's idealized-self on behalf of the individual's actual-self), T-3-1 (which states that as long as each type of  $Bob_i$  satisfies axioms A-3-1 to A-3-4, there exists an expected utility function  $U_{Bob_i}$  such that  $Bob_i$  strictly prefers  $x$  to  $y$  if and only if the expected utility of  $x$  is greater than the expected utility of  $y$ ), A-3-1 (i.e. each type of  $Bob_i$ 's strict preference relation is both asymmetric and negatively transitive), A-3-2 (i.e. compound lotteries can be reduced to simple lotteries), A-3-3 (i.e. Independence), A-3-4 (i.e. Archimedean) and by our assumption that  $Bob_+$  is the idealized-self of  $Bob_i$  ( $i = S, G$ ). (I will omit the details.)  $\square$

Given a choice between this lottery and securing a Moderate Life for sure,  $Bob_G$  would gladly take the lottery rather than securing a Moderated Life. This is because the expected utility of the lottery (which is  $a$ ), is greater than the expected utility of a Moderate Life (which is  $p^G$ ) for  $Bob_G$ . Indeed, in such cases, we can say that  $Bob_G$  is *mad* (in Hobbes's sense); he is too much infatuated by a basic passion for glory! In short,  $Bob_G$ 's preferences are *substantially irrational*.

Our framework captures this fact very nicely. According to P4, strictly preferring the lottery [ $a$ ·Glorified Life +  $(1 - a)$ ·Death] to securing a Moderate Life would be substantially rational if and only if  $Bob_+$  strictly prefers the lottery to securing a Moderate Life, which happens if and only if the expected utility of the lottery [ $a$ ·Glorified Life +  $(1 - a)$ ·Death] for  $Bob_+$  is greater than the expected utility of securing a Moderate Life.

As we can see, this is not the case. For  $Bob_+$ , the expected utility of the lottery, which is again  $a$ , is lesser than the expected utility of Moderate Life, which, for  $Bob_+$ , is  $p^+$  where  $p^+ > \frac{2}{3}$ . So, given the two options, the idealized-version of Bob would prefer to secure a Moderate Life for sure rather than to choose the lottery.

This means that in this situation, the glory-hungry version of the actual-Bob,  $Bob_G$ , will be preferring to do what his idealized-self would advice him not to do. And this, according to our current framework, indicates that  $Bob_G$ 's preferences are substantially irrational in the Hobbesian sense.

We can confirm that in choosing between lotteries, the self-preserving version of the actual-Bob,  $Bob_S$ 's preferences will always be substantially rational. This is because, for any lottery  $p \in P$ , the expected utility of  $p$  will be identical for both  $Bob_S$  and  $Bob_+$ . This comes from the fact that both versions of Bob are motivated by the right kind of basic passion (i.e. the basic passion for self-preservation) so that when they are provided with the objective probabilities of each option they are able to assess its merits in the same way.

This relates back to our working definition of substantial (Hobbesian) rationality (C1) which states: one's preferences are (substantially) rational if and only if they are (a) well-

considered and (b) well-balanced. One's preferences are well-considered when they are based on sufficient information about the relevant facts about each option. In our current framework, this, most importantly, implies that one's estimations of probabilities for each of the consequences in the options that one is facing are correct. However, in our current framework, the objective probabilities of  $p \in P$  is already given.

This means that, in our current framework, each type of *Bob<sub>i</sub>'s* preferences are guaranteed to be well-considered. Therefore, if any type of *Bob<sub>i</sub>'s* preferences are substantially irrational in the Hobbesian sense, this implies that such preferences were irrational mainly because they were unbalanced; that is, they were generated when Bob was primarily being influenced by the wrong kind of basic passion. This is exactly the case for *Bob<sub>G</sub>*, who is primarily being influenced by a basic passion for glory, which, according to Hobbes, is one of the major forms of madness.<sup>35</sup> So, our current reconstruction fits very well with Hobbes's main text.

I would like to end this section by mentioning the formal (or minimal) rationality of the three types of Bobs. Again, D2 claims that one's preferences are formally (or minimally) rational if and only if they satisfy the axioms of decision theory. In our current framework, which aims to provide an expected utility representation, the axioms of decision theory are the four axioms A-3-1 to A-3-4. We assumed that all three types of Bob satisfy these four axioms. Therefore, according to D2, all three types of Bobs are formally (or minimally) rational. This applies to even *Bob<sub>G</sub>* who is substantially irrational in the Hobbesian sense.

### **5.3.3 A Note on One's Attitudes Towards Risk**

Before I move on I would like to say something about one's attitude towards risk. I have explained that when one is the type of person who tries to maximize the expectation of some external value, this requires one to be risk-neutral about the external value in question. We

---

<sup>35</sup>“The passion whose violence or continuance maketh madness is either great vain-glory, which is commonly called *pride* and *self-conceit*, or great dejection of mind.”Hobbes [1994, *Leviathan*: Chapter VIII, Section 18]

have seen that all three types of *Bobs* are expected utility maximizers of their respective utility functions. This implies that each type of  $Bob_i$  is *risk-neutral* about the values that his particular utility function  $U_{Bob_i}$  produces. However, this does not imply that all three types of *Bobs* are risk-neutral in general.

The fact that  $Bob_+$  and  $Bob_S$  will be indifferent between a lottery [ $p \cdot$ Glorified Life +  $(1 - p) \cdot$ Death] and the sure outcome Moderate Life only when the value of  $p$  is very high (i.e. close to 1) indicates that both  $Bob_+$  and  $Bob_S$  are generally very risk-averse towards variances in life. Then, what exactly happened here? How could a *risk-averse* person, such as  $Bob_+$  and  $Bob_S$ , be seen as an expected utility maximizer, which implies that the person is *risk-neutral* about utility?

The answer is that when we construct the respective utility functions  $U_{Bob_+}$ ,  $U_{Bob_S}$ , and  $U_{Bob_G}$  by trying to find the value of  $p$  that would make each type of  $Bob_i$  indifferent between the lottery [ $p \cdot$ Glorified Life +  $(1 - p) \cdot$ Death] and the sure outcome Moderate Life, each type of  $Bob_i$ 's attitude towards risk became naturally incorporated into the construction of the utility functions themselves during the process.

Consider a very rash and risk-loving individual  $Bob_G$  who would strictly prefer to play the lottery [ $p \cdot$ Glorified Life +  $(1 - p) \cdot$ Death] to securing a Moderate Life for sure for values of  $p$  that are lesser than  $2/3$ . We have seen that  $Bob_G$  is indifferent to the two options when  $p$  is  $p^G$  ( $p^G > \frac{2}{3}$ ), which makes  $U_{Bob_G}(Moderate\ Life) = p^G$ . That is, the utility of a Moderate Life for  $Bob_G$  would be  $p^G$ , which will be much less than the utility of Moderate Life for  $Bob_+$  and  $Bob_S$  (i.e.  $U_{Bob_{+,S}}(Moderate\ Life) = p^{+,S}$ ), which, as we have seen, is at the very least greater than  $2/3$ , and probably very close to 1. That is,  $U_{Bob_G}(Moderate\ Life) = p^G < U_{Bob_{+,S}}(Moderate\ Life) = p^{+,S}$ .

Generally speaking, when we construct expected utility functions for different individuals, it is very likely that different utility numbers will be assigned to the same sure-outcomes for different individuals' utility functions depending on each individual's attitude towards risk. The utility numbers are derived from each individual's preferences towards risky gam-

bles, and since the preferences of each individual towards risky gambles reflect his or her attitude towards risk, the utility numbers that are assigned by the construction process will naturally reflect the individual's attitudes towards risk as well.

For instance, the fact that  $U_{Bob_G}(Moderate\ Life) = p^G < U_{Bob_{+,S}}(Moderate\ Life) = p^{+,S}$  indicates that  $Bob_G$  will gladly choose the lottery  $[a \cdot \text{Glorified Life} + (1 - a) \cdot \text{Death}]$  ( $a \in [0, 1]$ ) over securing a Moderate Life for much lesser values of  $a$  than  $Bob_{+,S}$ , and this, in turn, indicates that  $Bob_G$  is much more risk-loving than  $Bob_{+,S}$  and  $Bob_S$ , and that both  $Bob_{+,S}$  and  $Bob_S$  are much more risk-averse than  $Bob_G$ . However, once the utility numbers are assigned by this construction process, all three types of  $Bob$ 's will be risk-neutral towards the values that are generated by their respective utility functions.

The way that people's attitude towards risk gets incorporated into their utility functions can be illustrated in a more instructive way when we look at people's preferences towards some predetermined quantified value; such as money.

As we have seen from the insurance example introduced in section 7.3.1, many people are *not* expected monetary-value maximizers. And there aren't any good reasons why they even *should be*. For, as we have already seen, requiring that one should maximize the expectation of monetary value would basically imply that one should be *risk-neutral* about money, and there are many situations where it makes perfect sense for one to be risk-averse (or even risk-loving.) about money. This is basically why people in general buy insurance.

Then, what does it exactly mean for one to risk-averse, risk-neutral, risk-loving about some standard of value, e.g. money? Basically, we can identify a person's attitude towards risk by observing how the person feels about the *variances* of different kinds of lotteries that contain prizes measured by the standard of value in question.

Let  $[a(p) + b(1 - p)]$  denote a gamble that gives prize  $a$  with probability  $p$  and prize  $b$  with probability  $1 - p$ . Now, consider the following series of gambles:  $[\$0(\frac{1}{2}) + \$0(\frac{1}{2})]$ ,  $[\$100(\frac{1}{2}) - \$100(\frac{1}{2})]$ ,  $[\$10,000(\frac{1}{2}) - \$10,000(\frac{1}{2})]$ ,  $[\$1,000,000(\frac{1}{2}) - \$1,000,000(\frac{1}{2})]$ . The expectation of all four gambles is the same: it is \$0. What they differ is in their *variance*.

I expect that most people will not feel exactly the same way towards all four gambles. Take the gamble  $[\$0(\frac{1}{2}) + \$0(\frac{1}{2})]$  as our reference point. This gamble give you \$0 for sure; it involves no risk. If one happens to prefer this sure outcome to any of the other gambles that gives the same expectation but involves risk, then one is said to be *risk-averse*. The underlying intuition is that given that the expectation is the same, what one hates is the variance that is involved in all of the other gambles and this is what determined one's preference between the sure outcome and all of the other gambles.

Conversely, if one happens to prefer the other gambles that involve risk to the risk-free sure-outcome, then one is said to be *risk-loving*. Again, the underlying intuition is that what made one prefer the riskier alternatives is the fact that one likes variance; that is, the fact that one can earn a lot of money if things go well means more than the fact that one can lose a lot of money if things go badly.

However, if one happens to be indifferent to all four gambles, then this implies that one is *risk-neutral* towards all of the monetary values listed above. The underlying intuition is that variance does not matter to him or her.<sup>36</sup>

Generally speaking, when we want to know somebody's attitude towards risk, we look at the person's preference over: (a) getting the expectation of the gamble (i.e.  $a_1p_1 + a_2p_2 + \dots + a_np_n$ . remember that this is a determinate value, not a gamble) for sure, and (b) playing the gamble (i.e.  $[a_1(p_1) + a_2(p_2) + a_n(p_n)]$ ) that involves risks.

One is *risk-averse* if and only if one strictly prefers getting the expectation of the gamble for sure to the gamble (i.e.  $a_1p_1 + a_2p_2 + \dots + a_np_n \succ [a_1(p_1) + a(p_2) + a(p_n)]$ ); one is *risk-neutral* if and only if one is indifferent between getting the expectation of the gamble for sure and the gamble (i.e.  $a_1p_1 + a_2p_2 + \dots + a_np_n \sim [a_1(p_1) + a(p_2) + a_n(p_n)]$ ); one is *risk-loving*

---

<sup>36</sup>Note that it is possible for one to be risk-loving for some monetary values, but is, at the same time, risk-neutral or risk-averse to others. For example, one might prefer  $[\$100(\frac{1}{2}) - \$100(\frac{1}{2})]$  to  $[\$0(\frac{1}{2}) + \$0(\frac{1}{2})]$ , but may be indifferent between  $[\$10,000(\frac{1}{2}) - \$10,000(\frac{1}{2})]$  and  $[\$0(\frac{1}{2}) + \$0(\frac{1}{2})]$ , and prefer  $[\$0(\frac{1}{2}) + \$0(\frac{1}{2})]$  to  $[\$1,000,000(\frac{1}{2}) - \$1,000,000(\frac{1}{2})]$ . This would imply that one is risk-loving when there is not that much money involved, but becomes risk-neutral for \$10,000, and then becomes risk-averse for monetary values exceeding \$10,000.

if and only if one prefers the gamble to the sure outcome (i.e.  $a_1p_1 + a_2p_2 + \dots + a_np_n \succ [a_1(p_1) + a(p_2) + a(p_n)].$ )

We have seen that a utility function ( $U$ ) is a measure that represents person's preferences. Applying this to one's various attitudes towards risk that we have summarized above, we get: one is *risk-averse* if and only if the utility of the expectation is greater than the expected utility of the gamble (i.e.  $U(a_1p_1 + a_2p_2 + \dots + a_np_n) > p_1U(a_1) + p_2U(a_2) + \dots + p_nU(a_n)$ ); one is *risk-neutral* if and only if the utility of the expectation is equal to the expected utility of the gamble (i.e.  $U(a_1p_1 + a_2p_2 + \dots + a_np_n) = p_1U(a_1) + p_2U(a_2) + \dots + p_nU(a_n)$ ); one is *risk-loving* if and only if the utility of the expectation is lesser than the expected utility of the gamble (i.e.  $U(a_1p_1 + a_2p_2 + \dots + a_np_n) < p_1U(a_1) + p_2U(a_2) + \dots + p_nU(a_n)$ .)

It is convenient to represent these results in a graph. Let  $U_1$  be a utility function of a risk-averse individual,  $U_2$  be a utility function of a risk-loving individual, and  $U_3$  be a utility function of a risk-neutral individual. Then, each individual's utility function will respectively have the following general form:

Figure 5.5: Risk-Averse

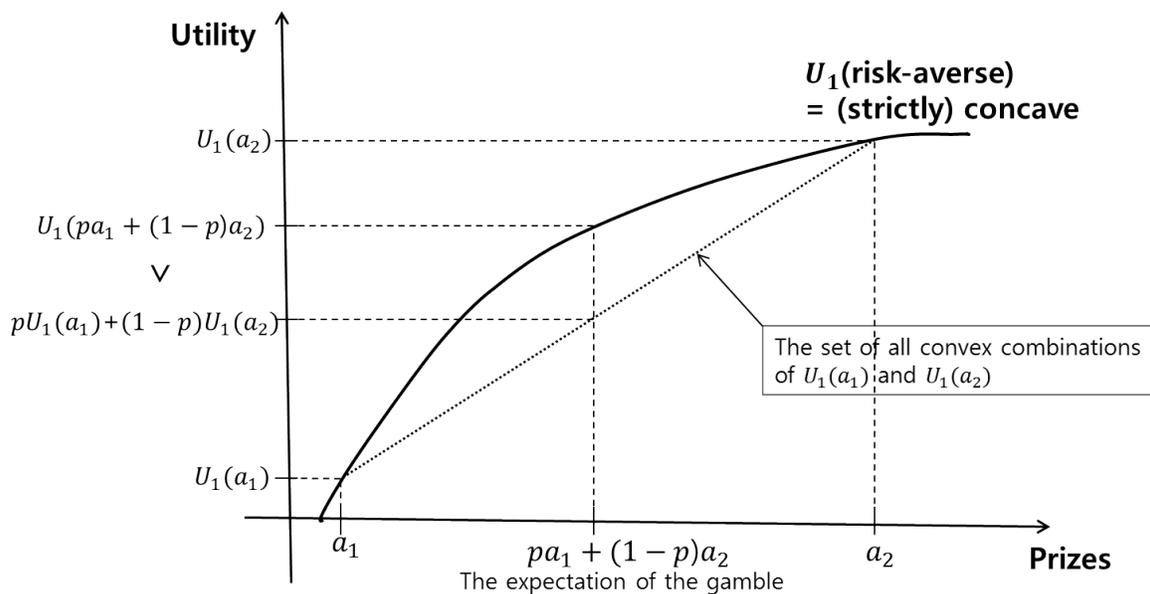


Figure 5.6: Risk-Loving

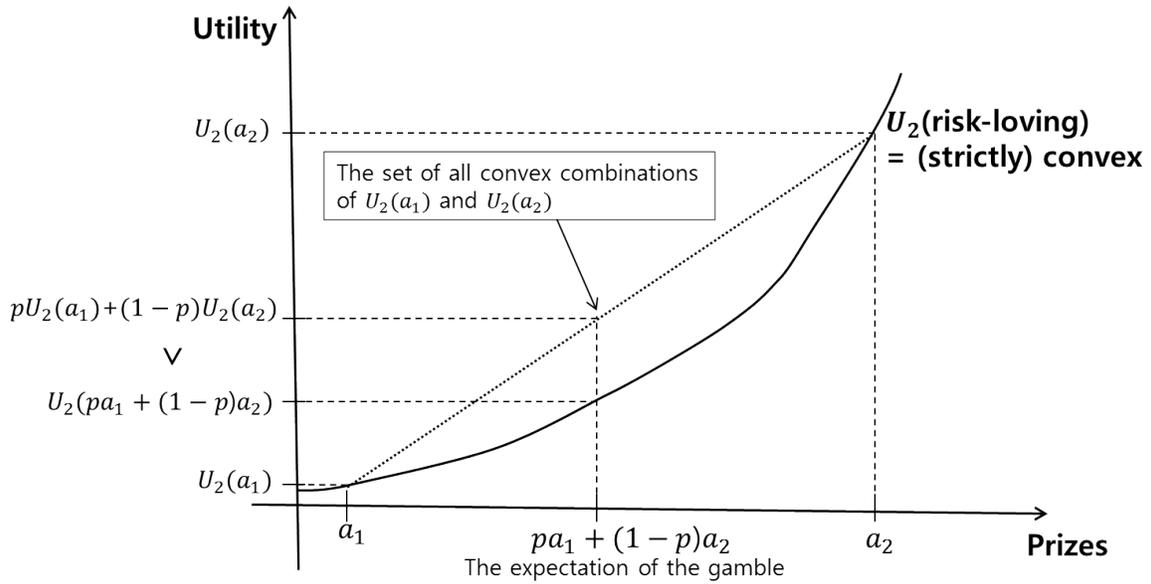
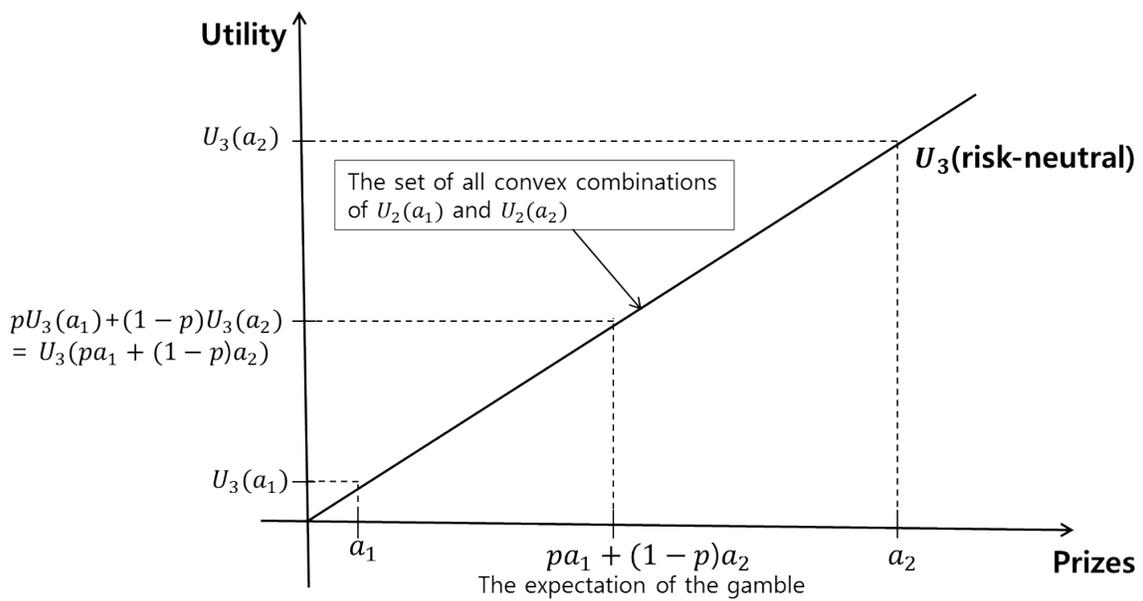


Figure 5.7: Risk-Neutral



So, when we construct utility functions for various people, we can say that each individual's attitude towards risk basically gets incorporated into *the curvature* of his or her utility function. An individual is *risk-averse* (towards the value in question) if and only if his or her utility function is *strictly concave*; An individual is *risk-loving* (towards the value in question) if and only if his or her utility function is *strictly convex*; and an individual is *risk-neutral* (towards the value in question) if and only if his or her utility function is *linear*.

However, once an individual's attitudes towards risk get incorporated into the curvature of his or her utility function, the individual's utility function for his or her utilities is guaranteed to be *linear*. This means that the individual is risk-neutral towards his or her utilities, which is required for the individual to be an expected utility maximizer just as expected utility theory claims him or her to be.

#### **5.3.4 *An Expected Utility Representation (Savage's Framework)***

In section 7.3.2, all three types of Bobs were required to express their preference towards lotteries in  $P$  where the objective probabilities of each lottery was already predetermined and provided. However, this is not what usually happens in real-life situations. We are never provided with explicit lotteries to choose from. We almost never calculate probabilities and there are no utility numbers written inside our heads that would make it possible for us to calculate the expected utility of any given option. Rather, we simply choose among the available actions open to us, and depending on the state of the world, our actions generate a specific consequence.

This would be the situation that the actual Bobs (i.e.  $Bob_S$  and  $Bob_G$ ) would be facing in the state of nature. That is, not only would they have no concept of utilities, but there will also be no objective probability distributions given to them when they are trying to decide how to act. What these people know are merely the fact that certain types of actions are available to them when they are making decisions and that these actions will generate a

certain consequence. They have preferences towards these available actions. And from these preferences we would need to extract probabilities and utilities that would let us explain why these people are acting in the way they do, and enable us to predict the future behaviors of these people.

In order to model the choice behaviors of decision makers in such situations we would need to use the (subjective) expected utility theory of Leonard Savage.<sup>37</sup> And the main gist of Savage's expected utility theory is that as long as people's preferences towards actions meet a certain set of conditions, then we are able to interpret their actions *as if* they had probabilities on states, utilities on outcomes, and that they were aiming to maximize expected utility.

We begin by describing the main framework of our current model. Here, is a list of entities that inhabit our current model:

- The set of consequences:  $C = \{\text{Death, Mortified Life, Moderate Life, Glorified Life}\}$
- The set of all states of the world:  $W$
- The set of all events (i.e. subsets of  $W$ ):  $E$
- The set of all acts:  $A$

Let me briefly explain each of these sets. Again, we have the set of consequences  $C$  which is the same set as in our earlier models. A state  $w \in W$  is a complete description of a particular state of the world; one may think of a single state  $w \in W$  as a conjunction of almost infinitely many propositions that describe all of the facts instantiated in a particular state of the world. Any subset  $a \subseteq W$  will be called an *event*. We can basically think of an event as a collection of several states. The set of all events, that is, the set of all subsets of  $W$  will be denoted by  $E$ . In our current framework (which mimics that of Savage), an action  $a$  is regarded as a *function* from the set of states  $W$  into the set of consequences  $C$ : that is,  $a : W \rightarrow C$ . The set of all such functions (i.e. acts) is denoted by  $A$ .

---

<sup>37</sup>See Savage [1972]

I believe that it would be easier to understand what is going on by looking at a specific example. Suppose that you are hungry. You go to your fridge. You discover a plate of salmon rolls that you bought yesterday morning and have put in the fridge, and a turkey sandwich that you bought today. You strongly prefer salmon rolls to a turkey sandwich given that both are fresh. However, the salmon rolls might have gone bad. You know that the turkey sandwich is fresh. How would we model this type of situation which respects the general spirit of Savage’s expected utility theory? Here is one such model.

- $C = \{\text{Feel Very Satisfied}(c_1), \text{Feel Barely Satisfied}(c_2), \text{Get Food Poisoned}(c_3)\}$
- $W = \{\text{The Salmon Rolls Got Bad}(w_1), \text{The Salmon Rolls Are Still Fresh}(w_2)\}$
- $A = \{\text{Eat Salmon Roll}(a_1), \text{Eat Turkey Sandwich}(a_2)\}$
- $c_1 \succ c_2 \succ c_3$
- $a_1(w_1) = c_3, a_1(w_2) = c_1$
- $a_2(w_1) = c_2, a_2(w_2) = c_2$

Here, we can see that the action “Eat Salmon Roll( $a_1$ )” is a function that generates the consequence “Get Food Poisoned( $c_3$ )” as its value when it takes the state “The Salmon Rolls Got Bad( $w_1$ )” as its argument, and generates the consequence “Feel Very Satisfied( $c_1$ )” as its value when it takes the state “The Salmon Rolls Are Still Fresh( $w_2$ ).” Similarly, the action “Eat Turkey Sandwich( $a_2$ )” is a function that generates the consequence “Feel Barely Satisfied( $c_2$ )” as its value regardless of which state it takes as its argument.

Now, there are some important things that one needs to be careful about when one tries to model this type of situation in this way.<sup>38</sup>

First of all, in Savage’s expected utility theory, *states are assumed to be independent of acts*. What this means is that performing a particular action should not influence the likelihoods of any particular states from realizing. Suppose that one was trying to model the arms

---

<sup>38</sup>Most of the things that I described below are inspired by Joe Halpern’s class notes on Decision Theory.

race between the US and the USSR during the cold war. Suppose that one was thinking from the US perspective and modeled the situation in the following way.

Table 5.3: The Arms Race

Acts \ States	War	Peace
Arm	Mutual Annihilation	Status Quo
Disarm	Become Communist State	Improved Society

Suppose that the preferences of the US is as follows: Improved Society  $\succ$  the Status Quo  $\succ$  Become Communist State  $\succ$  Mutual Annihilation. Described in this way, it seems obvious how the US should act; it should “Disarm.” For the act “Disarm” *dominates* the act “Arm”; that is, the act “Disarm” generates a more preferred consequence regardless of what particular state of the world instantiates. However, this might not be the correct prescription because the fact that the US unilaterally disarms might make it much more likelier for the USSR to initiate a war.

In other words, what is wrong about this “US-USSR Arms Race Model” is that the states are not described in a way that they are independent of the acts. In our “What Should I Eat?” model, the likelihood of whether or not the salmon rolls had gone bad does not seem to depend on whether one chooses to eat the salmon rolls or whether one choose to eat the turkey sandwich. So, our “What Should I Eat?” model seems to meet the requirement that the states should be independent of the acts.<sup>39</sup>

Second, the states should be described in detail enough for the acts to be properly regarded as a *function*. Suppose that the salmon rolls really did get bad. However, suppose that the strength of the germs and bacteria in the bad salmon rolls are weak enough for some-

---

<sup>39</sup>The widely known “Newcombe’s Problem” is primarily concerned about the issue of whether or not states should be interpreted as independent of acts in that particular setting.

body who has an extraordinarily strong stomach to be immune from suffering food poison by taking them. If this is the case, eating the salmon rolls when they are bad will not generate a single-valued consequence. That is, it might be either  $a_1(w_1) = c_3$  (get food poisoned) or  $a_1(w_1) = c_1$  (feel very satisfied), which would make the act of eating the salmon rolls (i.e.  $a_1$ ) no longer a function.

There are two basic ways to rectify this. One way is to enrich the state space  $W$  in such a way that the individual acts do turn out to be a function. So, instead of  $S = \{\text{The Salmon Rolls Got Bad}(w_1), \text{The Salmon Rolls Are Still Fresh}(w_2)\}$ , we might take

- $S^* = \{\text{The Salmon Rolls Got Bad and You Have a Strong Stomach}(w_1), \text{The Salmon Rolls Got Bad and You Have a Weak Stomach}(w_2), \text{The Salmon Rolls Are Still Fresh}(w_3)\}$

as our state space.

Another way to rectify the problem is to say that the act, Eat Salmon Roll( $a_1$ ), generates, *not* a single consequence in  $C$ , but a *probability distribution* over  $c_1$  and  $c_3$ , when it takes the state “The Salmon Rolls Got Bad( $w_1$ )” as its argument. In that case,  $a_1(w_1)$  will take the probability distribution  $p \cdot c_1 + (1 - p) \cdot c_3$  as its value (where  $p$  denotes the probability that you have a strong stomach) instead of a single consequence  $c_3$ . Any of these two ways will solve the problem.

Third, it is important in our current framework, (which follows that of Savage’s expected utility theory), that the set of states  $W$  is *infinite*. And how  $W$  can be properly seen as a set that has infinitely many elements needs some explanation.

Remember that a single state must include a *complete description* of a particular state of the world. I have been speaking a little sloppily by explaining that  $w_1$ ,  $w_2$ , and  $w_3$  are individual states that are members of  $W$ ; but, technically speaking,  $w_1$ ,  $w_2$ , and  $w_3$  do not qualify as individual states because they do not include a complete description of a particular state of the world. In this sense, it might be better to think of  $w_1$ ,  $w_2$ , and  $w_3$  as *events*, (that is, a *set of states*), which each share a common proposition that describes the world.

Now, it can be easily shown that a single event  $e$  (which is a subset of  $W$ ) can be divided into arbitrarily many events as one wishes. Consider the event  $w_1$  “The Salmon Rolls Got Bad *and* You Have a Strong Stomach.” This event can be arbitrarily divided into the event  $w_{1.1}$  “The Salmon Rolls Got Bad *and* You Have a Strong Stomach *and* Your Next Door Neighbor is Wearing a red Sweater” and the event  $w_{1.2}$  “The Salmon Rolls Got Bad *and* You Have a Strong Stomach *and* Your Next Door Neighbor is Wearing a blue Sweater.” Each of the event can further be arbitrarily divided into  $n$  events (for any number  $n$ ); for example, let  $w_{1.1.n}$  be the event “The Salmon Rolls Got Bad *and* You Have a Strong Stomach *and* Your Next Door Neighbor is Wearing a red Sweater *and* there are exactly  $n$  grains of sand in Waikiki Beach.”

So, by arbitrarily dividing individual events in this way, we can, in effect, partition the set of states  $W$  into any number of  $n$ -parts<sup>40</sup> as we wish. As we will soon see, Savage is going to require us to divide  $W$  into  $n$  parts for any number  $n$  in order to derive the main representation theorem. And such process can be legitimately performed only when the set of states  $W$  is infinite.

Now, let’s get back to our original framework. As I have already explained, unlike the Von-Neumann and Morgenstern’s framework which we used in the last section, within Savage’s framework, each type of  $Bob_i$  is not provided with objective probability distributions to choose from. Instead, they simply have preferences towards the available actions open to them (that is, they have preferences on the set of actions  $A$ ). Based on this information alone, we would have to, somehow, derive subjective probability distributions that each type of Bob (non-consciously) associates with each course of action that would make it possible for us to interpret their behaviors as *maximizing expected utility given their subjective probabilities*. And, this is the project that Savage’s expected utility theory aims to accomplish.

Again, we can see that the framework makes more practical sense; we are hardly ever

---

<sup>40</sup>Formally, a partition of a set  $X$  is a set of non-empty subsets of  $X$  such that: (a) the union of its members equals  $X$  (i.e. the individual partitions are *collectively exhaustive*), and (b) the intersection of any two members is empty (i.e. the individual parts are *mutually exclusive*.)

provided with objective probability distributions in real life choice situations. The framework is also more ambitious; it tries to accomplish the same thing as Von-Neumann and Morgenstern's expected utility theory even when predetermined objective probabilities are not given.

However, the cost for this is that the theory relies on more axioms - (it requires seven rather than four of what Savage calls "postulates") - and the process of deriving the main representation theorem is extremely complicated. Again, I will not go into the details of each step of the proof. Those who are really interested should read [Savage, 1972] or the last chapter of [Fishburn, 1970].<sup>41</sup> What is important for us is to understand the overall picture and the general strategy of the theory rather than the minute details of each step of the proof. My exposition of Savage's expected utility theory will follow the general flow of [Fishburn, 1970, Chapter 14].

Before we begin, we would first need to add some more preliminary definitions to be used in our framework.

**[DEFINITION(D-4-1)] (Conditional Preferences):** For  $i = S, G, +$  and  $f, g \in A, e \in E (e \subseteq W)$ ,  $f \succ_{Bob_i} g$  given  $e \equiv_{df} f' \succ_{Bob_i} g'$  whenever  $f = f'$  and  $g = g'$  on  $e$ , and  $f' = g'$  on  $e^c$

Let me explain what this is saying. The definition is trying to characterize what it means for a given type of Bob to strictly prefer act  $f$  to act  $g$  given that event  $e$  has occurred (i.e. conditional on event  $e$ .) According to the definition,  $Bob_i$  strictly prefers act  $f$  to act  $g$  conditional on event  $e$  if and only if: (i) there are two acts  $f'$  and  $g'$  that generate the same consequence outside of event  $e$ , (ii) act  $f'$  generates the same consequence as act  $f$  when event  $e$  occurs and act  $g'$  generates the same consequence as act  $g$  when event  $e$  occurs, and (iii)  $Bob_i$  strictly prefers act  $f'$  to act  $g'$ .

<sup>41</sup>Savage explains the proof in roughly 100 pages. Fishburn condenses it into less than 20 pages. However, it takes about the same amount of time to read through each.

**[DEFINITION(D-4-2)] (Null Event):** For  $i = S, G, +$  and  $f, g \in A$ ,  $e \in E$  ( $e \subseteq W$ ),  $e$  is *null*  
 $\equiv_{df} f \sim_{Bob_i} g$  whenever  $f = g$  on  $e^c$

The intuition behind this definition is that an event  $e$  is *null* if and only if one does not care at all about what happens when event  $e$  occurs. Suppose that there is a fair coin and somebody offers you the following two options to choose from: (a) the person pays you \$1 if the coin lands heads and you pay the person \$1 if the coin lands tails and *the person pays you* \$1,000,000 if the coin stays in mid-air for a 27 minutes and starts to break into exactly 39 pieces (call this event  $n$ ), and (b) the person pays you \$1 if the coin lands heads and you pay the person \$1 if the coin lands tails and *you pay the person* \$1,000,000 if event  $n$  occurs.

Presumably, many people will feel indifferent between the two options (a) and (b). Of course, the two options are different; with option (a), you would get paid \$1,000,000 when event  $n$  happens, and with option (b), you would have to pay \$1,000,000 when event  $n$  happens. But, presumably, it is impossible for event  $n$  to happen; it is a *null event*. So, what happens on event  $n$  does not influence your preference between the two options; you simply do not care!

However, if you happen to be the person who actually does care between option (a) and option (b) (say, you prefer (a) to (b)), then this just shows that you think that there actually is a possibility (no matter how slight) for event  $n$  to happen. In such case, event  $n$  will no longer be a null event *for you*. However, if you truly think that a certain event is *null* (i.e. it will not happen), then it seems plausible to assume that you would not care about what (whatever that is) happens on that event. And this is what the definition is saying.

**[DEFINITION(D-4-3)] (Constant Acts):** For  $i = S, G, +$  and  $f, g \in A$ ,  $x, y \in C$ ,  $x \succ_{Bob_i} y$   
 $\equiv_{df} f \succ_{Bob_i} g$  when  $f = x$  and  $g = y$  on  $W$

Remember that in our current framework, each type of Bob is assumed to have preferences towards the acts that are open to him. However, even if certain types of actions are not practically available to  $Bob_i$  at a particular moment, it is still possible for us to ask him

whether *he would prefer one hypothetical act to another hypothetical act if he were to have a choice*. And since  $A$  is the set of *all* actions, any kind of hypothetical act is a member of  $A$  as long as it is a function that maps states into consequences.

The definition is basically stating that there exists such acts that generate the same consequence in all possible states. And, that it is from  $Bob_i$ 's preferences from these “constant act” which we derive his preferences towards sure consequences, and *vice versa*. We know that, for each type of  $Bob_i$ ,  $\text{Glorified Life} \succ_{Bob_i} \text{Moderate Life} \succ_{Bob_i} \text{Mortified Life} \succ_{Bob_i} \text{Death}$ . So, if there are four acts,  $f, g, h, k$  such that  $f : W \rightarrow \{\text{Glorified Life}\}$ ,  $g : W \rightarrow \{\text{Moderate Life}\}$ ,  $h : W \rightarrow \{\text{Mortified Life}\}$ , and  $k : W \rightarrow \{\text{Death}\}$ , we know that  $f \succ_{Bob_i} g \succ_{Bob_i} h \succ_{Bob_i} k$  for all  $i$ .

**[DEFINITION(D-4-4)] (Subjective Likelihood):** For  $i = S, G, +$  and all  $f, g \in A$ ,  $e, d \in E$  ( $e, d \subseteq W$ ),  $e \succ_{Bob_i}^* d$  (read as: “ $Bob_i$  subjectively believes that event  $e$  is more likelier to happen than event  $d$ ”)  $\equiv_{df} f \succ_{Bob_i} g$  whenever  $x \succ_{Bob_i} y$ ,  $f = x$  on  $e$ ,  $f = y$  on  $e^c$ ,  $g = x$  on  $d$ ,  $g = y$  on  $d^c$ .

The definition explains how we are able to extract one’s subjective beliefs concerning the likelihoods of events from his or her preferences (which are revealed by his or her choice behaviors.) The basic thought is something like this.

Suppose that there are two sports teams, X and Y, who will compete against each other this following weekend. Suppose that you are trying to choose between the following two actions: (a) “Bet on X”: you win \$100 if X wins and lose \$100 if X loses, and (b) “Bet on Y”: you win \$100 if Y wins and lose \$100 if Y loses. Presumably, you would prefer to win \$100 rather to lose \$100. Suppose that think about the two options for a moment and realize that you strictly prefer act (a) (i.e. “Bet on X”) to act (b) (i.e. “Bet on Y”).

What does your preference between the two acts tell us about your subjective beliefs concerning which team has a higher chance of winning? Obviously, given that you want to win \$100 rather than to lose \$100, the fact that you strictly prefer to bet on X rather than

Y suggests that you *subjectively believe* that X has a higher chance of winning than Y. (Of course, your beliefs about the likelihoods of X winning might be vastly inaccurate.) This is the sense in which subjective probabilities are derived from one's preferences in Savage's framework. And, this is basically what definition D-4-4 is saying.

We are done with the preliminary definitions. We now state Savage's 7 axioms (which are more commonly known as Savage's 7 *postulates*):

For  $i = +, S, G$ , and  $f, f', g, g' \in A$ , and  $x, x', y, y' \in C$ , and  $e, d \subseteq W$  ( $e, d \in E$ ):

**[AXIOM(A-4-1)]:**  $\succ_{Bob_i}$  on  $A$  is *asymmetric* and *negatively transitive*.

Again,  $\succ_{Bob_i}$  is our old preference-relation; only, this time the relation is defined on the set of acts,  $A$ . As we have seen from D-4-3 (the definition for *constant acts*), each type of Bob's preferences on  $A$  induces each type of Bob's preferences on  $C$ . Once again, the other relations,  $\succsim_{Bob_i}$  (weak preference) and  $\sim_{Bob_i}$  (indifference) are defined in a similar way as in D-2-1 and D-2-2.

**[AXIOM(A-4-2)]:** Suppose  $f = f'$  and  $g = g'$  on  $e \subseteq W$ ,  $f = g$  and  $f' = g'$  on  $e^c \subseteq W$ .

Then,  $f \succ_{Bob_i} g$  if and only if  $f' \succ_{Bob_i} g'$ .

This reminds us with the "independence axiom" of Von-Neumann and Morgenstern's expected utility framework that we have seen in the previous section. What it is basically saying is that given that two acts generate exactly the same consequences (whatever they are) outside a specific event, each type of Bob's preferences between two acts are determined by what sort of consequences these two acts generate when the specific event in question does occur.

**[AXIOM(A-4-3)]:** Suppose event  $e \subseteq W$  is *not null*,  $f = x$  on  $e$  and  $g = y$  on  $e$ . Then,

$f \succ_{Bob_i} g$  given  $e$  if and only if  $x \succ_{Bob_i} y$ .

This axiom, along with axiom A-4-2, extends the independence aspect of each type of  $Bob'_i$ 's preferences towards *acts* to each type of  $Bob'_i$ 's preferences towards *consequences*.

Here is a specific example. Suppose that there are two acts: (a) that generates Moderate Life when event  $e$  happens and generates Glorified Life when event  $e$  does not happen, and (b) that generates Mortified Life when event  $e$  happens and generates Glorified Life when event  $e$  does not happen. All three types of Bobs strictly prefer a Moderate Life to a Mortified Life. Therefore, the axiom says that all three types of Bobs strictly prefer act (a) to act (b) *given event  $e$* . (By the definition of *conditional preferences* (i.e. D-4-1), we can actually just simply say that all three types of Bobs strictly prefer act (a) to act (b).)

Now, suppose that there are two other acts: (a') that generates Moderate Life when event  $e$  happens and generates Death when event  $e$  does not happen, and (b') that generates Mortified Life when event  $e$  happens and generates Death when event  $e$  does not happen. Both act (a') and (b') differ with act (a) and act (b) only in respect to what happens outside of event  $e$ . Then, the fact that all three types of Bobs strictly prefer a Moderate Life to a Mortified Life along with axiom A-4-3 and axiom A-4-2 jointly imply that all three types of Bobs strictly prefer act (a') to act (b') (*given event  $e$* ) just as they strictly prefer act (a) to act (b) (*given event  $e$* .)

**[AXIOM(A-4-4)]:**  $(x \succ_{Bob_i} y, f = x \text{ on } e \text{ and } f = y \text{ on } e^c, g = x \text{ on } d \text{ and } g = y \text{ on } d^c)$  and  
 $(x' \succ_{Bob_i} y', f' = x' \text{ on } e \text{ and } f' = y' \text{ on } e^c, g' = x' \text{ on } d \text{ and } g' = y' \text{ on } d^c)$  imply  
 $(f \succ_{Bob_i} g \Leftrightarrow f' \succ_{Bob_i} g')$

The notation is slightly complicated. But, the thought behind it is simple. Let's go back to the two sports team example. If you bet on team X, then you win \$100 if team X wins and lose \$100 if team X loses. If you bet on team Y, then you win \$100 if team Y wins and lose \$100 if team Y loses. Obviously, you prefer winning \$100 to losing \$100. Again, suppose that you prefer betting on team X rather than betting on team Y.

Now, suppose that somebody offers you the following two alternate bets. Again, you can bet either on X or Y. However, this time, if you bet on team X, then you win \$500 if team X wins and lose \$500 if team X loses, and if you bet on team Y, then you win \$500 if team Y

wins and lose \$500 if team Y loses.

What axiom A-4-4 is basically saying is that if you prefer winning \$500 to losing \$500, and if you previously preferred to bet on X rather than Y when the stakes were \$100, then your preference between the two bets should not change (that is, you should still prefer to bet on X rather than to bet on Y) even when the stakes have changed to \$500.

The intuition behind this is this. By definition D-4-4, the fact that you preferred to bet on X rather than on Y indicates that you *subjectively believe* that it is likelier for team X (rather than team Y) to win. If this is so, then axiom A-4-4 says that you should prefer any action that generates a better consequence when team X wins (whatever that happens to be) to any action that generates a better consequence when team Y wins (whatever that happens to be).

In other words, the axiom is basically saying that your subjective beliefs about the likelihoods of events should not be influenced by the particular consequences that are generated on those events. I think that this is at least a plausible assumption normatively speaking. If one subjectively thought that team X had a greater chance of beating team Y when the stakes were \$100, but, then changed his/her mind after the stakes when up to \$500 and thought that, now, team Y had a greater chance of beating team X, we would naturally think that he/she is not thinking consistently about probabilities.

**[AXIOM(A-4-5)]:**  $x \succ_{Bob_i} y$  for some  $x, y \in C$

What the axiom is saying is that that each type of Bob is not completely indifferent to all of the consequences in  $C$ . In our current framework, this axiom is already satisfied; since for all  $i = S, G, +$ , Glorified Life  $\succ_{Bob_i}$  Moderate Life  $\succ_{Bob_i}$  Mortified Life  $\succ_{Bob_i}$  Death.

The result of this axiom is that it prevents the whole set of states  $W$  from being a *null event*, which would make the framework very uninteresting. Furthermore, the axiom prevents the subjective likelihood relation (i.e.  $\succ_{Bob_i}^*$ ) of each type of Bob from being *reflexive*. For if it were the case that  $x \sim_{Bob_i} y$  for every  $x, y \in C$ , then this would make the antecedent of the definition of subjective likelihood (i.e. D-4-4) vacuously true, and thereby, render every

event  $e \subseteq W$  subjectively likelier than itself. This is not how we would want our subjective likelihood relation to behave. We would want our subjective likelihood relation to be both asymmetric and negatively transitive. And, the axiom prevents our subjective likelihood relation to behave in any weird ways that contradict our general intuition of probabilities. Again, the axiom is already satisfied in our current framework.

**[AXIOM(A-4-6)]:** Suppose  $f \succ_{Bob_i} g$  and  $x$  is any consequence of  $C$ . Then, there *exists* a finite partition of  $W$  such that, if  $e$  is any event in the partition, then  $(f' = x$  on  $e, f' = f$  on  $e^c$ ) implies  $f' \succ_{Bob_i} g$  and  $(g' = x$  on  $e, g' = g$  on  $e^c$ ) implies  $f \succ_{Bob_i} g'$ .

This is where our assumption that the set of states  $W$  is infinite comes to play. The fact that  $W$  is infinite means that we can arbitrarily partition  $W$  into any number of  $n$ -partitions. As  $n$  becomes greater and greater, the partition of  $W$  will become finer and finer, and, as a result, one's subjective beliefs concerning the likelihoods of the realization of any single partition of  $W$  (which is an event) will become smaller and smaller - that is, as  $n$  becomes arbitrarily large, one will start to believe that it is quite unlikely for any single partition (i.e. an event) to happen.

What the axiom is saying is that given that all three types of Bob strictly prefer (a) living a Moderate Life for sure to (b) living Mortified Life for sure, there *is*, for each type of Bob, a finite partition of  $W$  such that, between the two options, (a') living a Moderate Life except on one single partion of  $W$  where he would face Death and (b) living a Mortified Life for sure, the given type of Bob would *still strictly prefer* (a') to (b). Similarly, the axiom also claims that there *is*, for each type of Bob, a finite partition of  $W$  such that, between the two options, (a) living a Moderate Life for sure and (b') living a Mortified Life except on one single partion of  $W$  where he would live a Glorified Life, the given type of Bob would still strictly prefer (a) to (b').

For instance, suppose that (a') is the option of living a Moderate Life except on the event

that Angelina Jolie is a full-time professor of philosophy at Cornell University in the year 2013 and blinks her eyes exactly 13 times between 02/02/2013 2:36 PM and 02/02/2013 2:37 PM in which case one faces Death, and that (b) is the option of living a Mortified Life for sure. I believe that it is reasonable to assume that each type of Bob would strictly prefer (a') to (b) if they are given the two options. (If you have any doubts on this, we can always make the event that one faced death much less likelier than what it is right now.) The same thing holds for each type of Bob's preference between (a) and (b').

We can see that this axiom closely resembles the Archimedean axiom (i.e. A-3-4) that we have encountered in the Von-Neumann and Morgenstern framework in the previous section. That is, what it is virtually saying is that no consequence can possibly be so bad (or so good) such that having the most slightest chance of experiencing such consequence along with a given option would completely change one's preference towards such option to another.

These six axioms are actually all that we need for our current purpose. But, in Savage's original work, there is one more axiom that he presents:

**[AXIOM(A-4-7)]:**  $f \succ_{Bob_i} g(w)$  given  $e$ , for all  $w \in e$  implies  $f \succsim_{Bob_i} g$  given  $e$ . ( $f \prec_{Bob_i} g(w)$  given  $e$ , for all  $w \in e$  implies  $f \precsim_{Bob_i} g$  given  $e$ .)

This axiom is needed for the more general case where one wants to derive an expected utility representation theorem for probability distributions that assign positive probabilities on more than finitely many consequences. However, we can see that such cases do not arise in our current framework where the set of consequences  $C$  has only four elements; Death, Mortified Life, Moderate Life, and Glorified Life. So, it is guaranteed that any probability distribution on  $C$  is able to assign positive probabilities on only a finite number of consequences; which is, at most, four. So, for us, we can safely ignore axiom A-4-7 in our current setting.

So, we have gone through all of the axioms that we need to derive an expected utility representation theorem. Again, I will not go through the details of each step of the proof; it is a very long and complicated process which requires a lot of patience. However, it is

still worthwhile to understand how one would generally proceed to derive Savage's expected utility representation theorem.

Remember that each type of Bob in our current framework is not provided with predetermined objective probabilities; we simply infer each type of Bob's subjective beliefs about the likelihoods of particular events by observing each type of Bob's preferences towards available acts. As we have seen from definition D-4-4, each type of Bob's subjective beliefs concerning the likelihoods of particular events is represented by the binary relation  $\succ_{Bob_i}^*$  which is defined in terms of each type of  $Bob_i$ 's preferences towards acts.  $e \succ_{Bob_i}^* d$  would imply that  $Bob_i$  subjectively believes that event  $e$  is more likelier to happen than event  $d$ .

However, in order for us to derive an expected utility representation theorem, we would need a *probability measure* (i.e. a *probability distribution*) that generates a concrete number (i.e. a quantification) for each event in a way that correctly represents each type of  $Bob_i$ 's subjective likelihood relation for those events. That is, instead of just  $e \succ_{Bob_i}^* d$ , we would need something more concrete like  $p^{Bob_i}(e) = .73$  and  $p^{Bob_i}(d) = .16$ , where  $p^{Bob_i}$  is a probability distribution on  $(W, E)$ .

I have already mentioned this in passing in section 7.3.2, but, a probability distribution  $p^{Bob_i}$  on  $(W, E)$  is a function  $p^{Bob_i} : E \rightarrow [0, 1]$  such that: (i)  $p^{Bob_i}(W) = 1$  and (ii) for all  $e, d \subseteq W$  (i.e.  $e, d \in E$ ), if  $e \cap d = \emptyset$ , then  $p^{Bob_i}(e \cup d) = p^{Bob_i}(e) + p^{Bob_i}(d)$ . We can easily confirm that (i) and (ii) imply: (iii)  $p^{Bob_i}(e^c) = 1 - p^{Bob_i}(e)$ .

So, what we want is for each type of  $Bob_i$ 's subjective likelihood relation  $\succ_{Bob_i}^*$  defined on  $E$  to be refined enough for there to be a probability distribution  $p^{Bob_i}$  representing it. So, what are the conditions that each type of  $Bob_i$ 's subjective likelihood relation  $\succ_{Bob_i}^*$  need to satisfy in order for there to exist a probability distribution  $p^{Bob_i}$  representing it? Basically, there are five conditions that are jointly necessary and sufficient for there to exist a probability distribution  $p^{Bob_i}$  representing  $Bob_i$ 's subjective likelihood relation  $\succ_{Bob_i}^*$  on  $E$ . And they are:

**[Conditions for there to exist a probability distribution on  $(W, E)$ ]**

For all  $c, d, e \subseteq W$  (i.e.  $c, d, e \in E$ )

(A)  $\succ_{Bob_i}^*$  is *asymmetric* and *negatively transitive*.

(B)  $\phi \lesssim_{Bob_i}^* e$  (where  $\lesssim_{Bob_i}^* \equiv_{df} \not\succeq_{Bob_i}^*$ )

(C)  $W \succ_{Bob_i}^* \phi$

(D) If  $d \cap c = e \cap c = \phi$ , then  $[d \succ_{Bob_i}^* e \Leftrightarrow (d \cup c) \succ_{Bob_i}^* (e \cup c)]$ <sup>42</sup>

(E) Suppose  $d \succ_{Bob_i}^* e$ . Then, there is a finite partition  $\{h_1, \dots, h_n\}$  of  $W$  such that  $d \succ_{Bob_i}^* e \cup h_k$  for every  $k = 1, \dots, n$ .

The five conditions from (A) to (E) are jointly necessary and sufficient for there to be a probability distribution  $p^{Bob_i}$ . The first important step in the proof of deriving the expected utility representation theorem in Savage's framework is to show that the seven axioms from A-4-1 to A-4-7 (actually, just the six axioms from A-4-1 to A-4-6) logically imply the five conditions (A) to (E) stated above. As it turns out, Savage's seven axioms (actually, the first six axioms) do imply these five conditions (From (A) to (E)) that are necessary and sufficient for there to be a probability distribution  $p^{Bob_i}$  on  $(W, E)$ . From this, we are able to derive the following lemma.

**[LEMMA(A-4-1)] (*Deriving Probabilities from Preferences*):** Suppose that each type of

$Bob_i$ 's preferences on  $A$  (i.e. the set of acts) satisfy the seven axioms from A-4-1 to A-4-7. Then, for each type of  $Bob_i$ , there exists a probability distribution  $p^{Bob_i}$  on  $(W, E)$  such that:

(a) For all  $d, e \in E$ ,  $d \succ_{Bob_i}^* e$  if and only if  $p^{Bob_i}(d) > p^{Bob_i}(e)$ .

(b) For all  $e \in E$  and  $0 \leq r \leq 1$ , there exists a  $d \subseteq e$  such that  $p^{Bob_i}(d) = rp^{Bob_i}(e)$ .

---

<sup>42</sup>The four conditions from (A) to (D) are enough to make the binary relation  $\succ_{Bob_i}^*$  on  $E$  (what is known as) a "qualitative probability." We can see that the four conditions from (A) to (D) summarizes our general intuition of what probability is. However, it has been proved that these four conditions are insufficient for there to be a "quantitative probability" (i.e. a probability distribution) on  $E$ .

(c) Moreover, for each type of  $Bob_i$ , the probability distribution  $p^{Bob_i}$  on  $(W,E)$  is *unique*.

What we have just done is this. We started with each type of  $Bob_i$ 's preferences on acts, and, from this, we derived each type of  $Bob_i$ 's subjective beliefs concerning the likelihoods of various events. We, then, showed that given that each type of  $Bob_i$ 's preferences on acts satisfy axiom A-4-1 to A-4-7, each type of  $Bob_i$ 's subjective beliefs concerning the likelihoods of various events are refined enough for there to be a quantified notion of probability representing them. Moreover, the axioms A-4-1 to A-4-7 guarantee that the probability associated with each event, which represents each type of  $Bob_i$ 's subjective belief of the likelihood of that event, is *unique*.

One should not confuse this as saying that the same probabilities will be uniformly associated with any given event for all three types of  $Bob_i$ .  $Bob_+$  might act *as if* he were assigning probability .73 to event  $e$ ; while  $Bob_S$  might act *as if* he were assigning probability .51 to event  $e$ ; while  $Bob_G$  might act *as if* he were assigning probability .22 to event  $e$ . That is, the three types of Bobs may be acting as if they were assigning completely different probabilities for any given event. However, what the lemma is claiming is that, as long as each type of  $Bob_i$ 's preference towards acts satisfy axioms A-4-1 to A-4-7, there can be only one probability associated with any given event for any given type of Bob.

Also, as I have already mentioned previously, lemma L-4-1 is not claiming that each type of Bob is consciously assigning probabilities for each event when he acts in a particular way. The probabilities are extrapolated from each type of  $Bob_i$ 's choice behaviors. What the lemma is claiming is that given that each type of  $Bob_i$ 's choice behaviors (which reveal his preference towards acts) satisfy axiom A-4-1 to A-4-7, it is possible for us to *interpret* each type of  $Bob_i$ 's behaviors *as if* he were assigning unique probabilities to events in set  $E$ .

The next major step in deriving the representation theorem is to *associate each act with a probability distribution on the set of consequences (i.e. C.)* Since we have already derived

each type of  $Bob_i$ 's subjective probabilities for events in the previous step, this process can be performed relatively easily. The strategy is to define the probability distribution on  $C$  associated with act  $f \in A$  (i.e.  $p_f^{Bob_i}$ ) as follows:

**[DEFINITION(D-4-5)] (Probability Distribution Associated with Acts):** For all  $x \in C$ ,  $f \in$

$$A, p_f^{Bob_i}(x) \equiv_{df} p^{Bob_i}(\{w : f(w) = x\})$$

In other words,  $Bob_i$ 's subjective probability of experiencing consequence  $x$  when he performs act  $f$  is identified with his subjective probability for *the event* where consequence  $x$  is realized when he performs act  $f$  to occur. Let's, once again, go back to the sports team betting example. Team X is going to compete against team Y, and suppose that there are no draws. Suppose that you prefer to bet \$100 on team X rather than to bet on team Y. Given that your preference towards actions (in this case, your preference towards various bets on sports teams) satisfy axioms A-4-1 to A-4-7, there is a unique probability associated with each event that represents your subjective beliefs concerning the likelihoods of those events.

Suppose that probability  $p$  represents your subjective beliefs concerning the likelihood of the event of team X winning. Then, according to definition D-4-5, *the act* "Bet \$100 on team X" can be identified with *the lottery* (i.e. probability distribution),  $[p \cdot (\text{Win } \$100) + (1 - p) \cdot (\text{Lose } \$100)]$ ; similarly, *the act* "Bet \$100 on team Y" can be identified with *the lottery* (i.e. probability distribution),  $[(1 - p) \cdot (\text{Win } \$100) + p \cdot (\text{Lose } \$100)]$ . So, when you preferred to bet \$100 on team X rather than on team Y, we can say that you were, in effect, preferring to play the lottery  $[p \cdot (\text{Win } \$100) + (1 - p) \cdot (\text{Lose } \$100)]$  to the lottery  $[(1 - p) \cdot (\text{Win } \$100) + p \cdot (\text{Lose } \$100)]$ .

What this means is that, with definition D-4-5, we can *translate* each type of  $Bob_i$ 's preferences towards *actions* in  $A$  into each type of  $Bob_i$ 's preferences towards various lotteries (i.e. probability distributions) defined on  $C$ ; that is, we can think of the set of all acts (i.e.  $A$ ) as virtually the same thing as the set of all probability distributions on  $C$ , which in the previous Von-Neumann and Morgenstern's framework was denoted by  $P$ . So, after deriv-

ing subjective probabilities from each type of  $Bob_i$ 's preferences towards *actions*, we have in effect arrived back to our previous Von-Neumann and Morgenstern's framework. What is left for us to show is that the seven axioms from A-4-1 to A-4-7 jointly imply the four Von-Neumann and Morgenstern's axioms (i.e. A-3-1 to A-3-4) that we have seen in section 7.3.2.

The next step is to show that, given that  $Bob_i$ 's preferences towards acts meet the seven axioms from A-4-1 to A-4-7, each type of  $Bob_i$  will be *indifferent* towards any two acts that are associated with the same probability distribution defined on  $C$ . I will simply state the lemma without proving it.

**[LEMMA(A-4-2)] (*Indifferent Acts*):** Suppose that each type of  $Bob_i$ 's preferences on  $A$  (i.e. the set of acts) satisfy the seven axioms from A-4-1 to A-4-7 and that  $p_f^{Bob_i}(x) = p_g^{Bob_i}(x)$  for all  $x \in C$  where  $f, g \in A$ . Then,  $f \sim_{Bob_i} g$ .

What this lemma is saying is that, in the end, what each type of  $Bob_i$  cares about are only the particular consequences and their final probabilities of occurring; not the particular states that the consequences are realized or the specific ways the consequences are randomized. Suppose that, besides betting \$100 on either team X or team Y, you are given a third option which is: your friend tosses a fair coin and if the coin lands heads you win \$100 with probability  $r$  and lose \$100 with probability  $1 - r$ , and if the coin lands tails you win \$100 with probability  $2p - r$  and lose \$100 with probability  $1 - (2p - r)$  (where  $1 + r > 2p > r > 0$ .)

If we calculate the probabilities of the third option, we can see that by choosing this third option, you will win \$100 with probability  $p$  and you will pay \$100 with probability  $(1 - p)$ . That is, the third option can basically be identified with the probability distribution  $[(1 - p) \cdot (\text{Win } \$100) + p \cdot (\text{Lose } \$100)]$  which is the same probability distribution associated with the act "Bet \$100 on team X." What the lemma is claiming is that, in such cases, you would be *indifferent* between the act "Bet \$100 on team X" and choosing the third option.

Here, the consequence of winning \$100 and losing \$100 are realized in different possible

states for the two acts in question; with the act “Bet \$100 on team X” the consequence of winning \$100 is realized when the event of team X winning occurs, and with the third option, the same consequence is realized when the fair coin that your friend tosses lands either heads or tails and the events (whatever they happen to be) with the relevant probabilities described above occur. However, the lemma claims that, as long as the consequences as well as their respective final probabilities of occurring are the same, the specific ways that these consequences are realized does not matter.

We can see that this lemma, although described differently, performs the same function as the axiom of “Reduction of Compound Lotteries” (i.e. A-3-2) we encountered in Von-Neumann and Morgenstern’s framework in section 7.3.2. The next step is to show that the other three Von-Neumann and Morgenstern’s axioms (i.e. A-3-1, A-3-3, and A-3-4) are implied by Savage’s seven axioms (i.e. from A-4-1 to A-4-7) as well.

Axiom A-4-1 is basically the same axiom as A-3-1; that each type of  $Bob_i$ ’s preference-relation is both asymmetric and negatively transitive. Furthermore, we have seen that axioms A-4-2 and A-4-3 resemble Von-Neumann and Morgenstern’s “Independence Axiom” (i.e. A-3-3) Lastly, we have seen that axiom A-4-6 closely resembles Von-Neumann and Morgenstern’s “Archimedean Axiom” (i.e. A-3-4.) So, it should not be a surprise that all of Von-Neumann and Morgenstern’s axioms are logically implied by Savage’s seven axioms. Again, I will just simply state the lemma without proving it.

**[LEMMA(A-4-3)] (*Savage’s Axioms imply VnM Axioms*):** Suppose that each type of  $Bob_i$ ’s preferences on  $A$  (i.e. the set of acts) satisfy the seven axioms from A-4-1 to A-4-7. Then, for each type of  $Bob_i$ , each act  $f \in A$  can be identified with the probability distribution  $p_f^{Bob_i}$  as in D-4-5, and each type of  $Bob_i$ ’s preferences on  $p_f^{Bob_i}$  for all  $f \in A$  satisfy Von-Neumann and Morgenstern’s four axioms from A-3-1 to A-3-4.

We are now ready to state (finally!) Savage’s representation along with its uniqueness theo-

rem.

**[THEOREM(T-4-1)] (Savage's Expected Utility Representation):** The binary relation  $\succ_{Bob_i}$  on  $A$  satisfies axioms A-4-1 to A-4-7 if and only if there exist a function  $u_{Bob_i} : C \rightarrow \mathbb{R}$  such that: For all  $f, g \in A$ ,  $f \succ_{Bob_i} g$  if and only if  $p_f^{Bob_i} \succ_{Bob_i} p_g^{Bob_i}$  if and only if  $\sum_{x \in C} p_f^{Bob_i}(x) u_{Bob_i}(x) > \sum_{x \in C} p_g^{Bob_i}(x) u_{Bob_i}(x)$ . Furthermore,  $u_{Bob_i}$  is unique up to positive affine transformation.

The proof can be performed by following the same general strategy that I have introduced in section 7.3.2. I have already mentioned this before, but I believe that the meaning of this will be more apparent now. The main gist of Savage's expected utility theory, which is summarized in theorem T-4-1, is this: Suppose that one did not even have the slightest conception of what probabilities or utilities are. One simply thinks that certain actions are more preferable than others. However, according to Savage, given that one's preference towards actions meet the seven axioms from A-4-1 to A-4-7, it is possible for outsiders to interpret one's behaviors *as if* one were assigning (subjective) probabilities on events, utilities on consequences, and were aiming to maximize expected utility.

The significance of this is that given that we assume that each type of Bob is minimally rational and acts consistently (which, in this context, means to act according to Savage's seven axioms), we can *reliably predict* how each type of Bob would react to a given situation, once we have extrapolated each type of Bob's subjective probabilities on events as well as his utilities on consequences from observing his previous choices on actions. This will become crucial when we try to model Hobbes's state of nature in the next chapter.

Once again, each type of  $Bob'_i$ 's utilities for the individual consequences in  $C$  can be calibrated in the same way as we did in section 7.3.2. Again, the respective utilities for the consequences in  $C$  for each type of Bob can be summarized as follows:

- $u_{Bob_+}(\text{Glorified Life}) = u_{Bob_S}(\text{Glorified Life}) = u_{Bob_G}(\text{Glorified Life}) = 1$

- $u_{Bob_+}(\text{Moderate Life}) = u_{Bob_S}(\text{Moderate Life}) = p^+ = p^S > \frac{2}{3} > p^G = u_{Bob_G}(\text{Moderate Life})$
- $u_{Bob_+}(\text{Mortified Life}) = u_{Bob_S}(\text{Mortified Life}) = q^+ = q^S > \frac{1}{3} > q^G = u_{Bob_G}(\text{Mortified Life})$
- $u_{Bob_+}(\text{Death}) = u_{Bob_S}(\text{Death}) = u_{Bob_G}(\text{Death}) = 0$

Remember that each type of  $Bob'_i$ 's utility functions, stated here, are *expected utility functions*, where the utility of any probability distribution (i.e. lottery) is equal to the expected utility of the sure consequences involved in that probability distribution (i.e. lottery.) The utilities, here, represent what each type of  $Bob_i$  as *a matter fact* prefers. But, we can derive what the two actualized-Bobs (i.e.  $Bob_S$  and  $Bob_G$ ) *should* prefer, *normatively speaking*, based on the things that we have established in chapter 6. This is stated in the following theorem which will be the main theorem for our current section.

**[THEOREM (T5)] (Subjective Expected Utility Representation of Hobbes's Theory of Real Good):**

For all acts  $f, g \in A$ , performing act  $f$  is *really better* (or *substantially better*) than performing act  $g$  for  $Bob_i$  ( $i = S, G$ ) if and only if  $f \succ_{bob_+} g$  if and only if  $P_f^{Bob_+} \succ_{Bob_+} P_g^{Bob_+}$  if and only if  $\sum_{x \in C} P_f^{Bob_+}(x) u_{Bob_+}(x) > \sum_{x \in C} P_g^{Bob_+}(x) u_{Bob_+}(x)$ . That is, performing act  $f$  is really better (or substantially better) than performing act  $g$  for  $Bob_i$  ( $i = S, G$ ) if and only if the expected utility of act  $f$  for  $Bob_+$  is greater than the expected utility of act  $g$  for  $Bob_+$ .

**Proof.** Again, the proof can be performed in a very similar way as in the proof of T4, which we saw in section 7.3.2. That is, theorem T5 follows from T2 (which states that what is really good for a given individual is the satisfaction of the preferences formed by the individual's idealized-self on behalf of the individual's actual-self), T-4-1 (which states that as long as each type of  $Bob_i$  satisfies axioms A-4-1 to A-4-7, there exists an expected utility function  $u_{Bob_i}$  such that  $Bob_i$  strictly prefers to perform act  $x$  to act  $y$  if and only if the expected utility of act  $x$  is greater than the expected utility of act  $y$ ), our seven Savage's axioms from A-4-1

to A-4-7, and by our assumption that  $Bob_+$  is the idealized-self of  $Bob_i$  ( $i = S, G$ ). (Again, I will omit the details.)  $\square$

The facts about the substantial rationality of each type of Bob's preferences could be summarized in the usual way as follows:

**[PROPOSITION (P5)]:** It is substantially rational for  $Bob_i$  ( $i = S, G$ ) to strictly prefer to perform act  $f$  rather than to perform act  $g$  if and only if  $p_f^{Bob_+} \succ_{Bob_+} p_g^{Bob_+}$  if and only if  $\sum_{x \in C} p_f^{Bob_+}(x) u_{Bob_+}(x) > \sum_{x \in C} p_g^{Bob_+}(x) u_{Bob_+}(x)$ .

In the previous Von-Neumann and Morgenstern's framework, each type of Bob was provided with objective probability distributions to express his preference for; the probabilities in each lottery for which each type of Bob expressed his preference were both objective and predetermined. This means that, in the previous Von-Neumann and Morgenstern's framework, none of the three types of Bob's preferences were based on inaccurate estimations of probability.

If we, for the current moment, assume that having accurate estimations of probability suffice for somebody's preference to be based on sufficient information about the world, we can say that, in the previous Von-Neumann and Morgenstern's framework, all three types of Bobs' preferences were at least *well-considered*. So, if any type of Bob's preferences were substantially irrational in the Von-Neumann and Morgenstern's framework, it would have been mainly because his preferences were *unbalanced*; that is, the irrationality of the preference would have to have been due to the given type of Bob's being influenced by the wrong kind of basic passion when he formed such preference.

We have seen that this was the case for  $Bob_G$ . And, as we have seen, the substantial irrationality of  $Bob_G$ 's preference, which was due to its being unbalanced, was reflected in  $Bob_G$ 's utilities for the individual consequences (more specifically, his utilities for Moderate Life and Mortified Life) in  $C$ ; his utilities for Moderate Life (i.e.  $p^G < \frac{2}{3}$ ) and Mortified Life (i.e.  $q^G < \frac{1}{3}$ ) were much lower than that of  $Bob_+$ 's or  $Bob_S$ 's (i.e.  $p^+ = p^S > \frac{2}{3}$  and

$q^+ = q^S > \frac{1}{3}$ ), which, in practical terms, suggests that  $Bob_G$  would gladly prefer to take risky gambles that  $Bob_+$  and  $Bob_S$  would surely avoid. And, this is exactly what one would expect a person who is infatuated by a passion for glory to be like.

The problem with the previous Von-Neumann and Morgenstern's framework is that there is no way to explain how the preferences of  $Bob_S$ , who is another version of Bob's actual-self, could ever be irrational in the substantial Hobbesian sense. For if the preferences of  $Bob_S$  were substantially irrational, it could not have been due to  $Bob_S$ 's being influenced by the wrong kind of basic passion -  $Bob_S$  is already assumed to be influenced by the right kind of basic passion, which is the basic passion for self-preservation. So, if the preferences of  $Bob_S$  were substantially irrational, it is because  $Bob_S$  was insufficiently informed, especially, about the likelihoods of various consequences that would be generated by performing a given course of action. In short, if the preferences of  $Bob_S$  were substantially irrational, they are substantially irrational because they were *unconsidered*, and *not* because they were *unbalanced*.

Unlike the previous Von-Neumann and Morgenstern's framework, our current Savage's framework leaves room for  $Bob_S$ 's preferences to be substantially irrational in this way. For instance, suppose that, in the state of nature, Bob made an agreement with Fred to come and help each other when the other party is being attacked by another third party. Suppose that Fred gets attacked by somebody and calls Bob for help. What should Bob do?

The two available actions that come to Bob's mind are: (1) *Help Fred* (act *h*), and (2) *Ignore Fred's Request* (act *i*). Based on Hobbes's assumption of rough equality among men<sup>43</sup>, it is very likely that Bob and Fred would be able to successfully defeat the attacker by their joint endeavor if they managed to cooperate.

Suppose that, once offered help, Fred will be very likely to provide help in return when Bob is in trouble in the future, but will not provide any help for Bob in the future if his

---

<sup>43</sup>"Nature hath made men so equal in the faculties of body and mind as that ... For as to the strength of body, the weakest has strength enough to kill the strongest..." [Hobbes 1994: *Leviathan*, Chapter XIII, Paragraph 1]

request is ignored by Bob this time. Suppose that in the state of nature, one will, sooner or later, encounter an attacker that one will not be able to defeat alone.

Considering all of these facts, suppose that, Bob's idealized-self,  $Bob_+$  associates the probability distribution [.99·Moderate Life + .01·Death] (i.e.  $p_h^{Bob_+}(\text{Moderate Life})=.99$ ,  $p_h^{Bob_+}(\text{Death})=.01$ ) with act  $h$ , and associates the probability distribution [.5·Moderate Life + .5·Death] (i.e.  $p_i^{Bob_+}(\text{Moderate Life})=.5$ ,  $p_i^{Bob_+}(\text{Death})=.5$ ) with act  $i$ . Obviously,  $h \succ_{Bob_+} i$ , since  $.99 \cdot u_{Bob_+}(\text{Moderate Life}) + .01 \cdot u_{Bob_+}(\text{Death}) = .99 \cdot p^+ > .5 \cdot p^+ = .5 \cdot u_{Bob_+}(\text{Moderate Life}) + .05 \cdot u_{Bob_+}(\text{Death})$ .

However, suppose the actual-Bob,  $Bob_S$ , decides to ignore Fred's request with the fear of getting involved in an unwanted fight along with some wishful thinking that he would somehow manage to avoid facing such attacker himself in the future; so,  $h \prec_{Bob_S} i$ . According to P5, this implies that  $Bob_S$ 's preference is substantially irrational. However, obviously, the substantial irrationality of  $Bob_S$ 's preference was not due to his being influenced by the wrong kind of basic passion; as a matter of fact, the primary reason why  $Bob_S$  preferred to perform act  $i$  rather than to perform act  $h$  is because he thought that act  $i$  was the better way to secure his self-preservation. So, the main reason why  $Bob_S$ 's preference was substantially irrational is because his preference was based on inaccurate estimations of the likelihood of relevant events. Specifically,  $Bob_S$  was overestimating the likely danger (i.e. facing Death) of helping Fred out, and was underestimating the likelihood of him facing a random attacker in the future.

Our current model captures the intuition behind this explanation very nicely. Remember that, for both  $Bob_+$  and  $Bob_S$ , the utilities of each of the sure consequences in  $C$  were exactly the same; this reflects the fact that both types of Bobs are primarily influenced by the basic passion for self-preservation, which, according to Hobbes, is the right kind of basic passion for people to be under the influence of. This means that in order for both  $h \succ_{Bob_+} i$  and  $h \prec_{Bob_S} i$  to be the case, it must be the case that  $p_h^{Bob_S}(\text{Moderate Life}) < p_i^{Bob_S}(\text{Moderate Life})$ , which implies that  $p_h^{Bob_S} \neq p_h^{Bob_+}$  and  $p_i^{Bob_S} \neq p_i^{Bob_+}$  since  $p_h^{Bob_+}(\text{Moderate Life}) = .99$

$> p_i^{Bob+}(\text{Moderate Life}) = .01$ .

What all of this is saying is that, unlike what his own idealized-self,  $Bob_+$  believed, the actual  $Bob_S$  believed that he actually had a better chance to achieve Moderate Life by ignoring Fred's help request. Such belief was false; and the resulting preference, was, thereby, unconsidered, and, therefore, substantially irrational.

In short, within to our current reconstruction, whereas *unbalancedness*, which is one form of substantial irrationality, is reflected in the discrepancies between the utilities of the sure-consequences of one's idealized-self and actual-self, *unconsideredness*, which is another form of substantial irrationality, is reflected in the discrepancies between the probability distribution that one's idealized-self and actual-self associates with each course of action. All of this discussion suggests that, with our framework, we can provide a formal definition of what it means for one's preferences to be *fully-considered* and *fully-balanced*:

**[DEFINITION(D-4-6)] (*Fully-Consideredness*):** For all  $f, g \in A$ ,  $Bob_i$  ( $i = +, S, G$ )'s preference between act  $f$  and act  $g$  is *fully-considered* if and only if  $p_f^{Bob+} = p_f^{Bob_i}$  and  $p_g^{Bob+} = p_g^{Bob_i}$ .

**[DEFINITION(D-4-7)] (*Fully-Balancedness*):** For all  $x \in C$ ,  $Bob_i$  ( $i = +, S, G$ )'s preferences are *fully-balanced* if and only if  $u_{Bob_+}(x) = u_{Bob_i}(x)$ .

As I have already explained in chapter 1 (see section 1.3 "Rational Preferences"), one's preference need not be fully-considered or fully-balanced in order for it to be substantially rational in the Hobbesian sense. According to P5, any actual-Bob's preference between any two acts is substantially rational if and only if it coincides with the idealized-Bob's preference between the two acts. This means that even if  $p_f^{Bob+} \neq p_f^{Bob_i}$  and  $p_g^{Bob+} \neq p_g^{Bob_i}$  (which, by definition, would imply that the actual-Bob's preferences are *short of being fully-considered*), and even if  $u_{Bob_+}(x) \neq u_{Bob_i}(x)$  for some consequence  $x$  (which, by definition, would imply that the actual-Bob's preferences are *short of fully-balanced*), the actual-Bob's preference between act  $f$  and act  $g$  may still be substantially rational as long as the subjective

probabilities as well as the utilities for individual consequences of the actual-Bob are *close enough* to that of the idealized-Bob's so that the actual-Bob would still manage to prefer to perform the act that the idealized-Bob advises him to perform. When the actual-Bob's subjective probabilities and utilities are close enough in this way, the resulting preferences, albeit being non-fully-considered and non-fully-balanced, would still be considered as well-considered and well-balanced, and, would, thereby, qualify as being substantially rational.

Note that this means that it is possible for the actual-Bob's preferences to be substantially rational for some pair-wise comparisons between acts while, at the same time, be substantially irrational for other pair-wise comparisons between acts. For example, suppose that  $f \succ_{Bob_+} g \succ_{Bob_+} h$  and  $f \succ_{Bob_S} h \succ_{Bob_S} g$ . Then,  $Bob_S$ 's preference between act  $f$  and act  $g$  as well as his preference between act  $f$  and act  $h$  would be substantially rational, while his preference between act  $g$  and act  $h$  would be substantially irrational.

## 5.4 An Additive Utility Representation of Hobbes's Theory of Real Good

We are almost there. One final touch that must be made to our reconstruction is to take account of the aspects of *time*. It should be noted that considerations of time is a very important factor within Hobbes's theory of real good. For instance, as we have seen in chapter 1 (section 1.4. "Irrational Preferences"), according to Hobbes, the major flaw in the practical reasoning of *the fool* is that he/she did not properly take into account *the long-term effects* of his/her behaviors when he/she decided to cheat; that is, he/she underestimated the probability of his/her cheating behaviors being detected and did not properly consider what would happen at a later time if others discovered that he/she cheated.

The importance of taking long-term effects that one's action would bring into consideration in determining the overall goodness of a given course of action is evident in the

passage where Hobbes expresses his inchoate idea of aggregating the goodness of various consequences that are spread throughout different time periods.

And because in deliberation the appetites and aversions are raised by foresight of the good and evil consequences and sequels of the action whereof we deliberate, the good or evil effect thereof dependeth on the foresight of a long chain of consequences, of which very seldom any man is able to see to the end. But for so far as a man seeth, *if the good in those consequences be greater than the evil, the whole chain is that which writers call apparent or seeming good. And contrarily, when the evil exceedeth the good, the whole is apparent or seeming evil* [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 57 emphasis added]

The idea that Hobbes is expressing here is something like this. Suppose that there are two acts ( $a_1, a_2$ ), two type of consequences (*good, evil*), and a total of five time periods ( $t_1, t_2, t_3, t_4, t_5$ ). Now, suppose that the chain of consequences associated with the two acts can be summarized as follows.

Table 5.4: The Goodness of Chains of Consequences

acts \ time periods	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$
$a_1$	<i>good</i>	<i>evil</i>	<i>good</i>	<i>good</i>	<i>evil</i>
$a_2$	<i>evil</i>	<i>good</i>	<i>good</i>	<i>evil</i>	<i>evil</i>

Remember that we have interpreted “apparent (or seeming) good” as the option (in this case, the chain of consequences that is associated with each act) that one just happens to prefer. As we already know, one’s preferences are represented by one’s utility function. This means that what Hobbes is saying above is that the utility of a given course of action can be

identified with *the sum* of the utilities of the individual consequences in the whole chain of consequences this is associated with that action.

So, what Hobbes is saying above is:  $a_1 \succ a_2$  (act  $a_1$  is apparently better than (i.e. strictly preferred to) act  $a_2$ ) if and only if  $U(a_1) > U(a_2)$  if and only if  $u_1(\text{good})+u_2(\text{evil})+u_3(\text{good})+u_4(\text{good})+u_5(\text{evil}) > u_1(\text{evil})+u_2(\text{good})+u_3(\text{good})+u_4(\text{evil})+u_5(\text{evil})$ . Suppose that  $u_i(\text{good})=1$  and  $u_i(\text{evil})=-1$  for  $i = 1, 2, 3, 4, 5$ . Then,  $U(a_1) = u_1(\text{good})+u_2(\text{evil})+u_3(\text{good})+u_4(\text{good})+u_5(\text{evil}) = 1 - 1 + 1 + 1 - 1 = 1 > U(a_2) = u_1(\text{evil})+u_2(\text{good})+u_3(\text{good})+u_4(\text{evil})+u_5(\text{evil}) = -1 + 1 + 1 - 1 - 1 = -1$ . Therefore,  $a_1 \succ a_2$ .

Adding the utilities of the individual consequences to determine the utility of the whole chain of consequences seems to be intuitively innocuous; at first glance, there seems to be no big deal about it. However, we should never take it for granted that people's preferences concerning consequences that have multiple components can have an additive utility representation that we have just encountered. This is because there are specific conditions that the individual components of a given consequence must satisfy in order for there to be an additively separable utility representation for those consequences. The most important condition is: *independence (among the individual components.)*

We have already encountered the independence condition in both the VnM and Savage's framework (i.e. Axioms A-3-3, A-4-2, A-4-3) in the previous section. The basic idea was that one's preferences concerning the consequences that are realized on a particular state (or event) should not be influenced by what kind of consequence is realized on another state (or event.) Or to put it in another way, there should be no interaction or interdependencies among the consequences realized in different states (or events) that would, in any way, influence how one feels about a given consequence realized in a particular state (or event.)

We can actually see that an expected utility representation is a certain form of an additively separable utility representation. Consider  $U_{Bob_i}(p_1x_1, \dots, p_nx_n) = p_1u_{Bob_i}(x_1) + \dots + p_nu_{Bob_i}(x_n)$ . Here, the utility of a given probability distribution is expressed by *the sum* of the utilities of the individual consequences weighted by their respective probabilities. And

the major reason why such additively separable utility representation was possible in the first place was because we had assumed that each type of  $Bob_i$  satisfied the independence axiom.

It would be easier to understand the necessity of independence for there to be an additively separable utility representation by looking at a specific example where independence among the individual components that comprise a given option fail.

Suppose that Jennifer is at a restaurant that serves a “complete meal” that consists of an appetizer and two main dishes. What will be served is already fixed but Jennifer is allowed to choose the type of meat that she would like to include in each of the dishes. Her choices of meat are from the set  $M = \{\text{Shrimp, Beef, Pork, Chicken, Vegan}\}$ . Suppose that Jennifer is not a vegetarian, and her preferences towards the various meat (and non-meat) are:  $\text{Shrimp} \succ \text{Beef} \succ \text{Pork} \succ \text{Chicken} \succ \text{Vegan}$ . Suppose that a vector  $(x_1, x_2, x_3) \in M \times M \times M$  denotes the choice of three types of meat that is included in the three respective dishes.

We have assumed that Jennifer most prefers Shrimp among the various meat (and non-meat) included in  $M$ . However, suppose that Jennifer, just like many people, also likes *variety*. Therefore, even though Jennifer likes Shrimp the most she might actually prefer the meal (Chicken, Shrimp, Beef) to (Shrimp, Shrimp, Shrimp).

Suppose that this is so. Then, it is very easy to see that there cannot be an additively separable utility representation of Jennifer’s preferences towards the various meal combination. To see this, let  $u(\text{Shrimp}) = s$ ,  $u(\text{Beef}) = b$ ,  $u(\text{Pork}) = p$ ,  $u(\text{Chicken}) = c$ , and  $u(\text{Vegan}) = v$  where  $s > b > p > c > v$  for  $s, b, p, c, v \in \mathbb{R}$ . Then, we can see that function  $u$  accurately represents Jennifer’s preferences towards various types of meat in the sense that  $x \succ y$  if and only if  $u(x) > u(y)$ .

Now, suppose that Jennifer’s preferences towards various *composite meals* had an additively separable representation. Then,  $u(\text{Shrimp, Shrimp, Shrimp}) = u(\text{Shrimp}) + u(\text{Shrimp}) + u(\text{Shrimp}) = s + s + s = 3s$ , which, for every value of  $s$  that satisfies the inequality  $s > b > p > c > v$ , is greater than  $u(\text{Chicken, Shrimp, Beef}) = u(\text{Chicken}) + u(\text{Shrimp}) + u(\text{Beef}) = c + s + b$ . This suggests that Jennifer prefers the meal (Shrimp, Shrimp, Shrimp) to the meal

(Chicken, Shrimp, Beef). However, we have already assumed that she prefers (Chicken, Shrimp, Beef) to (Shrimp, Shrimp, Shrimp) because she likes variety. So, here, we can see that there cannot be an additively separable utility function that correctly represents Jennifer's preferences over composite meals.

I hope that nobody, at this point, is tempted to say that although Jennifer *does* prefer (Chicken, Shrimp, Beef) to (Shrimp, Shrimp, Shrimp), she *should* prefer (Shrimp, Shrimp, Shrimp) to (Chicken, Shrimp, Beef) since the additive utility of (Shrimp, Shrimp, Shrimp) is greater than that of (Chicken, Shrimp, Beef). Again, such remark is based on reading aspects of utilitarianism into the notion of utilities and utility functions, which I have already warned the reader to be very cautious about.

A utility function is simply a thermometer of a person's preference. So, it is the person's preferences that a utility function must adapt to, not the other way around. Saying that Jennifer should prefer (Shrimp, Shrimp, Shrimp) to (Chicken, Shrimp, Beef) since the additive utility of (Shrimp, Shrimp, Shrimp) is greater than that of (Chicken, Shrimp, Beef) implies that there is a normative reason for her to disregard variety in her meal. I doubt that there is any such normative reason.

So, there is no additively separable utility representation of Jennifer's preferences concerning composite meals. And this is mainly because her preference for what kind of meat she would like to have on one of the dishes is *dependent upon* what kind of meat she is going to get on the other two dishes of the whole meal. In short, Jennifer's preferences towards the consequences that are realized in one component of the vector  $(x_1, x_2, x_3) \in M \times M \times M$  is *not independent of* the consequences that are realized in other components of the vector.

So, if we are aiming to find an additive utility representation of Hobbes's theory of real good for multiple periods of time as implied in the passage that we have seen in the beginning of this section, we would first have to verify that each type of *Bob<sub>i</sub>*'s preferences towards the consequences that are realized in each time period is independent of the consequences that are realized in other time periods.

There are many ways to define and test independence and show that it leads to an additively separable utility representation. In this section, I will introduce the method developed in [Fishburn, 1965] and apply it to our current framework.

We start by describing the main components of our framework. Now, the set of consequences is the product set,  $C_1 \times \dots \times C_n$  where  $C_i = \{Death_i, MortifiedLife_i, ModerateLife_i, GlorifiedLife_i\}$ . The subscript  $i$  designates the specific time period. An element  $\mathbf{c}$  in the set of consequences is a vector  $(x_1, \dots, x_n)$  that denotes a given chain-of-consequences where the consequence  $x_i$  is realized in the  $i$ -th period of time. For example, the vector  $(ModerateLife_1, GlorifiedLife_2, MortifiedLife_3, \dots, Death_n)$  denotes a chain of consequences where one experiences Moderate Life at time period 1, Glorified Life at time period 2, Mortified Life at time period 3, ..., and Death at time period  $n$ .

So, now, what each type of  $Bob_i$  is considering is, not just simply individual consequences, but rather, *chains of consequences* in the set  $C_1 \times \dots \times C_n$ . Since each set  $C_i$  consists in four elements, the product set  $C_1 \times \dots \times C_n$  would normally consist in a total of  $4^n$  elements.

Of course, among the chains of consequences in  $C_1 \times \dots \times C_n$ , there will be some chains of consequences that are practically impossible; such as  $(ModerateLife_1, Death_2, Death_3, GlorifiedLife_4, \dots)$ .  $(ModerateLife_1, Death_2, Death_3, GlorifiedLife_4, \dots)$  is the chain of consequence where  $Bob_i$  experiences Moderate Life in the first period, gets killed in the second period, stays dead in the third period, and suddenly revives and experiences Glorified Life in the fourth period, and so on.

Needless to say, any rationally-minded person would not assign any positive probability to such chain of consequences when he/she tries to figure out what kind of consequences would unfold by performing a given course of action. More generally, any chains of consequences where a non-Death consequence appears in any time period after the first occurrence of Death would be practically impossible. Nevertheless, I believe that it is still not entirely meaningless to think about how each type of  $Bob_i$  would feel about such impractical chains

of consequences *if* such chains of consequences *were* actually realized. Therefore, it still makes sense to consider the whole product set  $C_1 \times \dots \times C_n$  as our set of (chains of ) consequences. Every chain of consequences in  $C_1 \times \dots \times C_n$  would be assigned utilities; but, not all of the chains of consequences will be assigned positive probabilities when each available act gets associated with a given probability distribution.

Everything else is left as it were in the previous VnM and Savage framework; that is, each type of  $Bob_i$  has preferences over acts, from each type of  $Bob_i$ 's preferences over acts we derive a unique probability distribution on the set of events (i.e. the set of all subsets of the set of states), each act is associated with a probability distribution on the set of (chains of) consequences, and the utility of any act equals the expected utility of the probability distribution that is associated with the given act in question. But, now, each consequence in a given lottery (i.e. probability distribution) is, not a single consequence, but a chain of consequence of the form  $(x_1, \dots, x_n)$ , and we would like to make the utility of this chain of consequence,  $u_{Bob_i}(x_1, \dots, x_n)$ , expressed in the additively separable form,  $u_{Bob_i}(x_1) + \dots + u_{Bob_i}(x_n)$ , just as Hobbes himself describes in the passage that we have seen in the beginning of this section.

As I have explained, doing this requires each type of  $Bob_i$ 's preferences concerning the consequences that are realized in any single time period to be independent of the consequences realized in other time periods. In what follows, I will introduce Fishburn's definition of independence, argue that each type of  $Bob_i$ 's preferences concerning various chains of consequences meet such definition of independence, and show that meeting such definition of independence leads to an additively separable utility representation of chains of consequences consisting of multiple components.

Let  $P$  be the set of all lotteries (i.e. probability distributions) on  $C_1 \times \dots \times C_n$ . Consider a pair of lotteries  $(l_1, l_2)$   $l_1, l_2 \in P$  where:

$$l_1 = (p_1c^1, p_2c^2, \dots, p_jc^j), \sum p_a = 1, c^a \in C_1 \times \dots \times C_n, \text{ for all } a = 1, \dots, j$$

$$l_2 = (q_1d^1, q_2d^2, \dots, q_kd^k), \sum q_b = 1, d^b \in C_1 \times \dots \times C_n, \text{ for all } b = 1, \dots, k$$

We now define the set  $G$  as follows:

- $G = \{(l_1, l_2) \mid l_1 \neq l_2 \text{ and both } l_1 \text{ and } l_2 \text{ have the same total probability for any } x_i \in C_i \text{ (} i = 1, 2, \dots, n \text{) that appears in either}\}$

It would be much easier to understand what sort of things belong in set  $G$  by looking at a concrete example. Consider the following two lotteries, when the total number of time periods is 3 ( $n = 3$ ):

$$l_1 = [\frac{2}{3}(\text{Moderate Life}_1, \text{Moderate Life}_2, \text{Moderate Life}_3), \frac{1}{3}(\text{Glorified Life}_1, \text{Glorified Life}_2, \text{Glorified Life}_3)]$$

$$l_2 = [\frac{1}{3}(\text{Glorified Life}_1, \text{Moderate Life}_2, \text{Moderate Life}_3), \frac{1}{3}(\text{Moderate Life}_1, \text{Glorified Life}_2, \text{Moderate Life}_3), \frac{1}{3}(\text{Moderate Life}_1, \text{Moderate Life}_2, \text{Glorified Life}_3)]$$

The two lotteries  $l_1$  and  $l_2$  are different;  $l_1$  gives the chains of consequences (Moderate Life<sub>1</sub>, Moderate Life<sub>2</sub>, Moderate Life<sub>3</sub>) and (Glorified Life<sub>1</sub>, Glorified Life<sub>2</sub>, Glorified Life<sub>3</sub>) each with probability 2/3 and 1/3, while  $l_2$  gives the chains of consequences (Glorified Life<sub>1</sub>, Moderate Life<sub>2</sub>, Moderate Life<sub>3</sub>), (Moderate Life<sub>1</sub>, Glorified Life<sub>2</sub>, Moderate Life<sub>3</sub>), (Moderate Life<sub>1</sub>, Moderate Life<sub>2</sub>, Glorified<sub>3</sub>) each with probability 1/3.

However, we can see that both lotteries give *the same total probability* for any time-specific consequence that appears in the lottery. For instance, in both lotteries, the total probability of each of the time-specific consequences Glorified Life<sub>1</sub>, Glorified Life<sub>2</sub>, Glorified Life<sub>3</sub> is 1/3, while the total probability of each of the time-specific consequences Moderate Life<sub>1</sub>, Moderate Life<sub>2</sub>, Moderate Life<sub>3</sub> is 2/3.

This means that the pair of lotteries  $(l_1, l_2)$  meets the criteria for set  $G$ 's membership, and is, therefore, a member of  $G$ . Given this, we now define our working notion of independence

of preferences among time-periods.

**[DEFINITION (D-5-1)] (*Independence Among Time-Periods*):** Given  $C_1 \times \dots \times C_n$  each type of  $Bob_i (i = +, S, G)$ 's preferences on the consequences that are realized in any given time period  $k = 1, \dots, n$  (i.e.  $Bob_i$ 's preferences on elements of  $C_k$ ) are independent of the consequences that are realized in any other time period  $j$  ( $j \neq k$  and  $j = 1, \dots, n$ ) if and only if  $l_1 \sim_{Bob_i} l_2$  for all  $(l_1, l_2) \in G$ .

In other words, the definition claims that if  $Bob_i$  feels that any two lotteries that give the same total probabilities for any time-specific consequence are equally preferable, then his preference towards the consequences that are realized in one specific time period is independent of the consequences (whatever they happen to be) realized in other time periods.

The intuition behind the definition is this. If you really are indifferent between any two lotteries that are different, but, which, nonetheless, give the same total probability for each time-specific consequence that occurs in them, this means that you *do not care about the specific way that each time-specific consequence is combined with other consequences* that occur in different time periods in a given chain of consequences. If this is the case, the, this means that your preference towards a given time-specific consequence is *independent of consequences* that occur in other time periods.

The next step in our current reconstruction is to see whether assuming that each type of  $Bob_i$ 's preferences towards each time-specific consequence is independent is consistent with Hobbes's own text. As we have already seen, Hobbes clearly has proposed an additive theory of preference and the good. And, such additive theory of preference and the good requires one's preferences towards the individual components that constitute a composite object to be independent of the other individual components in the composite object. So, whether Hobbes himself was actually aware of it or not, his theory of preference and the good are already committed to the notion of independence that we have just dealt. In other words, Hobbes's theory of preference and the good imply independence.

However, it might be better for our purpose to give some positive textual evidence that actually suggests that Hobbes had claimed something very similar to the notion of independence that we are dealing with. For this purpose, let's go back to the passage that I have quoted in the beginning of this section.

And because in deliberation the appetites and aversions are raised by foresight of the good and evil consequences and sequels of the action whereof we deliberate, the good or evil effect thereof dependeth on the foresight of a long chain of consequences, of which very seldom any man is able to see to the end. But for so far as a man seeth, *if the good in those consequences be greater than the evil, the whole chain is that which writers call apparent or seeming good. And contrarily, when the evil exceedeth the good, the whole is apparent or seeming evil* [Hobbes 1994: *Leviathan*, Chapter VI, Paragraph 57 emphasis added]

According to Hobbes, one perceives that a given chain of consequences is apparently good (i.e. something that one, as a matter of fact, prefers) when one perceives that the amount of good individual consequences that are included in the chain of consequences is greater than the amount of bad individual consequences included in the chain of consequences. Here, Hobbes's own wording suggests that he thought that the goodness (or the badness) of a given time-specific individual consequence is determined separately from what sort of consequences are located in other time periods.

Furthermore, Hobbes claims, here, that the good or evil effect depends on *the foresight* of a long chain of consequences. Based on our current framework, I believe that the term "foresight", here, can be translated into meaning one's *subjective probabilities*. Combining this with what we have established in the previous paragraph suggests that, according to Hobbes, the overall preferability (i.e. apparent goodness) of a given chain of consequences is determined by the total sum of the apparent goodnesses included in the individual parts, which, in turn, is determined by one's subjective probabilities for those individual consequences.

This implies that if the total (subjective) probabilities for each and every individual consequences located in two different chains of consequences are exactly the same, the apparent goodness of the two different chains of consequences would also be exactly the same as well. If one remember that apparent goodness is interpreted as what one happens to prefer, we can easily see that this means that the agent will be indifferent between the two chains of consequences in question. And, this is exactly what definition D-5-1 is claiming. So, regardless of whether or not independence (as defined in D-5-1) is a plausible normative assumption, we can, at least, say that Hobbes was committed to it.

The only thing left for us to show is that such notion of independence leads to an additively separable utility representation for different chains of consequences in our current framework. First of all, definition D-5-1 implies the following lemma:

**[LEMMA (L-5-1)] (*Even Chance Notion of Independence*):** Given  $C_1 \times \dots \times C_n$  each type of  $Bob_i (i = +, S, G)$ 's preferences on the consequences that are realized in any given time period  $t = 1, \dots, n$  (i.e.  $Bob_i$ 's preferences on elements of  $C_t$ ) are independent of the consequences that are realized in any other time period  $j$  ( $j \neq t$  and  $j = 1, \dots, n$ ) if and only if  $(\frac{1}{2}c^1, \frac{1}{2}c^2) \sim_{Bob_i} (\frac{1}{2}c^3, \frac{1}{2}c^4)$  whenever  $c^1, c^2, c^3, c^4 \in C_1 \times \dots \times C_n$  and any  $x_t \in C_t$  ( $t = 1, \dots, n$ ) that appears once (twice) in  $(\frac{1}{2}c^1, \frac{1}{2}c^2)$  also appears once (twice) in  $(\frac{1}{2}c^3, \frac{1}{2}c^4)$ .

It is easy to see that definition D-5-1 implies lemma L-5-1. Definition D-5-1 claims that  $Bob_i$ 's preferences towards consequences realized in any specific time period is independent of the consequences realized in any other time periods if and only if  $Bob_i$  feels indifferent towards two different lotteries that give the same total probability towards any time-specific consequence that occur in the lotteries. If this is the case, then it is obvious that  $Bob_i$  would feel indifferent towards any two lotteries that both gave a total probability of either 1/2 or 1 to every single time-specific consequence that occur in them. And this is basically what lemma L-5-1 is claiming.

From lemma L-5-1, we are able to derive the additively separable utility representation that we were seeking to achieve.

**[THEOREM (T-5-1)] (Additively Separable Utility Representation):** Given  $C_1 \times \dots \times C_n$  each type of  $Bob_i (i = +, S, G)$ 's preferences on the consequences that are realized in any given time period  $t = 1, \dots, n$  (i.e.  $Bob_i$ 's preferences on elements of  $C_t$ ) are independent of the consequences that are realized in any other time period  $j$  ( $j \neq t$  and  $j = 1, \dots, n$ ) if and only if there exist utility functions  $u_{tBob_i}$  on  $C_t$   $t = 1, \dots, n$  such that

$$U_{Bob_i}(x_1, x_2, \dots, x_n) = u_{1Bob_i}(x_1) + u_{2Bob_i}(x_2) + \dots + u_{nBob_i}(x_n) \quad (5.1)$$

for all  $(x_1, x_2, \dots, x_n) \in C_1 \times \dots \times C_n$  with  $u_{tBob_i}$  unique up to the simultaneous transformations  $u_{*tBob_i}(x_t) = u_{tBob_i}(x_t) + b_t$  for all  $x_t \in C_t$ ,  $t = 1, \dots, n$ ,  $\sum b_t = 0$  when  $U_{Bob_i}$  is fixed in origin and scale unit.

44

Previously, in our expected utility framework, we have derived each type of  $Bob_i$ 's (car-

---

<sup>44</sup>**Proof.** (The proof follows the general strategy of [Fishburn, 1965, pp. 42-43]) First, pick an arbitrary chain of consequences  $(x_1^0, x_2^0, \dots, x_n^0) \in C_1 \times \dots \times C_n$  and assign values to  $u_{tBob_i}^k(x_k^0)$  such that it satisfies the equation:

$$U_{Bob_i}(x_1^0, x_2^0, \dots, x_n^0) = u_{1Bob_i}(x_1^0) + u_{2Bob_i}(x_2^0) + \dots + u_{nBob_i}(x_n^0)$$

To make this more concrete, let  $x_t^0 = Death_t$  and  $u_{tBob_i}(Death_t) = 0$  for  $t = 1, \dots, n$ . Then,

$$\begin{aligned} &U_{Bob_i}(Death_1, Death_2, \dots, Death_n) \\ &= u_{1Bob_i}(Death_1) + u_{2Bob_i}(Death_2) + \dots + u_{nBob_i}(Death_n) \\ &= 0 + 0 + \dots + 0 = 0 \end{aligned} \quad (5.2)$$

Now, define each  $u_{tBob_i}$  as:

$$u_{tBob_i}(x_t) = U_{Bob_i}(Death_1, \dots, Death_{t-1}, x_t, Death_{t+1}, \dots, Death_n) - \sum_{j \neq t} u_{jBob_i}(Death_j) \quad (5.3)$$

for all  $x_k \in C_k$   $k = 1, \dots, n$ . We can see that the definition of  $u_{tBob_i}$  is simply a rearrangement of  $U_{Bob_i}(Death_1, \dots, Death_{t-1}, x_t, Death_{t+1}, \dots, Death_n) = u_{1Bob_i}(Death_1) + \dots + u_{t-1Bob_i}(Death_{t-1}) + u_{tBob_i}(x_t) + u_{t+1Bob_i}(Death_{t+1}) + \dots + u_{nBob_i}(Death_n)$ . However, we know that  $u_{tBob_i}(Death_t) = 0$  for  $t = 1, \dots, n$ . There-

dinal) utilities for the sure consequences in  $C$ . However, in the previous frameworks, we did not have the notion of time included in our model, and, thereby, the four consequences  


---

fore, the definition of  $u_{tBob_i}$  can be restated as:

$$u_{tBob_i}(x_t) = U_{Bob_i}(Death_1, \dots, Death_{t-1}, x_t, Death_{t+1}, \dots, Death_n) \quad (5.4)$$

Now, add both sides of (7.4) over all  $t = 1, \dots, n$ . Then, we get:

$$\sum_{t=1}^n u_{tBob_i}(x_t) = \sum_{t=1}^n U_{Bob_i}(Death_1, \dots, Death_{t-1}, x_t, Death_{t+1}, \dots, Death_n) \quad (5.5)$$

When I was beginning to explain our current framework, I have explained everything, (except for the fact that our set of consequences is, now, a multi-component product set  $C_1 \times \dots \times C_n$ ), is left intact as it were in the previous  $VnM$  and *Savage's* framework. This means that each type of  $Bob_i$ 's preferences towards various probability distributions on  $C_1 \times \dots \times C_n$  satisfy the four  $VnM$  axioms (from A-3-1 to A-3-4), which guarantees for there to exist an expected utility function representing each type of  $Bob_i$ 's preferences where the utility of a probability distribution equals the expected utility of its sure consequence.

Applying this fact to lemma L-5-1, with  $c^1 = (x_1, \dots, x_t, Death_{t+1}, \dots, Death_n)$ ,  $c^2 = (Death_1, \dots, Death_t, x_{t+1}, Death_{t+2}, \dots, Death_n)$ ,  $c^3 = (x_1, x_2, \dots, x_t, x_{t+1}, Death_{t+2}, \dots, Death_n)$ ,  $c^4 = (Death_1, \dots, Death_n)$ , we get:

$$\begin{aligned} & \frac{1}{2}U_{Bob_i}(x_1, \dots, x_t, Death_{t+1}, \dots, Death_n) + \frac{1}{2}U_{Bob_i}(Death_1, \dots, Death_t, x_{t+1}, Death_{t+2}, \dots, Death_n) \\ &= \frac{1}{2}U_{Bob_i}(x_1, x_2, \dots, x_t, x_{t+1}, Death_{t+2}, \dots, Death_n) + \frac{1}{2}U_{Bob_i}(Death_1, \dots, Death_n) \end{aligned}$$

Multiplying 2 to each side of the equation we get:

$$\begin{aligned} & U_{Bob_i}(x_1, \dots, x_t, Death_{t+1}, \dots, Death_n) + U_{Bob_i}(Death_1, \dots, Death_t, x_{t+1}, Death_{t+2}, \dots, Death_n) \\ &= U_{Bob_i}(x_1, x_2, \dots, x_t, x_{t+1}, Death_{t+2}, \dots, Death_n) + U_{Bob_i}(Death_1, \dots, Death_n) \end{aligned} \quad (5.6)$$

By summing both sides of this equation from  $k = 1, \dots, n - 1$ , we get:

$$\begin{aligned} & \sum_{t=1}^{n-1} U_{Bob_i}(x_1, \dots, x_t, Death_{t+1}, \dots, Death_n) + \sum_{t=1}^{n-1} U_{Bob_i}(Death_1, \dots, Death_t, x_{t+1}, Death_{t+2}, \dots, Death_n) \\ &= \sum_{t=1}^{n-1} U_{Bob_i}(x_1, x_2, \dots, x_t, x_{t+1}, Death_{t+2}, \dots, Death_n) + \sum_{t=1}^{n-1} U_{Bob_i}(Death_1, \dots, Death_n) \end{aligned} \quad (5.7)$$

By cancelling the terms and rearranging, we get:

$$\begin{aligned} & U_{Bob_i}(x_1, x_2, \dots, x_n) \\ &= \sum_{t=1}^n U_{Bob_i}(Death_1, \dots, Death_{t-1}, x_t, Death_{t+1}, \dots, Death_n) - (n-1)U_{Bob_i}(Death_1, \dots, Death_n) \end{aligned}$$

Since,  $U_{Bob_i}(Death_1, \dots, Death_n) = 0$ , we get:

$$\begin{aligned} & U_{Bob_i}(x_1, x_2, \dots, x_n) \\ &= \sum_{t=1}^n U_{Bob_i}(Death_1, \dots, Death_{t-1}, x_t, Death_{t+1}, \dots, Death_n) \end{aligned} \quad (5.8)$$

in  $C$  were *timeless*. Now that the consequences in our model are not just simple timeless consequences, but *chains of consequences* the individual components of which are dispersed throughout time, we would need to find a way to assign utilities for the individual consequences that occur in each time period. After doing this, we would be able to derive the utility for *the entire chain* of consequences *by summing up* the utilities that are assigned to the individual consequences realized in each time period. The legitimacy of such process is validated by theorem T-5-1.

In section 7.3, I have normalized each type of  $Bob_i$ 's utility scale by assigning 0 to the worst outcome (i.e. Death) and assigning 1 to the best outcome (i.e. Glorified Life.) From this, we were able to approximate the utilities for the other two outcomes (i.e. Moderate Life and Mortified Life) by finding the lottery that involves the two prizes Glorified Life and Death to which each type of  $Bob_i$  would feel indifferent to getting either Moderate Life or Mortified Life for sure. Again, each type of  $Bob_i$ 's utilities for these four timeless consequences can be summarized as follows:

- $u_{Bob_+}(\text{Glorified Life}) = u_{Bob_S}(\text{Glorified Life}) = u_{Bob_G}(\text{Glorified Life}) = 1$
- $u_{Bob_+}(\text{Moderate Life}) = u_{Bob_S}(\text{Moderate Life}) = p^+ = p^S > \frac{2}{3} > p^G = u_{Bob_G}(\text{Moderate Life})$
- $u_{Bob_+}(\text{Mortified Life}) = u_{Bob_S}(\text{Mortified Life}) = q^+ = q^S > \frac{1}{3} > q^G = u_{Bob_G}(\text{Mortified Life})$

---

Combining the results of (7.5) and (7.8), we finally get:

$$U_{Bob_i}(x_1, x_2, \dots, x_n) = u_{1Bob_i}(x_1) + u_{2Bob_i}(x_2) + \dots + u_{nBob_i}(x_n) \quad (5.9)$$

which is the additive utility representation that we were seeking to achieve. What the representation is saying is that, for each type of  $Bob_i$ , the utility of any chain of consequence in  $C_1 \times \dots \times C_n$  equals *the total sum* of the utilities of each individual consequence that is realized in each time period.

Adding  $b_t$  to each  $u_{tBob_i}(x_t)$  such that  $\sum_{t=1}^n b_t = 0$  will keep the equation (7.9) intact. Therefore, if  $u_{tBob_i}(x_t)$  represents  $Bob_i$ 's preferences towards the consequences that are realized in time period  $t$ , then  $u_{tBob_i}(x_t) + b_t$  ( $\sum_{t=1}^n b_t = 0$ ) also represents  $Bob_i$ 's preferences towards the consequences that are realized in time period  $k$ . This proves the uniqueness part of the theorem and, thereby, completes the proof.  $\square$

- $u_{Bob_+}(\text{Death}) = u_{Bob_S}(\text{Death}) = u_{Bob_G}(\text{Death}) = 0$

Now, each of these four consequences can, now, be realized in each time period from  $t = 1$  to  $t = n$ . Given *independence* of each type of  $Bob_i$ 's preferences towards the four consequences realized in different time periods, we are able to think of each time period *separately* in determining the relative distances among the four time-specific consequences in each type of  $Bob_i$ 's preference-ordering.

We *normalize* each time-specific utility function by assigning  $u_{tBob_i}(\text{Death}_t) = 0$  and  $u_{tBob_i}(\text{Glorified Life}_t) = 1$  for all  $t = 1, \dots, n$ . (Recall that during the proof of theorem T-5-1, I have assigned  $u_{Bob_i}^k(\text{Death}_k) = 0$  for all  $k = 1, \dots, n$ .) We, then, find the probabilities  $p_t^{+,S,G}$  to which  $Bob_{+,S,G}$  would feel indifferent between the sure consequence of Moderate  $\text{Life}_t$  and the lottery  $[p_t^{+,S,G} \text{Glorified Life}_t, (1 - p_t^{+,S,G}) \text{Death}_t]$  and assign  $u_{tBob_{+,S,G}}(\text{Moderate Life}_t) = p_t^{+,S,G}$ . Similarly, we find the probabilities  $q_t^{+,S,G}$  to which  $Bob_{+,S,G}$  would feel indifferent between the sure consequence of Mortified  $\text{Life}_t$  and the lottery  $[p_t^{+,S,G} \text{Glorified Life}_t, (1 - p_t^{+,S,G}) \text{Death}_t]$  and assign  $u_{tBob_{+,S,G}}(\text{Mortified Life}_t) = p_t^{+,S,G}$ .

Now, it might be the case that a given type of  $Bob_i$  does not value the same type of consequence realized in different time periods in the same way. For instance, it is very likely that  $Bob_G$  value a  $\text{Glorified Life}_t$  realized in an earlier time period much more highly than a  $\text{Glorified Life}_t$  realized in a much later time period. This means that  $u_{nBob_G}(\text{Glorified Life}_n) > u_{tBob_G}(\text{Glorified Life}_t)$  when  $n > t$ . We would need to incorporate this fact into our additively separable representation of each type of  $Bob_i$ 's preferences.

A convenient way to do this is to express the additively separable utility representation in *weighted form* as follows:

$$U_{Bob_i}(x_1, x_2, \dots, x_n) = \alpha_1^i u_{1Bob_i}(x_1) + \alpha_2^i u_{2Bob_i}(x_2) + \dots + \alpha_n^i u_{nBob_i}(x_n) \quad (5.10)$$

Here,  $\alpha_t^i$  ( $t = 1, \dots, n$ ) denotes the relative weights for each time-specific utility function

of  $Bob_i$ . To determine the relative weights for any two time periods, say  $t = 1, 2$ , we simply let  $Bob_i$  compare one option, which gives the best consequence in  $t = 1$  and the worst consequence in  $t = 2$  (i.e. (Glorified Life<sub>1</sub>, Death<sub>2</sub>)) with another option which gives the worst consequence in  $t = 1$  and the best consequence in  $t = 2$  (i.e. (Death<sub>1</sub>, Glorified Life<sub>2</sub>)). (Remember that (Death<sub>1</sub>, Glorified Life<sub>2</sub>) denotes the purely hypothetical farfetched chain of consequences where  $Bob_i$  dies at time period 1 and suddenly revives and experiences Glorified Life at time period 2.)

If  $Bob_i$ 's preference is: (Glorified Life<sub>1</sub>, Death<sub>2</sub>)  $\succ_{Bob_i}$  (Death<sub>1</sub>, Glorified Life<sub>2</sub>), then this implies that  $\alpha_1^i u_{1Bob_i}(Glorified\ Life_1) + \alpha_2^i u_{2Bob_i}(Death_2)$   
 $> \alpha_1^i u_{1Bob_i}(Death_1) + \alpha_2^i u_{2Bob_i}(Glorified\ Life_2)$ . Rearranging the terms, we get  
 $\alpha_1^i \{u_{1Bob_i}(Glorified\ Life_1) - u_{1Bob_i}(Death_1)\}$   
 $> \alpha_2^i \{u_{2Bob_i}(Glorified\ Life_2) - u_{2Bob_i}(Death_2)\}$ , which implies that  $\alpha_1^i > \alpha_2^i$   
since  $u_{tBob_i}(Glorified\ Life_t) = 1$  and  $u_{tBob_i}(Death_t) = 0$  for all  $t = 1, \dots, n$ .

This means that  $Bob_i$  values experiencing Glorified Life at time period 1 more than he values experiencing Glorified Life at time period 2. If the inequality was the other way around, then this would imply that  $Bob_i$  values experiencing Glorified Life at time period 2 more than he values experiencing Glorified Life at time period 1. If it is an equality, then this means that  $Bob_i$  values experiencing Glorified Life at time period 1 and experiencing Glorified Life at time period 2 equally.

By performing every pair-wise comparisons for all time periods  $t = 1, \dots, n$  in this way, we will be able to determine the relative weights for all  $\alpha_t^i$  ( $t = 1, \dots, n$ ). The relative weight  $\alpha_t^i$  ( $t = 1, \dots, n$ ) signifies the relative importance that  $Bob_i$  puts to a given time period  $t$ . So, when  $\alpha_1^i > \alpha_2^i > \dots > \alpha_n^i$ , this means that the given type of  $Bob_i$  cares more about short-term consequences than long-term consequences, and how much he cares about the consequences realized in a given time period diminishes by each increment time.

According to Hobbes, putting too much emphasis on short-term gains while neglecting long-term effects of one's actions is one of the major causes that makes an agent form ir-

rational preferences. According to Hobbes, this generally happens when the agent is being influenced by the wrong kind of basic passion - such as the basic passion for glory - that later becomes what Hobbes calls “perturbations” in the agent’s deliberation process.

Emotions or *perturbations* of the mind are species of appetite and aversion (. . .)  
 They are called perturbations *because they frequently obstruct right reasoning.*  
 (. . .) Therefore, although *the real good* must be sought *in the long term*, which is the job of reason, appetite seizeth upon a *present good* without foreseeing the greater evils that necessarily attach to it. Therefore appetite perturbs and impedes the operation of reason; whence it is rightly called a *perturbation.*  
 [Hobbes 1991: *De Homine*, Chapter XII, Section 1 emphasis added]

This would almost certainly be the case for  $Bob_G$  when he deliberates. That is, one of the major reasons why  $Bob_G$  would gladly take very risky gambles between Glorified Life and Death, which both  $Bob_+$  and  $Bob_S$  would quite obviously want to avoid, is because, the basic passion for glory, which  $Bob_G$  is infatuated with, would make  $Bob_G$  concentrate only on the immediate gains while making him overlook the long-term harms that such action would very likely bring. The result would be preferences that are irrational because they are *unbalanced*.

By utilizing the notion of relative weights in the additive representation of (7.10), we are able to characterize this type of *temporal unbalancedness*. That is, when  $Bob_G$ , by being influenced by a basic passion for glory, *overvalues* the importance of short-term gains while neglecting the importance of long-term losses, this would mean that  $\alpha_t^G$  is much greater than  $\alpha_t^{S,G}$  in earlier time periods while  $\alpha_t^G$  is much smaller than  $\alpha_t^{S,G}$  in later time periods. This means that, for  $i = S, G$ ,  $Bob_i$ ’s preferences concerning the consequences realized in a specific time period  $t$  is *temporally non-fully-balanced* if and only if  $\alpha_t^{S,G} \neq \alpha_t^+$ .

However, in the previous section, I have defined the notion of “fully-balancedness” in terms of utilities, and, it would, therefore, be nicer to find a way to characterize this type

of temporal unbalancedness in terms of utilities (and not relative weights) for the sake of a more unified approach.

Doing this is actually not very hard. In the weighted additive utility representation of (7.10), we can just simply incorporate the relative weights into each time-specific utility function itself. This would render each  $Bob_i$ 's time-specific utilities for each time-specific consequence as:

- $u_{tBob_{+,S,G}}(\text{Glorified Life}_t) = \alpha_t^{+,S,G}$
- $u_{tBob_{+,S,G}}(\text{Moderate Life}_t) = \alpha_t^{+,S,G} p_t^{+,S,G}$
- $u_{tBob_{+,S,G}}(\text{Mortified Life}_t) = \alpha_t^{+,S,G} q_t^{+,S,G}$
- $u_{tBob_{+,S,G}}(\text{Death}_t) = 0$

From this, we are able to express the utility for a given chain of consequences  $(x_1, x_2, \dots, x_n)$  in the additively separable form of (7.8) without using separate terms to denote relative weights as in (7.10).

From this, we extend the definition of fully-balancedness (i.e. D-4-7) to incorporate the notion of temporal non-fully-balancedness of  $Bob_i$ 's preferences that we have just discussed as follows.

**[DEFINITION(D-5-2)] (*Extension of Fully-Balancedness*):** For all chains of consequences

$(x_1, x_2, \dots, x_n) \in C_1 \times \dots \times C_n$ ,  $Bob_i$  ( $i = +, S, G$ )'s preferences are *fully-balanced* if and only if  $U_{Bob_+}(x_1, x_2, \dots, x_n) = U_{Bob_i}(x_1, x_2, \dots, x_n)$  if and only if  $u_{1Bob_+}(x_1) + u_{2Bob_+}(x_2) + \dots + u_{nBob_+}(x_n) = u_{1Bob_i}(x_1) + u_{2Bob_i}(x_2) + \dots + u_{nBob_i}(x_n)$ .

Moreover, for all time periods  $t = 1, \dots, n$ ,  $Bob_i$  ( $i = +, S, G$ )'s preferences concerning consequences that are realized in any given time period are *temporally fully-balanced* if and only if  $u_{tBob_+}(x_t) = u_{tBob_i}(x_t)$ .

With all of this in mind, we can now state our main theorem concerning Hobbes's theory of real good, as well as our main proposition concerning the notion of substantial rationality in our additively separable framework as follows.

**[THEOREM (T6)] (*Additive Utility Representation of Hobbes's Theory of Real Good*):** For

all chains of consequences  $(x_1, x_2, \dots, x_n), (y_1, y_2, \dots, y_n) \in C_1 \times \dots \times C_n$ , the chain of consequences  $(x_1, x_2, \dots, x_n)$  is *really better (or substantially better)* than the chain of consequences  $(y_1, y_2, \dots, y_n)$  for  $Bob_i$  ( $i = S, G$ ) if and only if  $(x_1, x_2, \dots, x_n) \succ_{bob_+} (y_1, y_2, \dots, y_n)$  if and only if  $U_{Bob_+}(x_1, x_2, \dots, x_n) > U_{Bob_+}(y_1, y_2, \dots, y_n)$  if and only if  $u_{1Bob_+}(x_1) + u_{2Bob_+}(x_2) + \dots + u_{nBob_+}(x_n) > u_{1Bob_+}(y_1) + u_{2Bob_+}(y_2) + \dots + u_{nBob_+}(y_n)$ .

That is, the chain of consequences  $(x_1, x_2, \dots, x_n)$  is really better (or substantially better) than the chain of consequences  $(y_1, y_2, \dots, y_n)$  for  $Bob_i$  ( $i = S, G$ ) if and only if, for  $Bob_+$ , the *total sum* of the utilities for the individual consequences realized in each time period of the chain of consequences  $(x_1, x_2, \dots, x_n)$  is greater than that of the chain of consequences  $(y_1, y_2, \dots, y_n)$ .

**[PROPOSITION (P6)]:** It is substantially rational for  $Bob_i$  ( $i = S, G$ ) to strictly prefer the chain of consequences  $(x_1, x_2, \dots, x_n)$  to the chain of consequences  $(y_1, y_2, \dots, y_n)$  if and only if  $(x_1, x_2, \dots, x_n) \succ_{bob_+} (y_1, y_2, \dots, y_n)$  if and only if  $U_{Bob_+}(x_1, x_2, \dots, x_n) > U_{Bob_+}(y_1, y_2, \dots, y_n)$  if and only if  $u_{1Bob_i}(x_1) + u_{2Bob_i}(x_2) + \dots + u_{nBob_i}(x_n) > u_{1Bob_i}(y_1) + u_{2Bob_i}(y_2) + \dots + u_{nBob_i}(y_n)$ .

I will omit the proof. The proof for both T6 and P6 can be performed in exactly the same way as it was performed in the previous sections.

## 5.5 The Role of Laws of Nature: Solving the Epistemic Problem

In our reconstruction of Hobbes's theory of real good, we have developed our theory on the basis of our interpretation that what is really good for a given individual (e.g. Bob) is to satisfy his/her *rational preferences* which can be identified with the type of preferences that the individual's *idealized-self* would form on behalf of the individual.

However, we have seen that, due to many practical constraints, the actual preferences of a given individual would, in many cases, be quite different from the type of preferences that the individual's idealized-self, who is fully-knowledgeable and who is, more than anything else, concerned about the individual's long-term self-preservation, would form on behalf of the actual individual. The individual might be insufficiently informed about the relevant facts about the world, which would make him/her associate the wrong probability distribution with a given course of action; the result will be preferences that are irrational because they are *unconsidered*. Or the individual might be primarily under the influence of the wrong kind of basic passion, which would make him/her assign wrong utilities for individual consequences as well as whole chains of consequences; the result will be preferences that are irrational because they are *unbalanced*.

So, the problem seems to be this. If what is really good for a given individual is the satisfaction of the type of preferences that his/her idealize-self would form on behalf of him/her, but, if, in very many cases, the preferences formed by actual individuals are, due to many practical constraints, not the type of preferences that Hobbes's theory of the good would urge them to form, then how can anyone manage to *know* what his/her idealized-self would advise him/her to do in any given situation, and, thereby, adjust his/her actual preferences accordingly? In short, given our reconstruction of Hobbes's theory of real good, we first encounter an *epistemic problem* of knowing what one's idealized-self would actually

advise one to do in any given situation.

I believe that what Hobbes calls the “laws of nature” are specifically designed to solve this very epistemic problem. According to Hobbes,

A LAW OF NATURE (*lex naturalis*) is a precept or general rule, found out by reason, by which a man is forbidden to do that which is destructive of his life or taketh away the means of preserving the same, and to omit that by which he thinketh it may be best preserved. [Hobbes 1994: *Leviathan*, Chapter XIV, Paragraph 3]

Again, the fact that the laws of nature, which are discovered by reason, are specifically designed to secure one’s self-preservation indicates that self-preservation is the major aim of one’s reason and rationality, as it is stated in (P1). The fact that the laws of nature provide prescriptive guidelines that help one secure one’s self-preservation indicates that the laws of nature are specifically designed to help one form *substantially rational preferences*.

In *Leviathan*, Hobbes introduces roughly twenty laws of nature. Here is a list of some of the laws of nature that Hobbes introduces:

- *The First Law*: “that every man ought to endeavour peace as far as he has hope of obtaining it, and when he cannot obtain it, that he may seek and use all helps and advantages of war.” [Hobbes 1994: *Leviathan*, Chapter XIV, Paragraph 4]
- *The Second Law*: “that a man be willing, when others are so too, as far-forth as for 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.” [Hobbes 1994: *Leviathan*, Chapter XIV, Paragraph 5]
- *The Third Law*: “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.” [Hobbes 1994: *Leviathan*, Chapter XV, Paragraph 1]

(...)

- *The Ninth Law*: “that every man acknowledge other for his equal by nature. The breach of this precept is *pride*.” [Hobbes 1994: *Leviathan*, Chapter XV, Paragraph 21]
- *The Tenth Law*: “that at the entrance into conditions of peace, no man require to reserve to himself any right which he is not content should be reserved to every one of the rest. ... the observers of this law are those we call *modest*, and the breakers *arrogant* men.” [Hobbes 1994: *Leviathan*, Chapter XV, Paragraph 22]

(and so on...)

The first law of nature urges one to choose peace when this is reliably expected and prepare for war when this is not. This means that in situations where the achievement of peace is reliably expected, acting peacefully best approximates what one’s idealized-self, who is, more than anything else, concerned about one’s long-term self-preservation, would advise one to do.

The second and the third laws of nature urge one to generally keep the agreements or covenants that one had made with others. This means that, in most cases, the act of keeping one’s agreements or covenants is associated with a probability distribution over chains of consequences that maximizes the expected utility of one’s idealized-self.

The ninth and tenth laws of nature are specifically designed to direct people’s attention away from any obsessions for glory and unnecessary power over other people. And, we can see how much important Hobbes thinks that it is to refrain from being obsessed with glory and power over others in order to secure one’s long-term self-preservation by observing his consistent emphasis on the importance of recognizing the natural equality of men in the other laws of nature.

My aim is not to go any deeper in analyzing Hobbes's laws of nature. My aim is simply to show how the laws of nature relate to our current reconstruction of Hobbes's theory of real goodness. As I have mentioned in the previous chapters, the laws of nature of nature constitute the contents of what Hobbes calls "moral science" or "moral philosophy."

And the science of them [the laws of nature] is the true and only moral philosophy. For moral philosophy is nothing else but the science of what is *good* and *evil* in the conversation and society of mankind. [Hobbes 1994: *Leviathan*, Chapter XV, Paragraph 40]

As we have seen in chapter 1, scientific knowledge is the culmination of the operation of one's combined rational faculties, and is something which Hobbes strongly advises one to incorporate into one's practical deliberation process. I believe that one of the main reasons why Hobbes did not just recommend people to incorporate their rational faculties into their deliberation process, but actually went a whole step further and presented a list of very concrete guidelines in the name of "laws of nature" for people to follow in their practical deliberations indicates that Hobbes was pretty much aware of the *epistemic gap* between what people *should* prefer in the substantially rational sense (i.e. the preferences that one's idealized-self would advise one to form) and what people actually do prefer in real life, and wanted to find a solution to this problem.

## Chapter 6

### A Bayesian Game-Theoretic

### Reconstruction of Hobbes's State of Nature

We now move on to one of the most important cornerstones of Hobbes's entire political philosophy; namely, his description of what would eventually happen in *the state of nature*; a state in which there is no centralized government power.

#### 6.1 Hobbes's State of Nature: State of War

Hobbes's justification for the existence of government entirely relies on the purported fact that, without a government, people's lives would be, not simply much worse, but utterly unbearable. This is because without a government that has sufficient power to enforce criminal laws and effectively regulate people's behaviors, the state of nature (which refers to a state where there is no government) will, according to Hobbes, inevitably dissolve into a state of universal war of all against all.

Hereby it is manifest that during the time men live without a common power to

keep them all in awe, they are in that condition which is called war, and such a war as is of every man against every man. [Hobbes 1994: *Leviathan*, Chapter XIII, Section 8]

Here, we must first get clear on what sort of state of affairs Hobbes intends to denote by the term “war.” For one thing, for Hobbes, the application of the term “war” is not restricted to the state of affairs where people are actually engaged in physical warfare with one another.

For WAR consisteth not in battle only, or the act of fighting, but in a tract of time wherein the will to contend by battle is sufficiently known. ... so the nature of war consisteth not in actual fighting, but in the known disposition thereto during all the time there is no assurance to the contrary. All other time is PEACE. [Hobbes 1994: *Leviathan*, Chapter XIII, Section 8]

In other words, for Hobbes, the state of war simply denotes a state of affairs where there is no sufficient guarantee of peace; that is, a state of affairs where expectations for a physical battle to break out at any given moment are quite high.

One of the major disadvantages of such precarious times is that it is impossible for social progress and prosperity to be stably achieved.

In such condition there is no place for industry, because the fruit thereof is uncertain, and consequently, no culture of the earth, no navigation, nor use of the commodities that may be imported by sea, no commodious buildings, no instruments of moving and removing such things as require much force, no knowledge of the face of the earth, no account of time, no arts, no letters, no society, and which is worst of all, continual fear and danger of violent death... [Hobbes 1994: *Leviathan*, Chapter XIII, Section 9]

Hobbes has famously summarized the life in the state of nature as:

... the life of man, solitary, poor, nasty, brutish, and short. [Hobbes 1994: *Leviathan*, Chapter XIII, Section 9]

## 6.2 The Five Axioms of The State of Nature (which leads to a state of universal war)

What are the specific conditions of the state of nature that makes Hobbes conclude that it will inevitably result in a state of universal war? Based on textual evidence, the characteristic conditions which, Hobbes thinks, would inevitably lead to a state of universal war can be summarized into the following five axioms (which I will refer to as “*the axioms of the state of nature.*”)

[AXIOM(A-6-1)] (*Equality*): People’s physical and mental capabilities are roughly equal.

The most important implication of this axiom is that, in the state of nature, even the weakest human being has enough power (either physical or mental) to kill the most powerful human being.<sup>1</sup> This means that everybody else in the state of nature can be a (potential) threat to one’s own self-preservation.

[AXIOM(A-6-2)] (*Competition Due To Scarce Resources*): In the state of nature, resources are scarce in such a way that there will inevitably arise situations in which two people would want to obtain the same object.

Coupled with the axiom of equality (i.e. A-6-1), this axiom implies that, in the state of nature, one would inevitably face situations in which one is in direct competition for a given resource with another person who has the potential to kill one. This is what Hobbes predicts in the following passage:

But the most frequent cause why men want to hurt each other arises when many want the same thing at the same time, without being able to enjoy it in common or to divide it. The consequence is that it must go to the stronger. But who is

---

<sup>1</sup>“For as to the strength of body, the weakest has strength enough to kill the strongest, ... , As as to the faculties of the mind ... I find yet a greater equality amongst men than that of strength.” [Hobbes 1994: *Leviathan*, Chapter XIII, Sections 1,2]

the stronger? Fighting must decide. [Hobbes 1997: *On the Citizen*, Chapter 1, Section 6]

And therefore, if any two men desire the same thing, which nevertheless they cannot both enjoy, they become enemies; and in the way to their end, which is principally their own conservation, and sometimes their delectation only, endeavour to destroy or subdue one another.” [Hobbes 1994: *Leviathan*, Chapter XIII, Section 3]

Now, one might question whether the axiom of competition (i.e. A-6-2) is really a reasonable assumption for the life in the state of nature. That is, is it really reasonable to assume that competition caused by scarcity of resources will inevitably arise in a situation where there is no government power?

The objection makes more sense if one remembers that the sole purpose of considering the state of nature in Hobbes’s political philosophy is to justify the existence of governments. If we absolutely need governments because, without governments, our lives in the state of nature would be unbearably miserable, and if one of the primary reasons why our lives in the state of nature would be unbearably miserable relies on the fact that competition will inevitably arise in the state of nature due to scarcity of resources, then it seems that our justification for the existence of governments would be significantly weakened in contemporary settings in which technological advances has, in many areas, overcome problems associated with scarcity of resources.

To this, it might be helpful to remember that, within Hobbes’s entire moral and political system, there is a resource that is guaranteed to be scarce: *power*.<sup>2</sup> One of the most signifi-

---

<sup>2</sup>Within Hobbes’s moral system, *power* is defined as a person’s “present means to obtain some future apparent good.” [Hobbes 1994: *Leviathan*, Chapter X, Paragraph 1] As we have seen previously, for Hobbes, a person’s *apparent good* is simply the satisfaction of the person’s current preferences. So, power, for Hobbes, generally means any present means that could be used to satisfy the preferences that a person has at a given moment. So, there could be many things that could count as power within Hobbes’s system. However, among the several things that could count as power, there is something that Hobbes regards as *the greatest*:

The greatest of human powers is that which is compounded of the powers of most men, united

cant characteristics of power<sup>3</sup> is that it is *zero-sum*; that is, one person's gain is necessarily coupled with another person's loss. This makes power a *scarce resource* that not everybody can fully enjoy. And, as we have seen, according to Hobbes, not only do people need power to secure their own self-preservation, but there is a certain portion of the human population that value power extremely highly, and, thereby, pursue power (not simply as a means for one's self-preservation, but) for its own sake. And, this leads us to the next two important axioms of the state of nature.

**[AXIOM(A-6-3)] (*The Existence Of Two Types of Men*):** In the state of nature, there exists two types of people: *the modest type* and *the vain-glorious type*. Furthermore, it is common knowledge that the state of nature is inhabited by these two types of people; that is, the modest type knows, and the vain-glorious type knows that the modest type knows that there is a certain portion of the entire human population who are *vain-glorious*, people who enjoy having superior power over others and pursue power, not as a means to secure one's self-preservation, but for its own sake.

**[AXIOM(A-6-4)] (*Human Psychology*):** *Not everybody's psychology is strictly egoistic.* The modest type, who compose the majority of the entire population, would strictly prefer to cooperate with other people given that these other people cooperate in return. This is not to say that the modest types are *perfect altruists* or *masochists*; they are not the type of people who would enjoy having their cooperative behaviors being taken advantage of by other people. By contrast, the vain-glorious types are the type of people who have a strictly egoistic psychology; they would

---

by consent in one person, natural or civil, that has the use of all their powers depending on his will, such as is the power of a commonwealth...[Hobbes 1994: *Leviathan*, Chapter X, Paragraph 3]

In other words, according to Hobbes, the greatest power that a person can possibly acquire is the power to use the powers of many other people according to one's will. This is power over other people. And, it is not hard to expect that the scarcity of such power cannot be alleviated by technological advancement.

<sup>3</sup>Here, I am denoting what Hobbes calls "the greatest of human powers" – namely, power over other people.

gladly enjoy taking advantage of other people's good intentions whenever it is to their advantage and increases their power.

The textual ground for these two axioms comes from the following passages:

Also, because there be some that taking pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their security requires, if others (that otherwise would be glad to be at ease within modest bounds) should not by invasion increase their power, they would not be able, long time, by standing only on their defence, to subsist. [Hobbes 1994: *Leviathan*, Chapter XIII, Section 4]

In the state of nature there is in all men a will to do harm, but *not for the same reason* or with equal culpability. One man practices the equality of nature (...) this is the mark of *modest man* (...) Another, supposing himself superior to other, wants to be allowed everything (...) that is the sign of *an aggressive character*. In his case, the will to do harm derives from *vainglory*. [Hobbes 1997: *On the Citizen*, Chapter 1, Section 4 emphasis added]

We can clearly see that Hobbes is *not* assuming that everybody in the state of nature is obsessed with power in a *vain-glorious way*; many people - those who Hobbes calls "modest" - would be quite content with having just enough power to subsist and secure their self-preservation. However, according to Hobbes, there are people who want more than enough power; not because they believe that having more power would help them better secure their self-preservation, but rather because they simply enjoy having conquered and having power over others.

It has been argued by many contemporary scholars that, not only does Hobbes assume psychological egoism in his description of the state of nature, but something very similar to psychological egoism is needed for Hobbes to properly explain the universal conflict, which,

he thinks, will inevitably emerge in the state of nature.<sup>4</sup> Psychological egoism is a doctrine that roughly claims that human beings *universally* have a strictly egoistic psychology - that is, according to psychological egoism, everybody is more or less what Gauthier calls “straight forward maximizers”<sup>5</sup> who always try to maximize his/her own self-interest. This is a contestable doctrine, and the more Hobbes’s political philosophy relies on psychological egoism, the less firm is its very foundation.

We can see from the above passages that at least Hobbes did not assume psychological egoism in his description of the state of nature. He concedes that there exist people who have a strictly egoistic psychology and “[takes] pleasure in contemplating their own power in the acts of conquest”; namely, people of the *vain-glorious type*. But he makes it clear that only “some” - and not all - human beings fit into this category.

If there are any substantial theories of human psychology to which Hobbes should be seen to be committed at this point, it is merely *the denial* of what may be called “psychological altruism” - a doctrine that claims that human beings are universally altruistic. This is a much plausible thing to assume than psychological egoism. As I have mentioned, it is quite contestable to think that people are universally selfish and care only about themselves; it is not hard to discover in our daily lives people displaying acts of benevolence and self-sacrifice. However, it seems quite evident from our everyday experiences that, at least, not everybody is a saint.

Some people might think that this is insufficient to generate the level and extent of universal conflict in Hobbes’s state of nature on which Hobbes’s justification for the existence of

---

<sup>4</sup>See [Butler, 1983, Hume, 1975, Broad, 1950]. Kavka [1986] thinks that some textual evidence does suggest that Hobbes was a psychological egoist, but thinks that only a weakened version of psychological egoism, which he calls “Predominant Egoism” is needed for Hobbes’s political philosophy to work. McNeilly [1966] thinks that Hobbes was at least committed to psychological egoism in his earlier works. Hampton [1986, pp. 20-24] interprets Hobbes as a psychological egoist who maintains that all of our desires are *caused by a “self-interested” bodily mechanism*, and opposes the idea of interpreting Hobbes as a psychological egoist who claims that all of our desires have *self-regarding content*. In other words, according to Hampton, Hobbes does allow people to have certain kinds of other-regarding desires. However, according to Hampton, these other-regarding desires play absolutely no role in Hobbes’s political argument that it is not entirely unreasonable to regard Hobbes as a psychological egoist when one is trying to understand his political philosophy.

<sup>5</sup>See [Gauthier, 1984, p. 167]

government ultimately relies. In what follows, I will argue that, concerning any substantial assumptions about human psychology and motivation, the denial of psychological altruism is all that we need to adequately explain how universal conflict in Hobbes's state of nature inevitably occurs.

Let's return back to our discussion of the axioms of the state of nature. If only modest people were in the state of nature, it might not have been so hard to achieve mutual cooperation without external enforcement. In such situations, people might be able to live in a peaceful anarchy and establishing a government might not be necessary. However, this is not the case for Hobbes's state of nature. As we have seen, in Hobbes's state of nature, it is a *known fact* that there are vain-glorious men who desire to conquer and seek power for its own sake. The major problem is that, in the state of nature, there is no reliable way to detect these vain-glorious men in advance. This leads us to our final axiom of the state of nature.

**[AXIOM(A-6-5)] (*Uncertainty*):** In the state of nature, people cannot reliably know other people's *types* or *beliefs*.

Let's think of this from the modest type's perspective. In the state of nature, a modest type will face at least one of the following two forms of uncertainties: (1) (Since the specific type of each person is not ingrained visibly on each person's forehead) in most cases, a modest type will be uncertain about his/her counterpart's *type* - that is, whether he/she is dealing with another modest type like him/her or whether he/she is dealing with a vain-glorious type. (2) Even when modest type *A* somehow gets to *know* that he/she is dealing with another modest type *B*, modest type *A* will still be uncertain about whether modest type *B* *believes* that he/she is dealing with another modest type like him/her or whether modest type *B* believes that modest type *A* believes that he/she is dealing with another modest type like him/her and so on - in other words, it is possible for modest type *A* to *falsely believe* that modest type *B* *believes* that modest type *A* is a vain-glorious type, or modest type *A* might *falsely believe* that modest type *B* *believes* that modest type *A* *believes* that modest type *B* is a vain-glorious

type (and so on), and thereby, adapt his/her actions accordingly.

According to Hobbes, these two kinds of uncertainties that permeate through out the state of nature causes *fear* and *diffidence*.

In men's mutual fear ... I mean by that word any anticipation of future evil. ... Even the strongest armies fully ready for battle, open negotiations from time to time about peace, because they fear each other's forces and the risk of being beaten. Men take precautions because they are afraid... [Hobbes 1997: *On the Citizen*, Chapter 1, p. 25]

And from this diffidence of one another, there is no way for any man to secure himself so reasonable as anticipation, that is, by force or wiles to master the persons of all men he can, so long till he see no other power great enough to endanger him. [Hobbes 1994: *Leviathan*, Chapter XIII, Section 4]

And, as it is indicated by this last passage, it emerges from all of these facts that initiating a preemptive attack becomes the optimum strategy for everybody living in the state of nature. This leads us to the main theorem of Hobbes's state of nature which can be stated as follows:

**[THEOREM(T-6-1)] (*War of Every Man against Every Man*):** The state of nature results in a state of war of every man against every man.

It is one thing to informally say that the state of nature will dissolve into a state of war by the five axioms of the state of nature; it is another thing to rigorously show that this is indeed the case. Our job in this chapter is to provide a game-theoretic model that is intended to show that theorem T-6-1, indeed, follows from the five axioms of the state of nature in a slightly more systematic and rigorous way.

Based on axiom A-6-5, we can see that *uncertainty* is one of the most important factors that causes unwanted conflict in Hobbes's state of nature. If there were no uncertainties in the state of nature, it would have been quite possible for the modest type of people to exclude the

vain-glorious type of people and cooperate with one another peacefully. The primary reason why this is not possible is because of uncertainty.

Unfortunately, uncertainty has mostly been neglected in contemporary scholarship that has tried to explain the universal conflict in the state of nature in the lights of contemporary game-theory. This deficiency is what I intend to supplement in this current chapter.

## **6.3 Hobbes’s State of Nature as a “Prisoner’s Dilemma (PD)” Game**

### **6.3.1 The Four Desiderata of Hobbes’s State of Nature**

Any game-theoretic model that attempts to represent Hobbes’s state of nature correctly must try to meet the following *desiderata*:

1. It must show that universal warfare is the *equilibrium* of the state of nature:
2. It must show that universal war is *sub-optimal* (i.e. *Pareto-inferior*): that is, it must show that there is a social state (i.e. universal peace) which everybody would strictly prefer to the state of universal war

The game-theoretic model would also have to incorporate the following factors into the model:

3. In the state of nature, there are two different types of people - the modest type and the vain-glorious type - who respectively have different motivations which are manifested in their respective preferences. (Axioms A-6-3 and A-6-4)
4. Each person in the state of nature is *uncertain* about his/her opponent’s *type* as well as his/her opponent’s *beliefs*. (Axiom A-6-5)

### 6.3.2 The PD (Prisoner's Dilemma) Game

Many people have tried to model Hobbes's state of nature as a PD (Prisoner's Dilemma) Game.<sup>6</sup> The PD game is a very attractive model to represent Hobbes's state of nature, and it might be useful at this point to briefly explain the main structure of the PD game.

A PD game represents a strategic situation consisting of two players each with an option to either cooperate or defect. Suppose that  $(x,y)$  denotes a state of the world where player 1 plays action  $x$  and player 2 plays action  $y$ . A strategic situation is a PD game if and only if player 1 and player 2 each have the following preference-orderings:

- Player 1's preference-ordering:

$(\text{Defect, Cooperate}) \succ (\text{Cooperate, Cooperate}) \succ (\text{Defect, Defect}) \succ (\text{Cooperate, Defect})$

- Player 2's preference-ordering:

$(\text{Cooperate, Defect}) \succ (\text{Cooperate, Cooperate}) \succ (\text{Defect, Defect}) \succ (\text{Defect, Cooperate})$

We can see that the preferences of player 1 and player 2 are *symmetric*. Both players *most prefer* to defect while the other player cooperates, and *least prefer* to cooperate while the other player defects. Both players also prefer mutual cooperation to mutual defection. The situation can be conveniently represented by the following game-matrix.

---

<sup>6</sup>See [Rawls, 1971, 1999, p. 269], [Taylor, 1976, Chapter 6], [Barry, 1965, pp. 253-254], [Gauthier, 1969, pp. 76-89].

Table 6.1: The PD Game

7

Player 1 \ Player 2	Cooperate	Defect
Cooperate	2,2	4,1
Defect	1,4	<b>3,3</b>

The number written on the left side of the comma signifies player 1’s order of preference while the number written on the right side of the comma signifies player 2’s order of

---

<sup>7</sup>Usually, there is a story that goes with the PD game. As a matter of fact the story is how the PD game got its name: “*Prisoner’s Dilemma*.” The story goes something like this:

Two suspects are arrested by the police. The police believe that the two suspects have jointly committed an egregious crime. However, the police lack sufficient evidence to charge the two suspects with the egregious crime that they quite confidently believe that the two suspect have committed; the police only have enough evidence to charge the two suspects with a minor offense. So, in order to charge the two suspects with the egregious crime, it is necessary for the police to receive confession from the two suspects. In order to induce confession, the police put the two suspects into two separate interrogation rooms rendering communication between the two suspects impossible. The police propose the following deal to each of the suspects: “If both of you remain silent, then both of you are going to each serve 1 year in prison for the minor offense that we are able to charge with our available evidence. However, if one of you confesses while the other remains silent, the one who confesses will get parole and will be immediately released for cooperating with the investigation, while the other who remained silent will be fully charged with the egregious crime and serve 10 years in prison. If both of you confess, then both of you will each serve 5 years in prison, which is a slightly reduced sentence for the egregious crime that you both have jointly committed. So, what are you going to do; confess? Or remain silent?”

The situation can be summarized by the following game-matrix.

Suspect 1 \ Suspect 2	Remain Silent	Confess
Remain Silent	1 year, 1 year	10 years, 0 years
Confess	0 years, 10 years	<b>5 years, 5 years</b>

Given that the two suspects care only about the number of years that each serves in prison, we can see that each suspect achieves a better outcome (i.e. serves lesser years in prison) by confessing regardless of what the other suspect decides to do. The result is that both suspects confess and achieve what each considers the second worst outcome.

Table 6.2: The Story of the Prisoner’s Dilemma

preference. Note that, here, the numbers are *not* meant to be *utilities*, but simply *the ranking* in each player's preference-ordering.

In a game of prisoner's dilemma, each player has a *strictly dominant strategy*; namely, to *defect*. As we can verify from the above game-matrix as well as each player's preference-orderings summarized above, each player obtains a more preferable outcome by defecting *regardless* of what the other player chooses to do.

The logic behind this reasoning is this: "The other player can either cooperate or defect. If the other player cooperates, I get a more preferred outcome by defecting. If the other player defects, I get a more preferred outcome by defecting. So, either way, I get a more preferred outcome by defecting." In this case, we say that the act of defecting *strictly dominates* the act of cooperating, which means that defecting generates a strictly preferable outcome for every action that the other player can take.

Knowing that defection strictly dominates cooperation, both players in the PD game will choose to defect and, thereby, (defect, defect) becomes the unique (Nash) *equilibrium* of the PD game. (I have indicated the equilibrium in the game-matrix by putting the numbers inside the cell **bold**.) An *equilibrium* of a game is a situation where every player is *best-responding* to every other player's actions, and everybody's actions are consistent with one another in the sense that nobody has any incentives to deviate from his/her current action given the actions of others. This makes the (Nash) equilibrium a *stable point*.

The interesting thing about the PD game is that the unique equilibrium of the game is *sub-optimal*. That is, as we can verify from the game-matrix and the preference-orderings of each player summarized above, we can see that *both* players prefer the social state (cooperate, cooperate) rather than the social state (defect, defect) which is the unique equilibrium of the game. This means that it is possible to enhance the situation of somebody (actually, in this case, both players) without worsening the situation of anybody else. In economics jargon, such enhancement is called a *Pareto-improvement*. A situation is sub-optimal whenever a

Pareto-improvement is possible.

And, this is exactly what makes the prisoner's dilemma a well-known *paradox*; the paradox of the PG game consists in the fact that *the social equilibrium state does not coincide with the social optimum state*.<sup>8</sup> Historically, the PD game has been used as a counter argument against Adam Smith's dictum that claims that society can always achieve social optimum (by the guidance of "the invisible hand") by simply letting individuals pursue their own preferences. We can see that this is not the case for the PD game; each player pursuing his/her own preferences results in an equilibrium that is sub-optimal. This is true for any situation that reflects the structure of the PD game.

### 6.3.3 Some Common Misunderstandings of the PD Game

Before we move on, I think that it would be helpful to correct some misunderstandings that are commonly associated with game theory (or "rational choice theory" in general, of which game theory is conceived to be a part) and, more specifically, the PD game. The misunderstandings generally stems from (mistakenly) thinking that game theory is committed to some highly contestable substantial theory of human nature and human motivation; that human beings either *are* or *should be* strictly egoistic and self-interest-maximizing beings.

Based on such general assumption about game theory, critics tend to make either of the following two objections.

- *Objection 1.* Game theory is defective as a normative theory of action; it urges one to care only about one's own self-interest when one ought to care about other things - such as morality, good citizenship, the common public good - as well.

---

<sup>8</sup>Let me elaborate this a little bit more. It is not simply the fact that both players end up in a sub-optimal situation that makes a PD game a paradox. It is no paradox in itself that people can end up in suboptimal situations. The paradox of the PD game consists in the fact that both players end up in a *sub-optimal* social state even when both players are perfectly rational in the sense that both are *best-responding* to each other's strategy given each of their preferences (note that this is exactly what is meant by saying that the social state is an *equilibrium*.) In short, the paradox consists in *the social equilibrium state not coinciding with the social optimum state*.

- *Objection 2.* Game theory is defective as a descriptive theory of action; it assumes that people, as a matter of fact, care only about their own self-interests even when they apparently do not.

Let me comment on each of these objections in turn.

### **Game Theory is Defective as a *Normative Theory of Action***

When one is first introduced to game theory and the PD game, it is very easy for one to understand game theory as recommending a certain *prescription* in the PD situation; that is, game theory might seem to say, in a PD game, defection is rational, and this might seem to imply that game theory recommends defection in the PD game.

Understood in this way, it seems that game theory is recommending people to be *selfish*; that is, it seems that it is urging people only to care about their *narrow self-interest* rather than to cooperate with other people even when such cooperation is possible. For instance, in his article, “The Rational Choice Approach to Politics: A Challenge to Democratic Theory”, Mark Petracca claims,

In the main, proponents of rational choice theory “assume that it is egoistically, individualistically, irrational not to maximize one’s satisfactions and seek one’s own greatest good.” [Petracca 1991, p. 296]

Many people find such conclusion rather distasteful. To them, even if it is true that defecting in the PD game would maximize one’s self-interest, there could be other considerations, such as a *moral reason*, that dictates one to cooperate rather than to defect in the PD game. (For instance, maybe, the two suspects in the original prisoner’s dilemma story made a *promise* not to confess if they happen to get interrogated by the police.) Some people might think that such moral reason should override any reason that stems from purely egoistic considerations. To them, game theory ignores such moral reasons or any other considerations that

are not directly relevant to maximizing one's own self-interest by relying on a very narrow conception of rationality. According to Petracca,

The influence of values, ethics, and ideas on individual motivation are alien to rational choice theories of human nature. By this account, public-spirited behavior or behavior motivated by other-regarding motives is not only irrational, but highly unlikely. [Petracca 1991, p. 297]

However, according to these people, urging people to become somebody who only cares about his/her own narrow self-interest is not a proper way to cultivate democratic citizenship to people who should rightfully care about things such as democratic deliberation and the common public good. And, hence, they think game theory is defective as a *normative theory of action* on which any political philosophy should be based.

Such objection against game theory and the PD game is misplaced. First of all, it is a mistake to think of game theory as a normative theory of action that tells people how they *should* act when they are in Prisoner's-Dilemma-like situations. Rather, the primary purpose of game theory is to learn what social state will emerge as a *stable equilibrium point* as a result of strategic interactions among two (or more) people who each have the respective preferences that the specific game assumes them to have, and to see what kind of properties that this stable equilibrium point has. If a strategic situation has a PD structure, then game theory shows that universal defection will be the stable equilibrium point, and that this stable equilibrium point will be sub-optimal. This is simply a conceptual (or a mathematical) truth: in a PD game, it is true, by definition (of Nash Equilibrium, Pareto-Optimality, and each player's preference-orderings), that both players defect and that this is the unique sub-optimal equilibrium of the game. Whether a real-life social situation truly has a PD game structure is something that we would have to determine outside of game-theory, empirically.

Second, although it is true that game theory assumes that it is *rational* for one to defect in

the PD game, we should be very careful not to interpret this as claiming that one *should* defect whenever one encounters a PD-game-like real-life situation. When game theory deems it rational for each player to defect, this is so *given that* the two players described in the PD game already have the preference structures that they are assumed to have. That is, what game theory is claiming is this: *given that* each player prefers the outcome that each player would obtain by defecting regardless of what the other player does, it is rational for each player to defect. This is similar to saying that it is rational to choose an apple over an orange *given that one prefers having an apple to having an orange*. However, one should be clear that this is not to say that one should prefer an apple over an orange in the first place. Similarly, game theory does not claim that it is rational for people to have a preference-ordering that renders defecting a strictly dominant strategy, and makes their interaction an instance of a PD game in the first place. Concerning the question of *what sort of preferences people should have*, game theory does not take any stance.

In short, game theory is *not* a normative theory of action that tells people what sort of preferences they should have. Rather, it is simply a mathematical model that represents a given strategic interaction between two or more people who are assumed to have particular preference-orderings. Other than requiring a minimum set of consistency requirements, game theory does *not* suggest what type of preference people *should* have; it only tells us what *would* happen if people do have those type of preferences that are already assumed in the model.

If there are any normative conclusions that we might be able to draw from the PD game, it would be that there can be certain social situations where the social structure itself could cause a sub-optimal social equilibrium to emerge, and that whenever we confront a social situation that resembles the structure of the PD game it might be recommendable to alter the incentive structure of the situation in order to restore the social optimum and achieve a Pareto-improvement.<sup>9</sup>

---

<sup>9</sup>Such things are usually done in the field that is now known as “mechanism design.”

## Game Theory is Defective as a *Descriptive Theory of Action*

To this, the objector might raise another objection of the following line: regardless of whether or not game theory is intended to be a normative theory of action, it is *even* defective as a *descriptive theory of action*. This is because, according to these critics, empirical evidence has shown that considerations of self-interests play a very marginal role in actual human beings' real-life actions and motivations.

... a growing body of empirical research in a variety of social science disciplines shows the explanatory limits of the rational choice approach to human nature.

... Tom Tyler's recently published study of why people obey the law shows that normative values about distributive and procedural justice matter in the motivation of individual behavior. In a study of randomly selected citizens in Chicago, Tyler made this important discovery:

"People obey the law because they believe that it is proper to do so, they react to their experiences by evaluating their justice or injustice, and in evaluating the justice of their experiences they consider factors unrelated to outcome, such as whether they have had a chance to state their case and been treated with dignity and respect. On all these levels people's normative attitudes matter, influencing what they think and do. (Tyler 2006, *Why People Obey the Law*, p. 178)"

[Pettracca 1991, PP. 300-301]

Similar empirical findings have been found in the study of PD games in real-life situations: it has been confirmed by many experiments that people participating in a game that mimics the structure of the PD game tend to cooperate far more often than what game theory predicts.

10

---

<sup>10</sup>See [Dawes and Thaler, 1988, Cooper et al., 1996, Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games]

However, what the results of these empirical experiments really show is *not* that there is any fault in game theory's predictive or descriptive power, but merely that many people do not have the preferences that would make their interaction in the experiments instances of a PD game.

For example, suppose that an experimenter randomly picks two people from a group and makes them play the following game: Each player can choose either to "cooperate" or "defect." When one player cooperates while the other defects, the person who cooperated *pays \$1* while the person who defected *receives \$2*. If both players cooperate, then both players *receive \$1*. If both players defect, then both players *receive nothing*. Moves are made simultaneously.

*Given* that the two players care only about the amount of money they receive, we can see that the experiment has exactly the same structure as the PD game. However, suppose that, after many trials of the experiment, it turned out there were very many cases where the two players chose to cooperate rather than to defect.

It is very easy to think that such experiment falsifies a major assumption as well as a general prediction of game theory; that people care only about promoting their own self-interest which would render the unique (Nash) equilibrium of the situation to be universal defection. On the contrary, what the experiment really shows is merely that money is not the only thing that people in general care about. And, the claim that people *do care* or *should care* only about money is *not* a part of game theory.

People participating in the experiments might care about their reputation, etiquette towards strangers, etc., and they might have thought that winning an extra dollar is not worth compromising any of these things. If this explanation is correct, then this means that the preference-orderings of the people who were participating in the experiments were very likely to be *not*:

- (Defect, Cooperate)  $\succ$  (Cooperate, Cooperate)  $\succ$  (Defect, Defect)  $\succ$  (Cooperate, De-

fect)<sup>11</sup>

but, something like:

- (Cooperate, Cooperate)  $\succ$  (Defect, Cooperate)  $\succ$  (Defect, Defect)  $\succ$  (Cooperate, Defect)

If this is so, then what the people participating in the experiment were playing was *not* a PD game, but something like the game of *Stag Hunt*.

As I have explained, game theory is simply a mathematical model, most of the central assertions of which are merely truths by definition. What game theory assumes is that people generally choose according to their preferences and that these preferences conform to a minimum set of consistency requirements (e.g. transitivity.) However, game theory is silent on the issue of what specific preferences people do have or should have.

Different set of preferences (among the players) results in a different game. If people's real-life preferences happen to roughly conform to the preferences of the players in a specific game-theoretic model, then the equilibrium of that specific game is a good predictor of what type of social situation will eventually emerge as a result of those people's interactions. However, (unsurprisingly) if people's preferences are misrepresented, then the resulting game-theoretic model will very likely give false predictions. This does not show that there is any intrinsic fault in game theory; it merely shows that we have chosen the wrong game to represent the situation.

In short, game theory is not committed to any substantial theory of human psychology; specifically, game theory does *not* claim that people *are* or that they *should be selfish* (e.g. that they (should) care only about money, reducing their years in prison, and so on.) Game theory does not deny that people's preferences can be based on other things - such as their

---

<sup>11</sup>Where (X, Y) denotes a situation where Player 1 plays action X while the other Player 2 plays action Y

moral or religious convictions, their sense of right and wrong, and certain types of other-regarding desires. Therefore, it is a mistake to object to game theory by claiming that it is defective either as a normative or a descriptive theory of human action on these grounds.

## 6.4 Why Hobbes's State of Nature is *Not* a PD Game

It is understandable why so many people have been attracted to the idea of modeling Hobbes's state of nature as a PD game.

First of all, in a PD game, the act of defection strictly dominates the act of cooperation and, thereby, universal defection is the unique equilibrium of the game. If the state of nature is seen as a PD game, then this explains very well why the state of nature, according to Hobbes, inevitably results in a state of universal war. So, modeling Hobbes's state of nature as a PD game meets the first desideratum that we have seen in the beginning of the previous section.

Secondly, the unique equilibrium of a PD game, (namely, the state (defect, defect)) is *sub-optimal*; that is, there is a state (namely, the state (cooperate, cooperate)) which both players in the game would strictly prefer over the equilibrium. This corresponds very well with the misery and the insecurity that Hobbes associates with the life in the state of nature, and supports Hobbes's own justification for establishing a government that has the power to enforce peace. This shows that modeling Hobbes's state of nature as a PD game meets the second desideratum as well.

What all this shows is that the PD game is a very attractive game to model Hobbes's state of nature. However, modeling Hobbes's state of nature by a PD game has the problem of misrepresenting what Hobbes deems to be the major cause of conflict in the state of nature.

It is true that Hobbes thinks that everybody in the state of nature has a tendency to initiate a preemptive attack and start a war of all against all. However, as we have already seen in section 8.2, Hobbes explicitly states that not everybody is inclined to initiate a preemptive

attack *for the same reason*.

Let's go back (despite pain of repetition) to some of the passages that come from Hobbes's major texts:

In the state of nature there is in all men a will to do harm, but *not for the same reason* or with equal culpability. One man practices the equality of nature (...) this is the mark of *modest man* (...) Another, supposing himself superior to other, wants to be allowed everything (...) that is the sign of *an aggressive character*. In his case, the will to do harm derives from *vainglory*. [Hobbes 1997: *On the Citizen*, Chapter 1, Section 4 emphasis added]

Also, because there be *some* that taking pleasure in contemplating their own power in the acts of conquest, which they pursue farther than their security requires, if *others* (*that otherwise would be glad to be at ease within modest bounds*) should not by invasion increase their power, they would not be able, long time, by standing only on their defence, to subsist. [Hobbes 1994: *Leviathan*, Chapter XIII, section 4 emphasis added]

We can see here that, according to Hobbes, the state of nature involves two different types of people: (a) the modest person and (b) the vain-glorious person.

The vain-glorious person is inclined to initiate a preemptive attack regardless of whether or not his/her opponent wishes to cooperate simply because he/she enjoys conquest and having power over others. The modest person would be very glad to cooperate with his/her opponent given that there is guarantee that his/her opponent will also cooperate in return. However, the modest person is also inclined to initiate a preemptive attack because he/she *lacks assurance* that his/her opponent will not initiate a preemptive attack against him/herself: either because his/her opponent is a vain-glorious person him/herself or because his/her opponent is a modest person who mistakenly believes that he/she is dealing with a vain-glorious

person.

And from this *diffidence*<sup>12</sup> of one another, there is no way for any man to secure himself so reasonable as anticipation, that is, by force or wiles to master the persons of all men he can, so long till he see no other power great enough to endanger him. [Hobbes 1994: *Leviathan*, Chapter XIII, section 4 emphasis added]

However, this is not the situation that is described in the PD game. In the PD game, both players have *exactly the same* preference-orderings; *both* players strictly prefer to defect even when there is a guarantee that the other player is going to cooperate. If we translate this to Hobbes's the state of nature, this would imply that everybody in Hobbes's state of nature would prefer to initiate a preemptive attack even when there is a guarantee that the other party will cooperate and seek mutual peace. In other words, modeling Hobbes's state of nature as a PD game implies that *everybody* in the state of nature is *vain-glorious*.

This directly conflicts with what Hobbes says in the passages that we have just seen above, which explicitly distinguishes between two types (i.e. the modest type and the vain-glorious type) of people. This means that modeling Hobbes's state of nature as a PD game fails to meet the third desideratum that we have seen in section 8.3.

Furthermore, the primary reason why people dwelling in Hobbes's state of nature lack assurance that the other party will not initiate a preemptive attack is, as we have seen, that people are *uncertain* both about the other party's *type* (i.e. whether the other party is modest or vain-glorious) as well as about the other party's *beliefs* (i.e. whether the other party *believes* that *I* am modest or vain-glorious.) This means that the game theoretic model that aims to represent Hobbes's state of nature should include aspects of uncertainty into the entire picture.

---

<sup>12</sup>Here, it is worth mentioning that Hobbes is using the word "diffidence" in the archaic sense in which it means suspicion or distrust.

However, one should note that there are no aspects of uncertainty involved in the PD game. The PD game (using the terminology of game theorists) is a *complete information* game; that is, each player is completely aware of the other player's preferences, pay-offs, what type of strategies are available to each player, how many times the game will be played in what sequence and so on. As we have seen, this is not how Hobbes describes the situation in the state of nature where uncertainty is one of its most characteristic features as well as the main cause of conflict. In short, the PD game fails to meet the fourth desideratum that we have seen in section 8.3.

What all this shows is that the PD game, despite having some notable features that could be used to explain the universal conflict in Hobbes's state of nature, does not fit very well with what Hobbes describes in his own text; it under-represents some of the key features (i.e. *different types of people* and *uncertainty*) which Hobbes deems to be the main source of conflict in the state of nature. In other words, although modeling Hobbes's state of nature as a PD game meets the first two desiderata, it fails to meet the third and fourth desiderata we have seen in the previous section.

However, independently of whether the PD game fits with Hobbes's original text well or not, it should be noted that modeling the state of nature as a PD game has an additional problem of significantly weakening the major purpose of Hobbes's political philosophy; which is to justify the existence of governments. As I have already mentioned in the previous section, many experiments that have been led by behavioral economists show that people tend to cooperate much more often in games that were designed to mimic the structure of the Prisoner's Dilemma. This suggests that people might not actually play a PD game were they situated in Hobbes's-state-of-nature-like situations where there is no government power to enforce laws.

This suggests that the argument that people will engage in universal warfare in the state of nature because they will be playing the PD game is quite likely to be at odds with empirical human psychology. The more one's justification for the existence of governments is based

on a premise that is at odds with empirical human psychology, the more it loses practical force and plausibility.

This last point suggests that even if Hobbes's own text really did suggest that the state of nature is a PD game, it might have been advisable for contemporary scholars to find alternate models simply to boost the plausibility of Hobbes's justification for the existence of governments by modeling Hobbes's state of nature in an alternate way. However, as we have seen, one does not even need to go that far, since there is more than enough textual evidence that shows that Hobbes did not think that the primary cause of universal warfare in the state of nature was due to everybody being dominated by a basic passion for vain-glory, which is required for the state of nature to be a PD game.

## 6.5 Other Alternative Models: *The Stag Hunt* and *The Iterated PD Game*

Before I present my own model, I would first like to point out that I am not the first person to express dissatisfaction of modeling Hobbes's state of nature as a PD game. Some people have suggested that Hobbes's state of nature is really not a one-shot-PD game; but, instead either a game of *Stag Hunt*<sup>13</sup> or a game of *iterated Prisoner's Dilemma*<sup>14</sup>. In this section, I will briefly explain why I think these two models are inadequate representations of Hobbes's state of nature.

---

<sup>13</sup>See [Skyrms, 2004, chapter 1], [Gauthier, 1969, p. 85]. Gauthier thinks that Hobbes's state of nature can be modeled as a PD game in the short-term, and a *Stag Hunt game* in the long-term.

<sup>14</sup>See [Kavka, 1986, Chapter 4], [Hampton, 1986, Chapter 3]

### 6.5.1 Problems with Modeling Hobbes’s State of Nature as a Game of *Stag Hunt*

The game of Stag Hunt can be summarized by the following game-matrix:

**Table 6.3: The Stag Hunt**

Player 1 \ Player 2	Cooperate	Defect
Cooperate	<b>1,1</b>	4,2
Defect	2,4	<b>3,3</b>

<sup>15</sup>

Again, the numbers signify each player’s order of preference (not utilities.) The game of Stag Hunt has two pure-strategy (Nash) equilibria (which are indicated in bold font); namely, (Cooperate, Cooperate) and (Defect, Defect), and one mixed-strategy (Nash) equilibrium which I will omit.

Just like the PD game, the Stag Hunt meets the first two desiderata of Hobbes’s state of nature; that is, (1) mutual defection is an equilibrium, and (2) mutual defection is sub-optimal.

However, what distinguishes the game of Stag Hunt from the PD game is that, unlike the PD game, mutual cooperation, along with mutual defection, is also an equilibrium. This means that if Hobbes’s state of nature is truly a game of Stag Hunt, it is quite unclear why the state of nature should inevitably dissolve into a state of universal war as Hobbes himself claims, rather than it turning out to be a state of mutual peace and harmony.

In his book, *The Strategy of Conflict*, Thomas Schelling has argued that when there are more than one equilibrium in a game, the actual equilibrium will turn out to be the one that

---

<sup>15</sup>Hampton follows Sen and calls the game an “Assurance Game.” See[Hampton, 1986, p. 67]

is *prominent* based on cultural, historical, conventional factors. Schelling has called such equilibrium a *focal point* of a game.<sup>16</sup> This means that if Hobbes's state of nature is a game of Stag Hunt, then individuals will be able to achieve peaceful harmony without government enforcement in some state of nature where there has historically been an *ethos* of mutual cooperation. As a result, in such situations, there would be no need for a government. This completely defies one of the main purposes of Hobbes's political philosophy; which is to justify the existence of governments for any population in any circumstances.

Furthermore, just like the PD game, the game of Stag Hunt does not incorporate one of Hobbes's major assumptions that in the state of nature there are two types of people (i.e. *the modest type* and *the vain-glorious type*) who respectively have distinct preference-orderings; we can see above that, in the game of Stag Hunt, there is only one type of player and the preferences of the two players are symmetric.

Also, just like the PD game, the game of Stag Hunt is a complete information game which incorporates no aspects of uncertainty. In short, not only does modeling Hobbes's state of nature as a game of Stag Hunt completely defies one of the major aims of Hobbes's political philosophy, it fails to meet the third and fourth desiderata that we have discussed previously.

### **6.5.2 Problems with Modeling Hobbes's State of Nature as an *Iterated PD Game***

What about modeling Hobbes's state of nature as an *iterated* PD game? An iterated PD game is a game where the two players play the PD game multiple times. When the PD game is played multiple time, it is possible for each player to either *reward* (by cooperating in the next round) or *punish* (by defecting in the next round) his/her opponent's behavior in the previous round. This changes the dynamics of the game significantly.

---

<sup>16</sup>See [Schelling, 1981]

If the game is played only *finitely* many times, then the game has only one equilibrium; namely, both players defecting in every period of the game.<sup>17</sup>

However, if the game is played *infinitely* many times, there are other equilibria besides the one where both players defect in every period of the game. One such equilibrium is where both players play a strategy known as *tit-for-tat*. The rule of *tit-for-tat* is simple; cooperate in your first move, and then copy what your opponent had done in the previous round. *Tit-for-tat* can be characterized as a strategy of both *punishment* and *forgiveness*: it punishes one's opponent by defecting in the current round if one's opponent defects in the previous round; but, it *forgives* and *rewards* one's opponent by cooperating in the next round if one's opponent cooperates in the current round.

There are a number of other equilibrium-strategy-pairs (besides *tit-for-tat* and *consistent defection*) in the infinitely repeated PD-game which I will not go through in detail.<sup>18</sup> What's important is that, unlike the case of an one-shot PD game, in an infinitely repeated PD game, it is possible for both players to reap the benefits of mutual cooperation for infinite number of periods by mutually employing the right kind of strategy.

Just like the one-shot PD game and the Stag Hunt, the iterated PD game meets the first two desiderata of Hobbes's state of nature.

However, besides meeting the first two desiderata, modeling Hobbes's state of nature as an iterated PD game has its own merits. The most significant merit is that it seems to explain the universal warfare that is characteristic of the state of nature, while, at the same time, show how people can *escape* the state of nature and successfully establish a government by themselves. As I have just briefly explained, although it is true that both players defecting in

---

<sup>17</sup>This can be proved by *backward induction*.

<sup>18</sup>These other equilibrium-strategy-pairs can be distinguished by *the severity* of the punishment that each strategy prescribes when one first encounters defection by the other player. The *grim-trigger* strategy (i.e. the strategy of no forgiveness) prescribes one to cooperate until one first encounters defection by the other player, in which case it prescribes to consistently defect afterwards. The strategy of *limited-punishment* prescribes one to initially cooperate, and, when one first encounters defection by the other player, it prescribes one to punish the other player by defecting for a given number (*n*) of periods. With an adequate discount rate, it can be shown that both players playing either the *grim-trigger* strategy or the strategy of *limited-punishment* can both be equilibria in an infinitely repeated PD game.

every period of the game is an equilibrium, there are other equilibria where both players are able to mutually cooperate throughout the game. These latter equilibria open possibilities for people to escape the predicament they face in the state of nature.<sup>19</sup>

However, the problem is not that simple. One problem is whether it is really plausible to think of the interaction among the people living in the state of nature as a repeated PD game. The iterated PD game requires each player to play the PD with *the same opponent* repeatedly.

I doubt that this would be the case for people living in the state of nature. In the state of nature, it would be far more likely for each person to randomly encounter a different opponent everytime they happen to interact with somebody. If this is so, then it might be more plausible to model Hobbes's state of nature as a one-shot game, rather than some repeated game.

Even if one happens to interact with the same person more than once, such interaction cannot be repeated *infinite number of times* in the state of nature. This is because, in the state of nature, interaction with other people can, in many case, result in the death of one of the parties. This means that Hobbes's state of nature can, at best, be modeled as a *finitely repeated* PD game.

However, we have seen that in a finitely repeated PD game, mutual defection for all periods of the game is the only equilibrium of the game.<sup>20</sup> This takes away a major attraction of modeling Hobbes's state of nature as an iterated PD game; namely, the fact that it shows how people can escape the state of nature and successfully establish a government by themselves.

Even if we concede that interaction in the state of nature can be repeated with the same

---

<sup>19</sup>Of course, the same thing can be claimed for the *Stag Hunt* as well. That is, in the game of *Stag Hunt*, the two players can achieve mutual cooperation if there happens to be widespread convention of cooperation which would make mutual cooperation the focal point of the game. However, the dynamic interaction of the iterated PD game seems to better explain how universal cooperation can emerge naturally by constant interaction between the players themselves.

<sup>20</sup>Things get a little more complicated in which the two players know that they are playing a finitely repeated PD game, but do not know the number of periods. If we assume that the probability that the game will end at the next period increases as the game moves on, we can again show, by backward induction, that the unique Nash equilibrium of the game is mutual defection for all periods of the game.

person infinite number of times, modeling Hobbes's state of nature as an infinitely repeated PD game has exactly the same problems that caused problems for the Stag Hunt game. That is, since there exist multiple equilibria where both parties can naturally achieve mutual cooperation in an infinitely repeated PD game, modeling Hobbes's state of nature as an infinitely repeated PD game significantly weakens Hobbes's major argument for justifying the necessity of government. Furthermore, modeling Hobbes's state of nature as an infinitely repeated PD game fails to meet the third and fourth desiderata by not incorporating the distinction between the two types of people (i.e. the modest type and the vain-glorious type) as well as aspects of uncertainty, which Hobbes clearly assumes to exist in the state of nature, into the model.

In short, although many people have been attracted to the idea of modeling Hobbes's state of nature as an infinitely repeated PD game, it fails to be an ideal game theoretic-model that is both faithful to Hobbes's original text and that could serve Hobbes's original intentions well.

## **6.6 Modeling Hobbes's State of Nature with Bayesian Game Theory**

We have just seen that most game theoretic models that have been hitherto presented to represent Hobbes's state of nature failed to provide an adequate representation by failing to meet the third and fourth desiderata that I have stated in section 8.3.1. If one is faithful to Hobbes's original text, it is not hard to realize that it was not, strictly speaking, people's egoistic psychology that Hobbes thought to be the primary cause of universal conflict in the state of nature.

Rather, the universal conflict in the state of nature, according to Hobbes, is primarily due to *uncertainty*. Therefore, any game theoretic model that does not model uncertainty is, I

claim, an incorrect model of Hobbes's state of nature. And, in order to model uncertainty, one would have to utilize, what is known as, *Bayesian game theory*.

When one encounters strategic situations with other people, one is not always *perfectly* or *completely informed* about the nature of the strategic interaction that one is facing. That is, one can be *uncertain* in many ways. Among the many ways that one can be uncertain about a given strategic interaction, game theorists generally distinguish between *imperfect information* and *incomplete information*.

In game theory, a player is said to face imperfect information when he/she does not know *the past moves* that were played by one's strategic opponents. However, in an imperfect information game, each player is still assumed to know every aspect that relates to the structure of the game itself; such as how many players there are, what kind of preferences the players have, what kind of strategies are open to the players, and so on. Simply put, in an imperfect information game, the players know *what kind of game* that they are playing; what they do not know is how the other players played out their moves in the previous stages of the game.

In an *incomplete information* game, the nature of the uncertainty is deeper. This happens when any one or more players are uncertain about any aspect that directly relates to the structure of the game itself: such as the preferences of the other players, the kind of strategies that the other players can play, what kind of beliefs that the other players have, and so on. Simply put, in an incomplete information game, the players might not even know what kind of game that they are playing.

Game theorists knew how to analyze imperfect information games. However, it took some time for game theorists to figure out how to analyze incomplete information games. The solution came from John Harsanyi in his three sequence papers, titled, "Games with Incomplete Information Played by "Bayesian" Players, I-III" published in 1967 and 1968<sup>21</sup>, and is now introduced in most standard introductory textbooks in game theory.<sup>22</sup>

---

<sup>21</sup>[Harsanyi, 1967, 1968a,b]

<sup>22</sup>See [Osborne, 2003, Dutta, 1999, Dixit and Skeath, 2004]

The basic idea is to include an additional player called “nature” that makes the first move related to the specific aspect that the players in the game are uncertain about. The move is usually characterized as a probability distribution on the various options that the uncertain parameter can possibly take. By doing so, the incomplete information game has, in effect, been reduced into an imperfect information game, where the players do not exactly know *nature’s* past moves.

After reducing an incomplete information game into an imperfect information game, we can now utilize the same tools that we have used to analyze imperfect information games to analyze the incomplete information game that we are interested in. This is what I plan to do in this section.

### **6.6.1 First Bayesian Model of Hobbes’s State of Nature: Uncertain About the Other Person’s Type**

Let’s start with two individuals living in Hobbes’s state of nature: Bob and Jill. There are two versions of Bob and Jill; the *modest versions* (denoted by  $Bob_S, Jill_S$ ) and the *vain-glorious versions* (denoted by  $Bob_G, Jill_G$ )<sup>23</sup> All types of Bob and Jill have two available actions: they can either *seek peace* or *initiate a preemptive attack*.

Different types of Bob and Jill have different preferences. The different preferences depend on both what type one is and what type one happens to be interacting with. The preferences of the four types of players are summarized in the following four game-matrices:

---

<sup>23</sup>Again, this is a continuation of our previous chapter.  $Bob_S$  is the type of Bob who is primarily influenced by a basic passion for self-preservation, and  $Bob_G$  is the type of Bob who is primarily influenced by a basic passion from glory. The same thing applies to Jill as well.

Table 6.4: Different Types of Bob and Jill's Preferences in the State of Nature

$Bob_S \setminus Jill_S$	Peace	Attack	$Bob_S \setminus Jill_G$	Peace	Attack
Peace	<b>1,1</b>	4,2	Peace	2,2	4,1
Attack	2,4	<b>3,3</b>	Attack	1,4	<b>3,3</b>

$Bob_G \setminus Jill_S$	Peace	Attack	$Bob_G \setminus Jill_G$	Peace	Attack
Peace	2,2	4,1	Peace	2,2	4,1
Attack	1,4	<b>3,3</b>	Attack	1,4	<b>3,3</b>

Again, the numbers in each cell signify each player's order of preference. Here is a short explanation for the situation. The vain-glorious types (i.e.  $Bob_G$  and  $Jill_G$ ), being obsessed with glory, prefers to initiate a preemptive attack regardless of his/her opponent's type as well as his/her opponent's actions. The modest types most prefers to seek peace and cooperate with other modest types provided that the other modest types reciprocate one's cooperation. However, the modest types would rather initiate a preemptive attack if the other modest type is going to attack, and would prefer to initiate a preemptive attack against other vain-glorious types for preventive purposes. We can see that the game played by  $Bob_S$  and  $Jill_S$  is a game of Stag Hunt, while the other three games played by other type-combinations of Bob and Jill are PD games.

It will be convenient to assign a pair of utilities for each cell in each of the game-matrices above. In order to use the results that we have established in the previous chapter, we would need to assign consequences for each of the cells in the game-matrices above that respects each types of Bob and Jill's preference-ordering. Remember that for all types of Bob and Jill: Glorified Life  $\succ$  Moderate Life  $\succ$  Mortified Life  $\succ$  Death.<sup>24</sup> The result of putting these

<sup>24</sup>Here is a very brief explanation of what each type of lives signifies; Death literally means death; Mortified Life means a life with insufficient power (i.e. being under the power and influence of other people); Moderate

consequences in the game-matrices becomes:

Table 6.5: The Consequences that Each Type of Bob and Jill will Face by Performing Their Respective Actions

<i>Bobs</i> \ <i>Jills</i>	Peace	Attack
Peace	<b>Glorified Life, Glorified Life</b>	Death, Moderate Life
Attack	Moderate Life, Death	<b>Mortified Life, Mortified Life</b>

<sup>25</sup>

<i>Bobs</i> \ <i>Jill<sub>G</sub></i>	Peace	Attack
Peace	Moderate Life, Moderate Life	Death, Glorified Life
Attack	Glorified Life, Death	<b>Mortified Life, Mortified Life</b>

<sup>26</sup>

Life means a life with just enough power to sustain one's self-preservation; Glorified Life means a life with more than enough power (i.e. having power over others) to sustain one's long-term self-preservation.

<sup>25</sup>The interpretation for this might be something like this. When two modest type of people cooperate, both will be able to reap the benefits of mutual cooperation and will be able to enjoy mutual prosperity. This, for them, would mean a *glorified life*. If one of the modest type initiates a preemptive attack, while the other modest type seeks peace, the modest type that attacks will be able to catch the other modest type that sought peace off-guard, and will be able to kill him/her. The fact that any human being has enough physical + mental power to kill another human being comes from *the axiom of equality* that we have seen in section 8.2. When both modest types decides to attack, then both would be able to avoid being killed by the other party since both parties would be prepared for such assault. However, after the fight both parties would be battered and would be worse-off than they would otherwise be if they had both sought mutual peace. This is simply an interpretation; it is pointless to argue about the minute details of it. What's important is that the two modest type of players are playing the game of Stag Hunt, and regards mutual peace as the most preferable outcome.

<sup>26</sup>The interpretation for this might be something like this. When a modest type knows that he/she is interacting with a vain-glorious type, he/she knows that the vain-glorious type would prefer to initiate a preemptive attack regardless how he/she acts. Knowing this, the modest type would prefer to initiate a preemptive attack, and, hopefully, catch the vain-glorious type off-guard. If the modest type succeeds in doing this, he/she will be able to kill the vain-glorious type as *the axiom of equality* suggests. Successfully killing a vain-glorious type will earn the modest type a reputation that he/she is a hard-liner against vain-glorious types, which would deter other vain-glorious types from attacking him/her. This added security (or power) would mean that the modest type would obtain a Glorified Life by killing a vain-glorious type in the state of nature. Of course, the modest type's preference for a Glorified Life over a Moderate Life would not be as strong as that of a vain-glorious type. Mutual peace between the modest type and the vain-glorious type would give both parties Moderate Life. Mutual attack between the two parties would give them Mortified Life, since both would be wounded despite not being killed. It's not hard to imagine that being attacked by a vain-glorious type while the modest type seeks peace will result in the modest type's death. Again, what's important is not the specific interpretation, but the fact that when a modest type and a vain-glorious type interact, they are playing a PD game.

$Bob_G \setminus Jill_S$	Peace	Attack
Peace	Moderate Life, Moderate Life	Death, Glorified Life
Attack	Glorified Life, Death	<b>Mortified Life, Mortified Life</b>

$Bob_G \setminus Jill_G$	Peace	Attack
Peace	Moderate Life, Moderate Life	Death, Glorified Life
Attack	Glorified Life, Death	<b>Mortified Life, Mortified Life</b>

We would now need to assign, not just simply *ordinal utilities*, but, what is known as *VnM* (short for “Von-Neumann and Morgenstern”) or *expected utilities*. Unlike ordinal utilities, which represents a person’s preferences in the sense that the person strictly prefers  $x$  to  $y$  if and only if the utility of  $x$  is greater than the utility of  $y$  (i.e.  $x \succ y$  iff  $U(x) > U(y)$ ), *VnM* or *expected utilities* represent a person’s preferences in the sense that the person strictly prefers  $x$  to  $y$  if and only if the *expected* utility of  $x$  is greater than the *expected* utility of  $y$  (i.e.  $x \succ y$  iff  $EU(x) > EU(y)$ ).

There is a set of specific conditions (known as the *VnM Axioms*<sup>27</sup>) that one’s preferences must meet in order for there to be an *expected utility function* that could represent them. I will not go through the VnM Axioms in this paper. I will simply assume that all types of Bob and Jill meet the VnM Axioms.

Assuming that all types of Bob and Jill meet the VnM axioms, there exist expected utility functions that represent each type of Bob and Jill’s preferences over various gambles that include the four type of lives (i.e. Glorified Life, Moderate Life, Mortified Life, Death) as their prizes. For our current model, I assign the following expected utilities for each of

---

<sup>27</sup>Some of the major VnM Axioms include *independence*, *Archimedean*, continuity (in some systems), *asymmetry*, *negative transitivity*, and so on. For those who are interested in expected utility theory in general and would like to know what these VnM Axioms signify, please refer to the following books. [Fishburn, 1970, Kreps, 1988, Luce and Raiffa, 1957, Resnik, 1987, Von Neumann and Morgenstern, 1944] or any standard graduate-level textbook in decision theory.

the four type of lives in the state of nature.<sup>28</sup>

$$\bullet U_{Bob_S, Bob_G, Jill_S, Jill_G}(\text{Death}) = 0$$

<sup>28</sup>The values that I assign here are not completely arbitrary. I will briefly explain the process by which I have determined these values.

Given that each type of Bob and Jill's meet the *VnM* axioms, we can find an expected utility function that represents each type of Bob and Jill's preferences over any gambles consisting of the four type of lives we have listed above. Furthermore, if we find any one expected utility function that represents a given type of Bob or Jill, we can find another expected utility function that represents the given type of Bob or Jill's preferences by performing a *positive affine transformation* of the original expected utility function.

Let's start assigning utilities for the four type of lives for each type of Bob (i.e.  $Bob_S$  and  $Bob_G$ ) and Jill (i.e.  $Jill_S$  and  $Jill_G$ .) I will start with the two types of Bob; exactly the same process can be applied to the two types of Jill afterwards.

First, start out by normalizing each type of Bob's utility scale by assigning  $U_{Bob_S, Bob_G}(\text{Glorified Life})=1$  and  $U_{Bob_S, Bob_G}(\text{Death})=0$  for both  $Bob_S$  and  $Bob_G$ . We then find the two values  $m_1^S$  and  $m_2^S$  ( $0 \leq m_1^S, m_2^S \leq 1$ ) that would make  $Bob_S$  indifferent to achieving a Moderate Life and playing the gamble  $[m_1^S \cdot \text{Glorified Life}; (1 - m_1^S) \cdot \text{Death}]$ , and would also make  $Bob_S$  indifferent to achieving a Mortified Life and playing the gamble  $[m_2^S \cdot \text{Glorified Life}; (1 - m_2^S) \cdot \text{Death}]$ . We then assign the utilities  $U_{Bob_S}(\text{Moderate Life}) = m_1^S$  and  $U_{Bob_S}(\text{Mortified Life}) = m_2^S$ .

We can see that the function  $U_{Bob_S}$  has the expected utility property - that is, the utility of a gamble is its expected utility (for instance,  $U_{Bob_S}([m_1^S \cdot \text{Glorified Life}; (1 - m_1^S) \cdot \text{Death}]) = m_1^S \cdot U_{Bob_S}(\text{Glorified Life}) + (1 - m_1^S) \cdot U_{Bob_S}(\text{Death}) = m_1^S$ ) - and represents  $Bob_S$ 's preferences in the sense that  $Bob_S$  strictly prefers gamble  $x$  to gamble  $y$  if and only if the expected utility of the gamble  $x$  is greater than the expected utility of gamble  $y$  (for example, we know that  $Bob_S$  is indifferent between Moderate Life and the gamble,  $[m_1^S \cdot \text{Glorified Life}; (1 - m_1^S) \cdot \text{Death}]$  and we can see that the expected utilities of Moderate Life and the gamble  $[m_1^S \cdot \text{Glorified Life}; (1 - m_1^S) \cdot \text{Death}]$  are both  $m_1^S$ .) We can assign  $U_{Bob_G}(\text{Moderate Life}) = m_1^G$  and  $U_{Bob_G}(\text{Mortified Life}) = m_2^G$ .

We, now, determine the the respective ranges that each value  $m_1^S, m_2^S, m_1^G, m_2^G$  can take. First, remember that  $Bob_S$  is the modest type of Bob who is being influenced by a basic passion for self-preservation. This implies that, within  $Bob_S$ 's preference-ordering, moving up from Death to Mortified Life would worth more than moving up from Mortified Life to Moderate Life *or* moving up from Moderate Life to Glorified Life. This implies that  $|U_{Bob_S}(\text{Mortified Life}) - U_{Bob_S}(\text{Death})| = m_2^S - 0 = m_2^S > |U_{Bob_S}(\text{Moderate Life}) - U_{Bob_S}(\text{Mortified Life})| = m_1^S - m_2^S$  and  $|U_{Bob_S}(\text{Mortified Life}) - U_{Bob_S}(\text{Death})| = m_2^S - 0 = m_2^S > |U_{Bob_S}(\text{Glorified Life}) - U_{Bob_S}(\text{Moderate Life})| = 1 - m_1^S$ .

Furthermore,  $Bob_S$  (unlike  $Bob_G$ ) would not attach much value to glory (i.e. power) *per se*;  $Bob_S$ 's concern for power would be limited to his interests in securing his self-preservation. Therefore, for  $Bob_S$ , moving up from Moderate Life to Glorified Life would not be worth more than moving up from Mortified Life to Moderate Life (which means moving up from an abject life to a decent life.) This implies that  $|U_{Bob_S}(\text{Moderate Life}) - U_{Bob_S}(\text{Mortified Life})| = m_1^S - m_2^S > |U_{Bob_S}(\text{Glorified Life}) - U_{Bob_S}(\text{Moderate Life})| = 1 - m_1^S$ .

Summarizing all of this we get:

- (a)  $m_1^S > m_2^S$  (i.e. a Moderate Life is strictly preferred to a Mortified Life)
- (b)  $2m_2^S > m_1^S$
- (c)  $m_1^S + m_2^S > 1$
- (d)  $2m_1^S > 1 + m_2^S$

Of course, there will be more than one set of values for  $m_1^S$  and  $m_2^S$  that satisfy the inequalities (a), (b), (c), (d). However, as one can easily confirm by simple algebra, there are *lower bounds* for both  $m_1^S$  and  $m_2^S$ , which are:

$$\bullet m_2^S > \frac{1}{3}$$

(Multiplying 2 to each side of (b) and connecting it with (d), we get:  $4m_2^S > 2m_1^S > 1 + m_2^S$ . From this,

- $U_{Bob_G, Jill_G}(\text{Mortified Life}) = 1/4$

---

we get:  $m_2^S > \frac{1}{3}$ .)

- $m_1^S > \frac{2}{3}$   
(From the fact that  $m_2^S > \frac{1}{3}$  and (d), we get:  $2m_1^S > 1 + m_2^S > 1 + \frac{1}{3} = \frac{4}{3}$ . From this, we get:  $m_1^S > \frac{2}{3}$ .)

Summarizing this, we get:

- $U_{Bob_S}(\text{Glorified Life}) = 1$ ,
- $U_{Bob_S}(\text{Moderate Life}) = m_1^S > \frac{2}{3}$
- $U_{Bob_S}(\text{Mortified Life}) = m_2^S > \frac{1}{3}$
- $U_{Bob_S}(\text{Death}) = 0$

We now go over the same process for  $Bob_G$ . Obviously,  $Bob_G$ , who is primarily influenced by the basic passion for glory, will have different preferences towards the various gambles that include Glorified Life and Death as their prizes.

It is not hard to expect that such type of Bob would value a glorified life extremely highly. So, for  $Bob_G$ , moving up from Moderate Life to Glorified Life would definitely worth more than moving up from Mortified Life to Moderate Life or moving up from Death to Mortified Life in  $Bob_G$ 's.

Furthermore, the fact that  $Bob_G$  strongly desires glory and honor implies that he would have a strong aversion against dishonor and mortification. Of course,  $Bob_G$ 's aversion against dishonor would not be so strong to the extent that he would rather prefer Death to a Mortified Life: but his preference for a Mortified Life would not be that strong. This implies that, for  $Bob_G$ , moving up from Death to Mortified Life would worth less than moving up from Mortified Life to Moderate Life.

Let  $m_1^G$  and  $m_2^G$  respectively be the values of  $p$  in the lottery [ $p$ ·Glorified Life +  $(1 - p)$ ·Death] to which  $Bob_G$  would be indifferent to securing a Moderate Life and a Mortified Life. By summarizing all of the facts above, we get the following four inequalities:

- (a)  $m_1^G > m_2^G$  (i.e. a Moderate Life is strictly preferred to a Mortified Life)
- (b)  $1 - m_1^G > m_1^G - m_2^G$
- (c)  $1 - m_1^G > m_2^G$
- (d)  $m_1^G - m_2^G > m_2^G$

Again, there will be more than one set of values of  $m_1^G$  and  $m_2^G$  that would satisfy these four inequalities. However, we are able to derive the *lower bounds* of  $m_1^G$  and  $m_2^G$  from these inequalities, which can be summarized below:

- $m_2^G < \frac{1}{3}$   
(Multiplying 2 to each side of (d) and connecting it with (b), we get:  $1 + m_2^G > 2m_1^G > 4m_2^G$ . From this, we get:  $m_2^G < \frac{1}{3}$ .)
- $m_1^G < \frac{2}{3}$   
(From the fact that  $m_2^G < \frac{1}{3}$  and (b), we get:  $2m_1^G < 1 + m_2^G < \frac{4}{3}$ . From this, we get:  $m_1^G < \frac{2}{3}$ .)

From this, we are able to summarize  $Bob_G$ 's utilities for the sure consequences as follows:

- $U_{Bob_S}(\text{Glorified Life}) = 1$ ,
- $U_{Bob_S}(\text{Moderate Life}) = m_1^G < \frac{2}{3}$
- $U_{Bob_S}(\text{Mortified Life}) = m_2^G < \frac{1}{3}$
- $U_{Bob_S}(\text{Death}) = 0$

- $U_{Bob_G, Jill_G}(\text{Moderate Life}) = U_{Bob_S, Jill_S}(\text{Mortified Life}) = 1/2$
- $U_{Bob_S, Jill_S}(\text{Moderate Life}) = 3/4$
- $U_{Bob_S, Bob_G, Jill_S, Jill_G}(\text{Glorified Life}) = 1$

The four game-matrices that we have just seen above can now be summarized as follows:

Table 6.6: Summary of the Four Games Played by Each Type of Bob and Jill

$Bob_S \setminus Jill_S$	Peace	Attack	$Bob_S \setminus Jill_G$	Peace	Attack
Peace	<b>1, 1</b>	0, 3/4	Peace	4/3, 1/2	0, 1
Attack	3/4, 0	<b>1/2, 1/2</b>	Attack	1, 0	<b>1/2, 1/4</b>

$Bob_G \setminus Jill_S$	Peace	Attack	$Bob_G \setminus Jill_G$	Peace	Attack
Peace	1/2, 3/4	0, 1	Peace	1/2, 1/2	0, 1
Attack	1, 0	<b>1/4, 1/2</b>	Attack	1, 0	<b>1/4, 1/4</b>

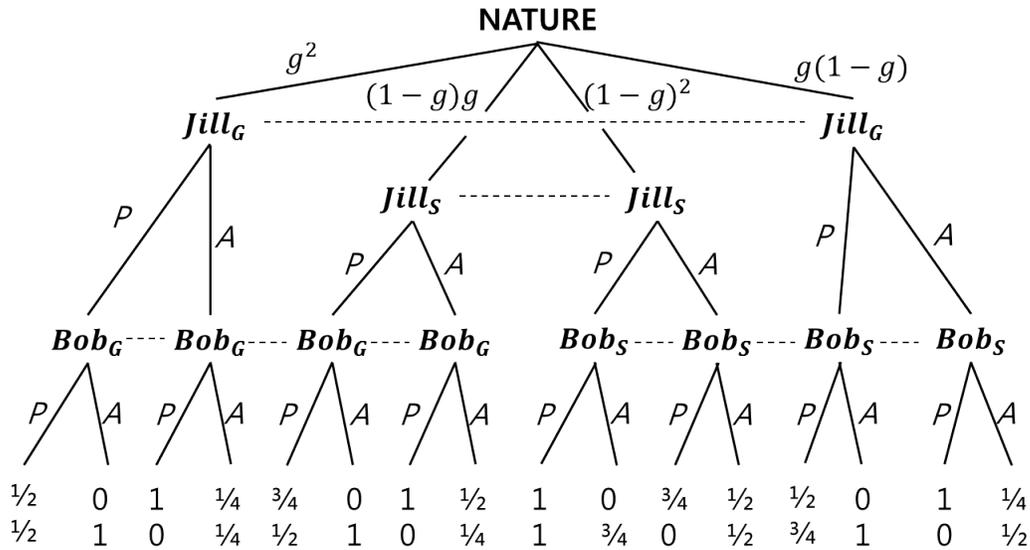
Remember that the numbers in the matrices are, now, *expected utilities*; this means that each type of Bob and Jill prefers option  $x$  to option  $y$  if and only if the expected utility of option  $x$  is greater than the expected utility of option  $y$  for each type of Bob and Jill in question. I now present the first Bayesian game-theoretic model for Hobbes's state of nature as follows:

---

In sum, the ranges that the utilities of Moderate Life and Mortified Life could possibly take for  $Bob_S$  and  $Bob_G$  would be:  $1/3 < U_{Bob_S}(\text{Mortified Life}) < 2/3 < U_{Bob_S}(\text{Moderate Life}) < 1$  and  $0 < U_{Bob_G}(\text{Mortified Life}) < 1/3 < U_{Bob_G}(\text{Moderate Life}) < 2/3$ . The same utility ranges will apply to both types of Jill as well.

For convenience, let us pick some numbers that fit within these ranges and assign them as the expected ( $VnM$ ) utilities of Mortified Life and Moderate Life for both types of Bob and Jill. (What follows in the main text of this paper are the values that I have assigned.) Note that this is not entirely arbitrary since the expected utilities that were assigned would have to fit within the range that we have just seen which we have established here.

Figure 6.1: The First Bayesian Game-Theoretic Model of Hobbes's State of Nature



- **Pay-offs:**  $Jill_i^j$
- $Bob_i^j$
- **P:** Seek Peace
- **A:** Initiate a Preemptive Attack
- **g:** The proportion of the entire population who are vain-glorious ( $0 < g < 1$ )

Let me briefly explain what the model is saying. At first, nature makes the first move and determines the proportion  $g$  of the entire population that is vain-glorious.  $g$  is a probability that is known to all types of players. One may think that people living in the state of nature know the value of  $g$  by their past experiences; that is, if one had encountered  $m$  vain-glorious persons among  $n$  people that one had interacted in the past, then one believes that  $g = m/n$ . Let's assume that the proportion  $g$  of the entire population that is vain-glorious is common knowledge.

The four branches that ramify from nature each correspond to the four possible states. The left-most branch corresponds to the state where both Jill and Bob are vain-glorious; the second branch corresponds to the state where Jill is modest and Bob is vain-glorious; the third branch corresponds to the state where both Jill and Bob are modest; and the fourth

branch corresponds to the state where Jill is vain-glorious and Bob is modest.

The nodes that are connected with a *dotted-line* denote a given *information set*; in which the player, given the information he/she has received, knows that he/she is located at one of the nodes in the information set, but does not completely know which particular node that he/she is located at. So, for example, the first information at the very top signifies that Jill knows that she is a vain-glorious type, but does not know whether she is dealing with a modest Bob or a vain-glorious Bob. In the bottom-left information set, Bob knows that he is a vain-glorious type, but does not know whether he is interacting with a modest Jill or a vain-glorious Jill, and also does not know what action Jill would decide to perform if she were a modest type.<sup>29</sup>

For each information set, we use the *Bayes's law* to determine the conditional probability that the player is located at a particular node *given* that he/she is in the information set. (This is why it is called *Bayesian game-theory*.) So, the probability that Jill is located at the upper leftest node (i.e. she is interacting with a vain-glorious Bob) given that she herself is vain-glorious (i.e. she is in the very first information set) would be:  $Pr(Bob_G | Jill_G) = \frac{Pr(Bob_G \cap Jill_G)}{Pr(Bob_G)} = \frac{g^2}{g^2 + g(1-g)} = g$ .

The pay-offs written at the bottom denote the *VnM* (expected) utilities of Jill (at the top) and Bill (at the bottom) for each “player-type + action” combinations. One can verify that the pay-offs correspond to the pay-offs written in the four game-matrices above.

Now, let's get into the interpretation of the model. We already know how the two vain-glorious types (i.e.  $Bob_G$  and  $Jill_G$ ) would act in the state of nature; they would initiate a preemptive attack regardless of the type and actions of their opponent. What we are curious about is how the modest types (i.e.  $Bob_S$  and  $Jill_S$ ) in the state of nature would act when they are uncertain about what type of person with whom they are interacting.

Remember that the modest types most prefer to mutually cooperate with another modest

---

<sup>29</sup>Note that Bob knows what action Jill would decide to perform if she were a vain-glorious type, (namely, initiate a preemptive attack), since initiating a preemptive attack strictly dominates seeking peace for a vain-glorious Jill.

type given that the other modest type cooperates as well. However, the modest types prefer to initiate a preemptive attack if *either* he/she is interacting with a vain-glorious type *or* he/she is interacting with a modest type who decides to attack.

Since the pay-offs denote each player's *VnM* (expected) utilities, each player prefers to perform a given act to another act if and only if the expected utility of the former act is greater than that of the latter act.

Suppose that one is the modest type of Jill (i.e.  $Jill_S$ .) For  $Jill_S$ , the expected utility of seeking peace (i.e. playing  $P$ ) would be: (the probability that Bob, who Jill is interacting with, is vain-glorious)  $\times$  (the utility of the outcome that is generated when  $Bob_G$  plays  $A$  while  $Jill_S$  plays  $P$ ) + (the probability that Bob, who Jill is interacting with, is modest)  $\times$  {(the probability that  $Bob_S$  plays  $P$ ) $\times$ (the utility of the outcome that is generated when  $Bob_S$  and  $Jill_S$  play  $P$ ) + (the probability that  $Bob_S$  plays  $A$ ) $\times$ (the utility of the outcome that is generated when  $Bob_S$  plays  $A$  and  $Jill_S$  plays  $P$ )}. Suppose that the probability of  $Bob_S$  playing  $P$  is  $pe^B$ . Calculating all of this, we get:

- $EU_{Jill_S}(P) = (1 - g) \cdot pe^B \dots$  (8.1)

Similarly, we can calculate  $Jill_S$ 's expected utility of initiating a preemptive attack (i.e. playing  $A$ ), which is:

- $EU_{Jill_S}(A) = \frac{1}{2} \cdot g + (1 - g) \left( \frac{3}{4} \cdot pe^B + \frac{1}{2} \cdot (1 - pe^B) \right) = \frac{1}{2} + \frac{1}{4} \cdot (1 - g) \cdot pe^B \dots$  (8.2)

The modest type of Jill,  $Jill_S$ , will initiate a preemptive attack if and only if the value of (8.2) greater than that of (8.1). With simple algebra, one can verify that this is so when:

- $pe^B \cdot (1 - g) < \frac{2}{3}$  (where  $0 < g < 1$  and  $0 \leq pe^B \leq 1$ )  $\dots$  (8.3)

Whenever the values of  $g$  and  $pe^B$  satisfy this inequality, the modest Jill will initiate a preemptive attack. Then, what does this inequality tell us? It tells us that it is not that hard for even a modest type of person, (such as  $Jill_S$ ), who would most prefer to cooperate and seek

peace with another modest type, to initiate a preemptive attack on the other party in the state of nature.

More specifically, the general intuition that is implied by the inequality is this: (a) The more  $Jill_S$  believes Bob to be vain-glorious (i.e. the larger the value of  $g$ ), the more is it likely for  $Jill_S$  to initiate a preemptive attack. (b) The less  $Jill_S$  believes Bob will choose to seek peace when he is modest (i.e. the lesser the value of  $pe$ ), the more likely it is for  $Jill_S$  to initiate a preemptive attack.

We can see that whether or not  $Jill_S$  (who is herself a modest type) would initiate a preemptive attack is determined by two parameters; the values of  $g$  and  $pe^B$ . However, we can see that when  $g$  is greater than  $1/3$  - that is, when the vain-glorious types constitute more than one third of the entire population, the modest Jill (i.e.  $Jill_S$ ) is *guaranteed* to initiate a preemptive attack; even when she initially believes that Bob will seek peace for sure given that he is modest. The same reasoning applies to the modest type of Bob (i.e.  $Bob_S$ ) as well.

Even when there are far fewer vain-glorious people in the entire population, the modest Jill will decide to initiate a preemptive attack if she believes that the probability that a modest Bob would seek peace is not that high; again, this is so, whenever the two values of  $g$  and  $pe^B$  satisfy the inequality (8.3) Again, exactly the same reasoning applies to the modest type of Bob (i.e.  $Bob_S$ ) as well.

What our first game-theoretic model shows is this. One does not need to assume that everybody in the state of nature is vain-glorious or egoistic (which is an assumption that is tacitly made by people who model Hobbes's state of nature as a PD game) in order to explain the universal conflict in Hobbes's state of nature. Even the modest type of people could very well initiate a preemptive attack in the state of nature, not because they do not prefer peace (in fact, they prefer peace more than anything else), but rather, because they are *uncertain* about whether they are interacting with another modest type like themselves or a vain-glorious type. The conclusion is that, (in very many cases), everybody (including both the modest types and the vain-glorious types) would decide to initiate a preemptive attack in

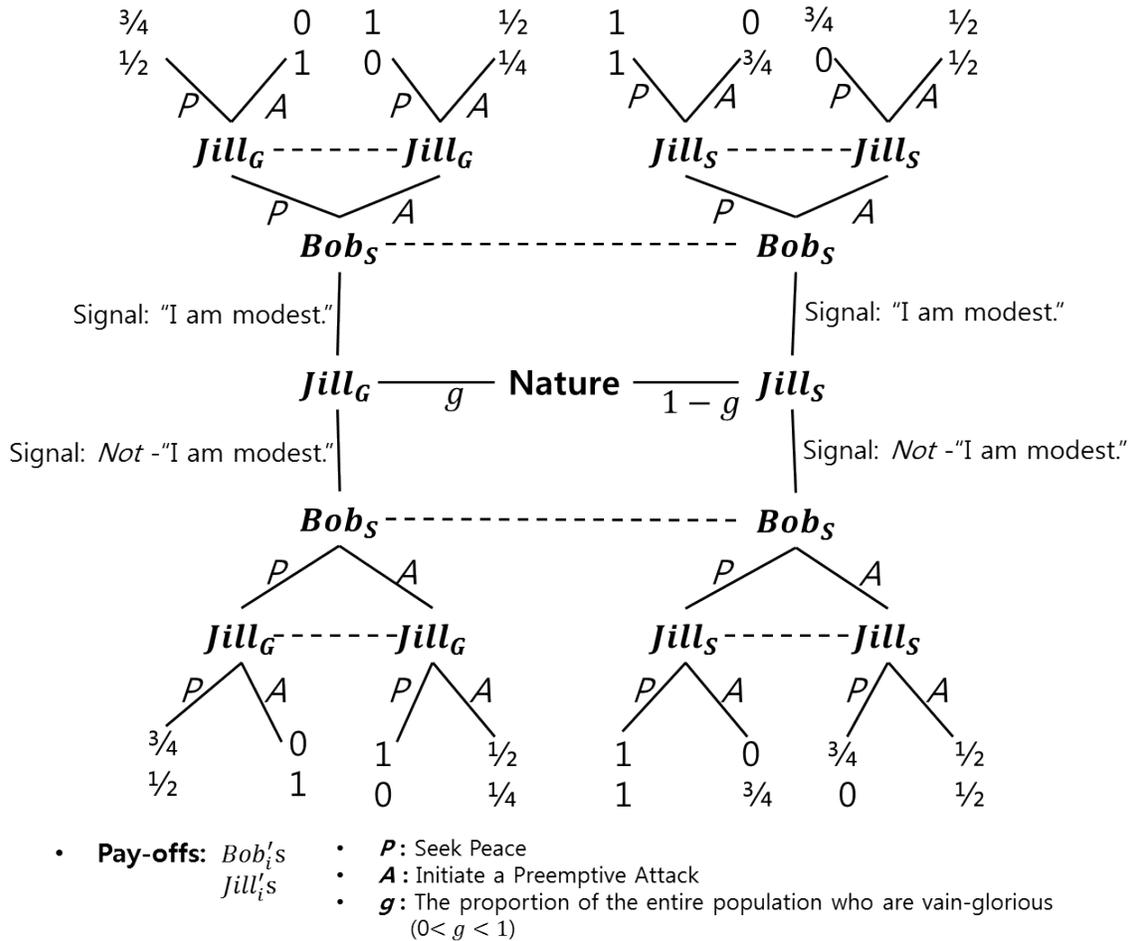
Hobbes's state of nature.

### **6.6.2 Second Bayesian Model of Hobbes's State of Nature: Is it Possible for One to Credibly Signal One's Type?**

In our first Bayesian game-theoretic model, we have seen that modest people living in Hobbes's state of nature can very well initiate a preemptive attack because they are uncertain about whether they are interacting with another modest person or a vain-glorious person. In other words, people in Hobbes's state of nature can initiate a preemptive attack because they are uncertain about the other person's *type*.

If this is so, then it would be quite handy if it is possible for the modest people to credibly *signal* their type to the other modest types so that they could coordinate their actions and seek peace with one another. The following variant of Bayesian games (usually known as *signaling-games*) models this type of situation.

Figure 6.2: The Second Bayesian Game-Theoretic Model of Hobbes's State of Nature



The model represents a situation where each type of Jill knows her own type and *also knows* that she is interacting with a modest Bob. However, the modest Bob does not know what type of Jill he is interacting with. Again, in such situations, the modest Jill would want to distinguish herself from the vain-glorious Jill and would want to *signal* that she is a modest type to the modest Bob; she might do so, by shouting, "I am modest!", to the modest Bob.

However, in order for the modest Bob to properly use this information to successfully coordinate his behaviors with the modest Jill, such signal must be *credible*. And such signal

is credible only when it is actually in vain-glorious Jill's own interest *not* to mimic modest Jill's signal and shout out, "I am also modest!", herself. When the vain-glorious Jill thinks that she would be better-off by copying the modest Jill's signal, she will do so, and, this would render the modest Jill's signal, namely, shouting out, "I am modest!", mere *empty words* that convey no information.

When the two different types of Jill can be properly distinguished by their different signals, the resulting equilibrium will be, what game theorists call, a "separating equilibrium." However, if the two different types of Jill cannot be properly distinguished by a credible signal because one of the types has an incentive to mimic the other type's signal, the resulting equilibrium will be, what game theorists call, a "pooling equilibrium."

I claim that there is no separating equilibrium in Hobbes's state of nature. Here is a very simple and informal way to think about it. Suppose, on the contrary, the modest Jill and the vain-glorious Jill can be properly distinguished by their different signals; that is, only the modest Jill sends the signal, "I am modest.", to the modest Bob. If so, then the modest Bob will know that he is interacting with a modest Jill whenever he receives the signal, "I am modest." from Jill, and will know that he is interacting with a vain-glorious Jill whenever he receive any other signal. So, after receiving the signal, "I am modest.", modest Bob will very likely seek peace, and after receiving any other signal, modest Bob will initiate a preemptive attack.

The catch to this is that the vain-glorious Jill will also know this quite too well. That is, the vain-glorious Jill will know that if she copies modest Jill's signal and shouts, "I am modest.", then she is very likely to face *Bob's* peace-seeking gesture, while if she doesn't, she will be facing *Bob's* preemptive attack for sure. In other words, by copying the modest Jill's signal, the vain-glorious Jill has a high chance to achieve Glorified Life, while she will achieve Mortified Life for sure if she doesn't. We can see that the vain-glorious Jill has every incentive to copy the modest Jill's signal. So, the vain-glorious Jill will also signal, "I am modest."

However, the modest Bob will know this very well also. So, after receiving the signal, “I am modest.” from Jill, the modest Bob will realize that such signal conveys no useful information. This means that the modest Bob’s beliefs about the probability that he will be facing a vain-glorious Jill will not be updated, and will, thereby, remain unchanged (i.e. it will remain as  $g$ ) after receiving the signal.

This, in effect, reduces our current signaling game into the right-half part of the first Bayesian game-theoretic model that we have seen in the previous subsection. The analysis is exactly the same as before; only this time, we will be reasoning in the shoes of the modest Bob instead of the modest Jill. The result is that the modest Bob will initiate a preemptive attack if and only if the values of  $g$  and  $pe^J$  (i.e. the probability that the modest Jill will seek peace) meet the following inequality:

- $pe^J \cdot (1 - g) < \frac{2}{3}$  (where  $0 < g < 1$  and  $0 \leq pe^J \leq 1$ )  $\dots$  (8.4)

Again, if the value of  $g$  is greater than  $1/3$  - that is, if the vain-glorious types constitute more than one third of the entire population in Hobbes’s state of nature - the modest Bob is guaranteed to initiate a preemptive attack. Even if the vain-glorious types constitute less than one third of the entire population, the modest Bob is still very likely to initiate a preemptive attack given that his beliefs about the probability that the modest Jill will seek peace is not too high. And given that the modest Bob initiates a preemptive attack, the modest Jill best-responds by initiating a preemptive attack also.

In short, our second Bayesian game-theoretic model shows another important reason why Hobbes’s state of nature is very likely to dissolve into a state of universal war even when the majority of the people prefer to seek mutual peace and cooperation. Not only are people in Hobbes’s state of nature uncertain about what type of person they are interacting with, but, in Hobbes’s state of nature, it is impossible for the modest types to successfully coordinate with other modest types by sending signals that would distinguish themselves from the vain-glorious types, because the vain-glorious types have a strong incentive to mimic the signals

of the modest types which would render any signal sent by the modest types mere empty words.

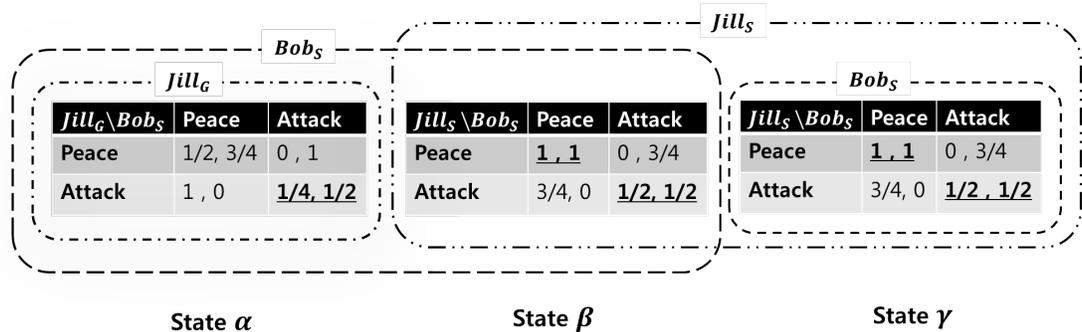
### **6.6.3 Third Bayesian Model of Hobbes's State of Nature: Uncertain About the Other Person's Beliefs**

Until now, we have seen that, in Hobbes's state of nature, even the modest types, who would most prefer to seek peace with other modest types, are quite likely to initiate a preemptive attack because they are uncertain about *what type of person* they are interacting with. That is, in Hobbes's state of nature, the modest types do not know whether they are interacting with another modest type like themselves or a vain-glorious type. Furthermore, we have seen that it is practically impossible for the modest types to credibly signal their identities to other modest types because the vain-glorious types have an incentive to mimic their signals and do exactly the same.

However, now, suppose that two modest types who encounter each other in the state of nature, somehow, got to know that each is interacting with another modest type. Would this guarantee that the two people will now succeed in coordinating their actions and achieve mutual peace? Not quite.

This is because although both modest people may know that he/she is interacting with another modest person, the person may still not know whether *the other person believes* that he/she is interacting with a modest person. In other words, conflict in the state of nature can still emerge because people are uncertain about the other party's *beliefs* even when they know the other party's *type*. The following Bayesian model demonstrates how people in the state of nature can fail to achieve mutual peace by being uncertain, not about the other person's type, but about the other person's beliefs.

Figure 6.3: The Third Bayesian Game-Theoretic Model of Hobbes's State of Nature



In this model, there are three possible states of the world;  $\alpha$ ,  $\beta$  and  $\gamma$ . Again, the dotted-lines indicate the information set in which each given player is located. In all three states,  $\alpha$ ,  $\beta$  and  $\gamma$ , Jill knows that she is interacting with a modest Bob,  $Bob_S$ . In state  $\gamma$ , the modest Bob,  $Bob_S$ , also knows that he is interacting with a modest Jill. However, in states  $\alpha$  and  $\beta$ , the modest Bob,  $Bob_S$ , does not know whether he is interacting with a modest Jill or a vain-glorious Jill.

A very interesting thing happens when the actual state is  $\gamma$ . Suppose that the actual state is  $\gamma$ . Then, both Bob and Jill know that they are interacting with another modest type. Furthermore, Bob also knows that Jill knows that she is interacting with another modest type. However, what  $Bob_S$  does not know is whether  $Jill_S$  knows that  $Bob_S$  knows that he is interacting with a modest Jill,  $Jill_S$ . This is because when the actual state is  $\gamma$ ,  $Bob_S$  knows that  $Jill_S$  knows that the true state is either  $\beta$  or  $\gamma$ . This means that there is a chance that  $Jill_S$  would think that the true state is  $\beta$  than  $\gamma$ . But, if the true state is  $\beta$ , then  $Jill_S$  knows that  $Bob_S$  knows that the true state is either  $\alpha$  or  $\beta$ . In other words, if the true state is  $\beta$ , then  $Jill_S$  knows that there is a chance that  $Bob_S$  would believe that he is interacting with a vain-glorious Jill rather than a modest Jill.

We can see, here, that Bob and Jill each has uncertainty in some part of his/her *higher-order beliefs*. And this is precisely how  $Bob_S$  and  $Jill_S$  may both decide to initiate a preemp-

tive attack even when they both know perfectly well that they are interacting with another modest person. That is, if the actual state is  $\gamma$ ,  $Bob_S$  might think that  $Jill_S$  thinks that  $Bob_S$  thinks that he is interacting with a vain-glorious Jill, which would make  $Jill_S$  think that  $Bob_S$  would initiate a preemptive attack, which, in turn, would make  $Jill_S$  initiate a preemptive attack, and, by knowing this,  $Bob_S$  would have no choice but to initiate a preemptive attack himself. This, again, results in universal conflict in Hobbes's state of nature.<sup>30</sup>

Our third Bayesian game-theoretic model shows us another very important reason why modest people that constitute the majority of the entire population can fail to coordinate their

---

<sup>30</sup>Here is the specific algebraic calculations. Suppose that one is  $Bob_S$  and is in the information set  $\alpha\beta$  -  $Bob_S$  knows that the true state is either  $\alpha$  or  $\beta$ , but he does not know whether the true state is  $\alpha$  or whether the true state is  $\beta$  (i.e.  $Bob_S$  does not know whether he is interacting with a vain-glorious Jill ( $\alpha$ ) or a modest Jill ( $\beta$ )).  $Bob_S$  knows that if the true state is  $\alpha$ ,  $Jill_G$  will initiate a preemptive attack for sure. Let the probability for the true state being  $\alpha$  given that  $Bob_S$  is in the information set  $\alpha\beta$  be  $a$ , and the probability that  $Jill_S$  would seek peace given that the true state is  $\beta$  be  $pe^J$ . Then, the expected utilities for each action, *Peace* and *Attack*, for  $Bob_S$  in state  $\alpha\beta$  becomes:

- $EU_{Bob_S}(Peace) = (1 - a) \cdot pe^J \dots$  (a)

- $EU_{Bob_S}(Attack) = \frac{1}{2} \cdot a + (1 - a) \left( \frac{3}{4} \cdot pe^J + \frac{1}{2} \cdot (1 - pe^J) \right) = \frac{1}{2} + \frac{1}{4} \cdot (1 - a) \cdot pe^J \dots$  (b)

$Bob_S$  initiates a preemptive attack if and only if (a) < (b). This is so, when,

- $pe^J \cdot (1 - a) < \frac{2}{3}$  (where  $0 < a < 1$  and  $0 \leq pe^J \leq 1$ )  $\dots$  (c)

Again, for all values of  $0 \leq pe^J \leq 1$ , inequality (c) is satisfied when  $\frac{1}{3} < a < 1$ . Suppose that this is so. That is, suppose that whenever  $Bob_S$  is in state  $\alpha\beta$  and is unsure about what type of Jill he will be facing, he believes that there is more than one third chance of meeting a vain-glorious Jill rather than a modest Jill. Let this fact be known to  $Jill_S$ .

Now, suppose that one is  $Jill_S$  who is in the information set,  $\beta\gamma$ . Exactly the same reasoning as before applies here as well. Let the probability that the true state is  $\beta$  given that  $Jill_S$  is in the information set  $\beta\gamma$  be  $b$ , and the probability that  $Bob_S$  would seek peace given that the true state is  $\gamma$  be  $pe^B$ . Note that if the true state is  $\beta$ , then  $Bob_S$  is guaranteed to attack by the assumption that we made in the previous paragraph. The expected utilities for each action, *Peace* and *Attack*, for  $Jill_S$  in state  $\beta\gamma$  becomes:

- $EU_{Jill_S}(Peace) = (1 - b) \cdot pe^B \dots$  (d)

- $EU_{Jill_S}(Attack) = \frac{1}{2} \cdot b + (1 - b) \left( \frac{3}{4} \cdot pe^B + \frac{1}{2} \cdot (1 - pe^B) \right) = \frac{1}{2} + \frac{1}{4} \cdot (1 - b) \cdot pe^B \dots$  (e)

$Jill_S$  initiates a preemptive attack if and only if (d) < (e). This is so, when,

- $pe^B \cdot (1 - b) < \frac{2}{3}$  (where  $0 < b < 1$  and  $0 \leq pe^B \leq 1$ )  $\dots$  (f)

Again, for all values of  $0 \leq pe^B \leq 1$ , inequality (f) is satisfied when  $\frac{1}{3} < b < 1$ .

Suppose, the common prior probabilities for each state,  $\alpha, \beta$ , and  $\gamma$ , are  $Pr(\alpha) = \frac{11}{70}$ ,  $Pr(\beta) = \frac{20}{70}$ ,  $Pr(\gamma) = \frac{39}{70}$ . Then, both inequalities (c), (f) are satisfied, and, in state  $\gamma$ , which is the true state, the two modest types of Bob and Jill will initiate a preemptive attack even when both perfectly know that each is interacting with another modest type. This is because, in state  $\gamma$ , Bob does not know whether Jill knows that Bob knows that Jill is modest; and Jill does not know whether Bob knows that Jill is modest.

behaviors to achieve universal peace in Hobbes's state of nature. Even when two modest people correctly identify that their counterpart is modest, they may still be uncertain about whether their counterpart knows that he/she is modest or whether his/her counterpart knows that he/she knows that his/her counterpart is modest, and so on. As our third Bayesian game-theoretic model shows, being uncertain about one's counterpart's beliefs (or one's higher-order beliefs) can very well cause each party to initiate a preemptive attack even when both parties perfectly know that his/her counterpart is modest.

## 6.7 Concluding Remarks

In the previous section, we have seen three Bayesian game-theoretic models that jointly illustrate how Hobbes's state of nature could dissolve into a state of war of all against all as Hobbes himself claims. We can see that the Bayesian game-theoretic models that I have provided nicely meet all of the four desiderata that I have introduced in section 8.3.

First, the models show that universal warfare is the equilibrium of the state of nature; of course, whether or not the modest types decide to initiate a preemptive attack depends on the proportion of vain-glorious people in the entire population as well as the probability that a given modest type would seek peace. However, we have seen that if the proportion of vain-glorious people exceeds a certain number (in our model,  $1/3$ ), the modest types are guaranteed to initiate a preemptive attack, which would lead to universal warfare.

Second, such universal warfare is *sub-optimal*. We can easily verify from our three Bayesian game-theoretic models that universal peace is a social state in which *everybody* (including the vain-glorious types) would prefer to rather be. This means that if there happens to be a powerful authority, such as a government, that has the power to enforce universal peace among the people living in the state of nature, then such a social institution would enhance the situation of everybody without worsening the situation of anybody.

However, many of the game-theoretic models that we have seen in the previous sec-

tions (including the PD game) all meet the first two desiderata. The real merit of our three Bayesian game-theoretic models comes from the fact that, unlike the other game-theoretic models that have tried to model Hobbes's state of nature, they meet the third and fourth desiderata as well.

As we can see, all three Bayesian game-theoretic models incorporate the existence of two distinct types of people (i.e. the modest type and the vain-glorious type) that Hobbes clearly assumes to dwell in the state of nature. This makes the three Bayesian game-theoretic models meet the third desideratum. Furthermore, all three Bayesian game-theoretic models directly model the relevant aspects of uncertainty into the model and show how uncertainty can be the primary cause of the universal conflict that emerge in Hobbes's state of nature. This makes the three Bayesian game-theoretic models meet the fourth desideratum.

The fact that our three Bayesian game-theoretic models meet all four desiderata of Hobbes's state of nature makes them have several advantages that the previous game-theoretic models that tried to model Hobbes's state of nature did not have.

First of all, the three Bayesian game-theoretic models better accord with what Hobbes actually states in his text. As we have already seen, Hobbes clearly did not think that the universal conflict which he thought the state of nature will inevitably dissolve into is primarily caused by everybody being strictly egoistic and having the preference structure of the players in the PD game. It is true that Hobbes thought that there *are* such egoistic and power-seeking people in the state of nature. However, the main reason why Hobbes thought that universal conflict would inevitably emerge in the state of nature is not because he thought that everybody is egoistic and glory-hungry in this way, but rather, because he thought that the modest people cannot properly distinguish themselves from the glory-hungry ones, and are unsure about the beliefs of other modest people. Our three Bayesian game-theoretic models, unlike previous game-theoretic models, accommodate Hobbes's original intentions very nicely.

Second, the fact that universal conflict in the state of nature is due to uncertainty rather than people having a strictly egoistic psychology has the advantage of freeing Hobbes from

*psychological egoism*. As I have argued in one of the previous chapters, psychological egoism (which roughly claims that all human actions are ultimately motivated by strictly egoistic concerns) is a very contestable doctrine of human psychology which many people think to be false. Remember that the primary role that the state of nature plays in Hobbes's political philosophy is to justify the existence of governments by illustrating the misery that people would face when they did not have one. However, if the universal conflict as well as the misery that accrues it can only be explained by assuming a theory of human psychology which many people think to be implausible, Hobbes's justification for the existence of governments would, to that very extent, be weakened.

Our three Bayesian game-theoretic models have the advantage of explaining the universal conflict of the state of nature without assuming that everybody has a strictly egoistic psychology. It shows that even when the majority of the people favor mutual peace and cooperation, universal conflict can still emerge primarily because of *uncertainty*. Such explanation is free of any contestable psychological assumptions, and is, thereby, more plausible and widely applicable to many actual human situations. This means that our three Bayesian game-theoretic models provide a much firmer foundation that Hobbes's political philosophy can stably rest on, and, thereby, provide a more plausible justification for the existence of governments.

I would like to say one more thing before I conclude this chapter. In section 6.5, I have explained that one of the reasons why many people are attracted to the idea of modeling Hobbes's state of nature as an infinitely repeated PD game comes from the fact that such model seems to explain how people can *escape* the state of nature. The reason why I did not include such attraction as one of the desiderata is because it was never part of Hobbes's intentions to provide an accurate historical account of how all of our particular governments have originated. We can see, in the following passage, Hobbes's rather apologetic attempt to provide real-life examples that roughly fit into his description of the state of nature as a way of responding to people who question the state of nature's historical authenticity:

It may peradventure be thought, there was never such a time nor condition of war as this; and I believe it was never generally so, over all the world. But there are many places where they live so now. For the savage people in many places of *America* ... have no government at all. ... Howsoever, it may be perceived what manner of life there would be where there were no common power to fear, by the manner of life which men that have formerly lived under a peaceful government use to degenerate into, in a civil war. But though there had never been any time wherein particular men were in a condition of war one against another ...  
[Hobbes [1997, *Leviathan*, Chapter XIII, Section 10, 11]]

Hobbes explains that at least some parts of the world, such as America, are currently in a state of nature, and that even if it is true that not all governments have originated from an actual state of nature, we can, at least, speculate how miserable life would be in a situation where there were no governments by observing what generally happens during civil wars. We can see that what Hobbes is emphasizing here is *the misery* of the state of nature rather than its *historical authenticity*. In other words, what Hobbes is virtually saying here is, “Even if my description of the state of nature is, in many cases, historically false, the fact that our lives will be miserable without a government is absolutely true!” And, as a matter of fact, this is all that is needed for Hobbes’s purpose.

We should remember that the sole purpose of Hobbes’s state of nature is to justify the existence of governments. And between the misery and the historical authenticity of Hobbes’s state of nature, it can be easily shown that the latter really plays no major role. That is, for Hobbes, the fact that a particular government did not go through a phase of state of nature, and was, thereby, not established by the mutual agreements made by its people, does not undermine its legitimacy.

In *Leviathan*, Hobbes distinguishes between two different ways that a commonwealth can

be established: (1) by *institution*, and (2) by *acquisition*.<sup>31</sup> A commonwealth is established by institution when people spontaneously form a government and transfer their rights to the sovereign by mutual agreement. A commonwealth is established by acquisition when people become part of an already existing commonwealth either by succession (which Hobbes calls “Paternal”<sup>32</sup>) or by physical force (which Hobbes calls “Despotical.”<sup>33</sup>) Between the two cases, only governments established in the first way (i.e. by institution) go through a phase of state of nature; that is, a phase of state of nature is absent in governments that were established in the second way (i.e. acquisition.) Nonetheless, Hobbes claims that the rights of the sovereign in both cases - that is, in cases of commonwealths by institution and in cases of commonwealths by acquisition - are *the same*.<sup>34</sup>

Of course, Hobbes makes it clear that, in both types of commonwealths, the right of the sovereign derives from the *covenants* made by its subjects.<sup>35</sup> However, it is quite clear that Hobbes is not confining the term “covenant” to mean only *actual expressed consent*. As it can be affirmed by the fact that Hobbes likens the rights of a sovereign over its subjects in a commonwealth established by natural succession (which is one form of commonwealth by acquisition) to the rights that parents have over their children<sup>36</sup>, the term “covenant”, within Hobbes’s moral and political philosophy, is used broadly to encompass the type of acts that may be called “*tacit consent*” or “*hypothetical consent*.”

This means that, for Hobbes, whether or not a given commonwealth has actually gone through a phase of state of nature and was established by the actual expressed consent of its people is immaterial for its justification. What is sufficient is that, *counterfactually speaking*, people’s lives *would have been* much worse if the commonwealth did not exist. And, I repeat, this is all that is needed for Hobbes’s general purpose. In this sense, it might be

---

<sup>31</sup> See Hobbes [1994, *Leviathan*, Chapter XVII, Section 15]

<sup>32</sup> Hobbes [1994, *Leviathan*, Chapter XX, Section 4]

<sup>33</sup> Hobbes [1994, *Leviathan*, Chapter XX, Section 10]

<sup>34</sup> Hobbes [1994, *Leviathan*, Chapter XX, Section 14]

<sup>35</sup> Hobbes [1994, *Leviathan*, Chapter XVIII, Section 1 and Chapter XX, Section 11 ]

<sup>36</sup> Hobbes [1994, *Leviathan*, Chapter XX, Section 4]

plausible to think of Hobbes's state of nature as a *hypothetical theoretical device* that occupies a similar place in Hobbes's political philosophy as what *the original position* occupies in Rawls' political philosophy.<sup>37</sup>

Or course, I am not denying that thinking of how people can actually escape the state of nature (if there were such a phase), and how they can successfully establish a government by mutual agreement, is, in itself, an interesting question to ask. However, it is not a question that Hobbes himself directly pursues, nor does an answer to this question required for Hobbes's general purpose of justifying the existence of governments. This means that there is really no theoretical need for a game-theoretical model that is designed to represent Hobbes's state of nature to show how people in the state of nature can *escape* it and successfully establish a government by mutual agreement. For the purpose of justifying the existence of governments, it suffices for the game-theoretic model to show that the state of nature results in a sub-optimal equilibrium.

However, for those who still think that showing how the people in the state of nature could possibly escape the state of nature is a major attraction of any game-theoretic model that represents it, it is not hard to see that our three Bayesian game-theoretic models can satisfy these people as well. As we have seen, whether or not a given modest type decides to initiate a preemptive attack in the stat of nature depends on two factors; (a) the proportion of vain-glorious people in the entire population, and (b) the probability that a given modest type is believed to seek peace.

When the proportion of vain-glorious people in the entire population is significantly low or when the modest types start to believe that there is a high chance that other modest type will seek peace, a given modest type will decide to seek peace as well. So, there are two basic ways to escape Hobbes's state of nature: reduce the number of vain-glorious people in the entire population, or encourage modest people to seek peace more often. How this may be achieved is left for the reader's own speculation.

---

<sup>37</sup>See Rawls [1971, 1999, A Theory of Justice].

I would like to end this paper by briefly saying something about the particular form of government that Hobbes thinks that his argument from the state of nature justifies. Hobbes thought that his argument supported a government with nearly absolute power - a government in which the liberty of everybody except the sovereign are severely restricted.<sup>38</sup> I believe that this is a major logical slip into Hobbes's, otherwise quite cogent, argument had fallen; I believe that such logical slip can be attributed to many external factors (such as the general political atmosphere) that Hobbes was experiencing in his own time. If we agree that giving governments excessive power is nothing that is desirable in itself, and that the sole purpose of governments is to prevent unwanted conflict and restore optimality, we must agree that the basic form of government that Hobbes's argument logically justify is a government that has much more limited powers than what Hobbes himself had suggested.

---

<sup>38</sup>See Hobbes [1994, Leviathan, Chapter XVIII]

# Bibliography

F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1), 1963.

Kenneth Arrow. *Social Choice & Individual Values*. Yale University Press, 1951.

Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1980.

Kurt Baier. *The Rationality and the Moral Order*. Open Court, 1995.

S. Barberà, P. Hammond, and C. Seidl, editors. *Handbook of Utility Theory-Volume I Principles*. Kluwer Academic Publishers, 1998.

S. Barberà, P. Hammond, and C. Seidl, editors. *Handbook of Utility Theory-Volume II Extensions*. Kluwer Academic Publishers, 2003.

Brian Barry. *Political Argument*. London: Routledge & Kegan Paul, 1965.

D. Bauchner and P Kelly, editors. *The Social Contract From Hobbes to Rawls*. Routledge, 1994.

Daniel Bernoulli. Exposition of a new theory on the measurement of risk. *Econometrica*, 22 (1), 1954.

Cristina Bicchieri. 10. rationality and game theory. In *The Oxford Handbook of Rationality*. Oxford University Press, 2004.

- Ken Binmore. *The Foundations of Analysis: A Straightforward Introduction-Book I Logic, Sets and Numbers*. Cambridge University Press, 1980.
- Ken Binmore. *The Foundations of Analysis: A Straightforward Introduction-Book II Topological Ideas*. Cambridge University Press, 1981.
- Ken Binmore. *Mathematical Analysis: A Straightforward Introduction (second edition)*. Cambridge University Press, 1982.
- Ken Binmore. *Game Theory and the Social Contract Volume I: Playing Fair*. MIT Press, 1994.
- Ken Binmore. *Game Theory and the Social Contract Volume II: Just Playing*. MIT Press, 1998.
- Ken Binmore. *Rational Decisions*. Princeton University Press, 2009.
- L. Blume. Lecture notes on "ordinal representations".
- L. Blume, A. Brandenburger, and E. Dekel. An overview of lexicographic choice under uncertainty. *Annals of Operations Research*, 19, 1989.
- Richard Brandt. The definition of an 'ideal observer' theory in ethics. *Philosophy and Phenomenological Research*, 15(3), 1955.
- Richard Brandt. *A Theory of the Good and the Right (Revised Edition)*. Prometheus Books, 1998.
- Michael E. Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press, 1987.
- C. D. Broad. Egoism as a theory of human motives. *Hibbert Journal*, LXXXIV(8), 1950.
- John Broome. *Weighing Goods*. Blackwell Publishing, 1991.

- John Broome. *Ethics out of Economics*. Cambridge University Press, 1999.
- John Broome. *Weighing Lives*. Oxford University Press, 2004.
- Joseph Butler. *Five Sermons (edited by Stephen Darwall)*. Hackett, 1983.
- John M. Cooper, editor. *Plato - Complete Works*. Hackett, 1997.
- Russell Cooper, Douglas Dejong, Robert Forsythe, and Thomas W. Ross. Cooperation without reputation: Experimental evidence from prisoner's dilemma games. *Games and Economic Behavior*, 12(13), 1996.
- Peter A. Danielson, editor. *Modeling Rationality, Morality, and Evolution*. Oxford University Press, 1998.
- Stephen Darwall. Normativity and projection in hobbes leviathan. *Philosophical Review*, 109(3), 2000.
- Stephen Darwall, Allan Gibbard, and Peter Railton, editors. *Moral Discourse and Practice: Some Philosophical Approaches*. Oxford University Press, 1996.
- Robyn M. Dawes and Richard H. Thaler. Anomalies: Cooperation. *The Journal of Economic Perspectives*, 2(3), 1988.
- A. Dixit and S. Skeath. *Games of Strategy (second edition)*. W. W. Norton & Company, 2004.
- Prajit Dutta. *Strategies and Games: Theory and Practice*. The MIT Press, 1999.
- Daniel Ellsberg. Classic and current notions of 'measurable utility'. *The Economic Journal*, 64(255), 1954.
- J. Elster and J. Roemer, editors. *Interpersonal Comparisons of Well-Being*. Cambridge University Press, 1991.

- Jon Elster. *Sour Grapes - Studies in the Subversion of Rationality*. Cambridge University Press, 1983.
- Jon Elster. *Ulysses and the Sirens - Studies in Rationality and Irrationality*. Cambridge University Press, 1984.
- Roderick Firth. Absolutism and the ideal observer. *Philosophy and Phenomenological Research*, 12(3), 1952.
- Roderick Firth. Reply to professor brandt. *Philosophy and Phenomenological Research*, 15(3), 1955.
- Peter C. Fishburn. Independence in utility theory with whole product sets. *Operations Research*, 13(1), 1965.
- Peter C. Fishburn. Utility theory. *Management Science*, 14(5), 1968.
- Peter C Fishburn. A general theory of subjective probabilities and expected utilities. *Management Science*, 14(5), 1969.
- Peter C Fishburn. *Utility Theory for Decision Making*. Wiley, 1970.
- Peter C. Fishburn. A study of lexicographic expected utility. *Management Science*, 17(11), 1971a.
- Peter C. Fishburn. Lexicographic orders, utilities and decision rules: A survey. *Management Science*, 20(11), 1971b.
- Peter C. Fishburn. The axioms of subjective probability. *Statistical Science*, 1(3), 1986.
- M. Fleurbaey, M. Salles, and J. Weymark, editors. *Justice, Political Liberalism, and Utilitarianism - themes for Harsanyi and Rawls*. Cambridge University Press, 2008.
- Wulf Gaertner. *A Primer in Social Choice Theory*. Oxford University Press, 2006.

- David Gauthier. *The Logic of Leviathan*. Oxford University Press, 1969.
- David Gauthier. Rational cooperation. *Noûs*, 8(1), 1974.
- David Gauthier. Thomas hobbes: Moral theorist. *The Journal of Philosophy*, 76(10), 1979a.
- David Gauthier. David hume, contractarian. *Philosophical Review*, 88(1), 1979b.
- David Gauthier. *Morals by Agreement*. Oxford University Press, 1984.
- David Gauthier. Review: Taming leviathan. *Philosophy and Public Affairs*, 16(3), 1987.
- Bernard Gert. Hobbes and psychological egoism. *Journal of the History of Ideas*, 28(4), 1967.
- Bernard Gert. Hobbess account of reason. *The Journal of Philosophy*, 76(10), 1979.
- Bernard Gert. Introduction. In *Man and Citizen (De Homine and De Cive)* (edited by Bernard Gert). Hackett, 1991.
- Allan Gibbard. *Wise Choices, Apt Feelings*. Harvard University Press, 1990.
- Alvin I. Goldman. *A Theory of Human Action*. Prentice-Hall, Inc., 1970.
- Ian Hacking. *The Emergence of Probability*. Cambridge University Press, 1975, 2006.
- F. Hahn and Martin Hollis, editors. *Philosophy and Economic Theory*. Oxford University Press, 1979.
- Ishtiyaque Haji. Hampton on hobbes on state of nature cooperation. *Philosophy and Phenomenological Research*, 51(3), 1991.
- Jean Hampton. *Hobbes and The Social Contract Tradition*. Cambridge University Press, 1986.

- Jean Hampton. Cooperating and contracting a reply to i. haji. *Philosophy and Phenomenological Research*, 51(3), 1991.
- Jean Hampton. Hobbes and ethical naturalism. *Philosophical Perspectives*, 6. Ethics, 1992.
- Russell Hardin. *Collective Action*. Johns Hopkins University Press, 1982.
- John C. Harsanyi. Ethics in terms of hypothetical imperatives. *Mind, New Series*, 67(267), 1958.
- John C. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii. part i. the basic model. *Management Science*, 14(3):159–182, November 1967.
- John C. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii. part iii. the basic probability distribution of the game. *Management Science*, 14(7):486–502, March 1968a.
- John C. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii. part ii. bayesian equilibrium points. *Management Science*, 14(5):320–334, January 1968b.
- John C. Harsanyi. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, 1977.
- John C. Harsanyi. Bayesian decision theory and utilitarian ethics. *Economics and Ethics (American Economic Association)*, 68(2), 1978.
- John C. Harsanyi. Does reason tell us what moral code to follow and, indeed, to follow any moral code at all? *Ethics*, 96(1), 1985.
- D. Hausman and M. McPherson. *Economic Analysis, Moral Philosophy, and Public Policy (second edition)*. Cambridge University Press, 2006.
- Thomas Hobbes. *Man and Citizen (De Homine and De Cive)*. Hackett, 1991.

Thomas Hobbes. *Leviathan (with selected variants from the Latin edition of 1668)*. Hackett, 1994.

Thomas Hobbes. *On The Citizen*. Cambridge University Press, 1997.

Thomas Hobbes. *The Elements of Law, Natural and Politic*. Dodo Press, 2009.

David Hume. *An Enquiry Concerning the Principles of Morals*. Oxford: Clarendon Press, 1975.

David Hume. *A Treatise of Human Nature*. Penguin, 1984.

Richard C. Jeffrey. On interpersonal utility theory. *The Journal of Philosophy*, LXVIII(20), 1971.

Richard. C. Jeffrey. *The Logic of Decision*. University of Chicago Press, 1983.

G. Jehle and P Reny. *Advanced Microeconomic Theory (second edition)*. Addison Wesley, 2000.

Niels Erik Jensen. An introduction to bernoulliam utility theory: 1 utility functions. *The Swedish Journal of Economics*, 69(3), 1967a.

Niels Erik Jensen. An introduction to bernoulliam utility theory ii - interpretations, evaluation and application; a critical survey. *The Swedish Journal of Economics*, 69(4), 1967b.

James Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 1979.

Gregory Kavka. Hobbess war of all against all. *Ethics*, 93(2), 1983.

Gregory Kavka. *Hobbesian Moral and Political Theory*. Princeton University Press, 1986.

- Gregory Kavka. The rationality of rule-following: Hobbes dispute with the foole. *Law and Philosophy*, 14(1), 1995.
- Isaac Kramnick. Writing politics. In Jonathan Monroe, editor, *Writing and Revising the Disciplines*. Cornell University Press, 2002.
- David H. Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of Measurement*. Dover, 1971.
- Jody S. Kraus. *The Limits of Hobbesian Contractarianism*. Cambridge University Press, 1993.
- David M. Kreps. *Notes on the Theory of Choice*. Westview Press, 1988.
- David Lewis. *Convention*. Blackwell Publishing, 2002.
- R. Duncan Luce and Howard Raiffa. *Games and Decisions Introduction and Critical Survey*. Dover, 1957.
- Alfred F. MacKay. Extended sympathy and interpersonal utility comparisons. *The Journal of Philosophy*, 83(6), 1986.
- John Mackie. *Ethics - inventing right and wrong*. Penguin, 1977.
- A. Mas-Colell, M. Whinston, and J. Green. *Microeconomic Theory*. Oxford University Press, 1995.
- F. S. McNeilly. Egoism in hobbes. *The Philosophical Quarterly*, 16(64), 1966.
- F. S. McNeilly. *The Anatomy Of Leviathan*. Macmillan, 1968.
- A. Mele and P Rawling, editors. *The Oxford Handbook of Rationality*. Oxford University Press, 2004.
- John Stuart Mill and Jeremy Bentham. *Utilitarianism and Other Essays*. Penguin, 1987.

- C. Morris and A. Ripstein, editors. *Practical Rationality and Preference*. Cambridge University Press, 2001.
- Paul Moser, editor. *Rationality in Action - Contemporary Approaches*. Cambridge University Press, 1990.
- Roger Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1997.
- Patrick Neal. Hobbes and rational choice theory. *The Western Political Quarterly*, 41(4), 1988.
- Robert Nozick. *Anarchy, State and Utopia*. Basic Books, 1974.
- Martin Osborne. *An Introduction to Game Theory*. Oxford University Press, 2003.
- Derek Parfit. *Reasons and Persons*. Oxford University Press, 1984.
- Mark P. Petracca. The rational choice approach to politics: A challenge to democratic theory. *The Review of Politics*, 53(2), 1991.
- Philip Pettit. *Rules, Reasons, and Norms*. Oxford, 2002.
- Peter Railton. Moral realism. *The Philosophical Review*, XCV(2), 1986a.
- Peter Railton. Facts and value. *Philosophical Topics*, XIV(2), 1986b.
- Peter Railton. *Facts, Values, and Norms: Essays Toward a Morality of Consequence*. Cambridge University Press, 2003.
- John Rawls. *A Theory of Justice (revised edition)*. Harvard University Press, 1971, 1999.
- John Rawls. *Political Liberalism (second edition)*. Columbia University Press, 1993, 2005.
- Tom Regan. The case for animal rights. In David Boonin and Graham Oddie, editors, *What's Wrong? – Applied Ethicists and Their Critics (Second Edition)*. Oxford University Press, 2010.

- Michael Resnik. *Choices - An Introduction to Decision Theory*. University of Minnesota Press, 1987.
- Fred S. Roberts. *Measurement Theory*. Addison-Wesley Publishing Company, 1979.
- John Roemer. *Theories of Distributive Justice*. Harvard University Press, 1998.
- Leonard Savage. *The Foundations of Statistics*. Dover, 1972.
- T. M. Scanlon. *What We Owe to Each Other*. Belknap Press of Harvard University Press, 1998.
- Thomas M. Scanlon. Preference and urgency. *The Journal of Philosophy*, Vol. 72(19), 1975.
- Thomas Schelling. *Micromotives and Macrobehavior*. W.W. Norton & Company, 1978.
- Thomas Schelling. *The Strategy of Conflict*. Harvard University Press, 1981.
- A. Sen and B. Williams, editors. *Utilitarianism and Beyond*. Cambridge University Press, 1982.
- Amartya Sen. *Collective Choice and Social Welfare*. North-Holland Publishing Company, 1970.
- Amartya Sen. Plural utility. *Meeting of the Aristotelian Society*, 1981.
- Amartya Sen. *Choice, Welfare and Measurement*. Harvard University Press, 1982.
- Brian Skyrms. *Evolution of The Social Contract*. Cambridge University Press, 1996.
- Brian Skyrms. *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, 2004.
- Michael Slote. An empirical basis for psychological egoism. *The Journal of Philosophy*, 61 (18):530–537, 1964.

- J.J.C. Smart and Bernard Williams. *Utilitarianism For & Against*. Cambridge University Press, 1973.
- Michael Smith, David Lewis, and Mark Johnston. Dispositional theories of value. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 63, 1989.
- Tom Sorell and Luc Foisneau, editors. *Leviathan After 350 Years*. Oxford University Press, 2004.
- Leo Strauss. *The Political Philosophy of Hobbes: Its Basis and Its Genesis*. University of Chicago Press, 1996.
- Nicholas Sturgeon. Hume on reason and passion.
- Robert Sugden. Rational choice: A survey of contributions from economics and philosophy. *The Economic Journal*, 101(407), 1991.
- A. E. Taylor. The ethical doctrine of hob. *Philosophy*, 13(52), 1938.
- Michael Taylor. *Anarchy and Cooperation*. London: Wiley, 1976.
- Tom R. Tyler. *Why People Obey the Law*. Princeton University Press, 2006.
- Peter Vallentyne. *Contractarianism and Rational Choice*. Cambridge University Press, 1991.
- Hal Varian. *Microeconomic Analysis (Third edition)*. W. W. Norton & Company, 1992.
- Hal Varian. *Intermediate Microeconomics (seventh edition)*. W. W. Norton & Company, 2006.
- John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Howard Warrender. *The Political Philosophy of Hobbes: His Theory of Obligation*. Oxford University Press, 2000.

John A. Weymark. Measurement theory and the foundations of utilitarianism. *Social Choice and Welfare*, Vol. 25, 2005.

Bernard Williams. Internal and external reasons. In *Moral Discourse & Practice* (edited by Darwall, Gibbard, and Railton) (1997). Oxford University Press, 1981.