A Study on Hume’s Theory of Justice: Evolution, Utilitarianism, and the Social Contract

Dissertation zur Erlangung des akademischen Grades eines Doktors der Philosophie

vorgelegt von
Fior Mota de Andrade, Pedro
an der

Universität Konstanz
Geisteswissenschaftliche Sektion
Fachbereich Philosophie

Tag der mündlichen Prüfung: 10.07.2020
Referent: Prof. Dr. Peter Stemmer
Referent: Prof. Dr. Jacob Rosenthal
Konstanz, 2020
Summary

Introduction .................................................................................................................. 1

Chapter 1 - Hume’s Moral Theory and the Virtue of Justice ......................... 7
  1.1 Some Preliminary Remarks ........................................................................... 7
  1.2 Virtue, Vice, and the General Point of View .............................................. 9
    1.2.1 Definition of Virtue and Vice ................................................................. 9
    1.2.2 Actions, Mental Qualities, and Character Traits ............................... 11
    1.2.3 The General Point of View ................................................................. 14
    1.2.4 Moral Sentiments ................................................................................... 20
  1.3 Sources of Moral Approbation and Disapprobation ............................. 23
  1.4 Sympathy and Benevolence ........................................................................ 34
  1.5 The Role of the Virtue of Justice .............................................................. 42
  1.6 Three Lines of Interpretation ..................................................................... 57

Chapter 2 - The Background Conditions for the Convention of Justice ....... 62
  2.1 The Background Conditions for the Convention of Justice ................. 62
    2.1.1 Instrumental Rationality ........................................................................ 62
    2.1.2 Artificiality .............................................................................................. 69
    2.1.3 Circumstances of Justice ....................................................................... 75
  2.2 Analysis of the Pre-Conventional Structure of Interaction .................... 88
    2.2.1 The Humean State of Nature ................................................................. 88
    2.2.2 The Stag Hunt ....................................................................................... 96
    2.2.3 Solutions for the Stag Hunt ................................................................. 102
Chapter 3 - Hume’s Convention of Justice

3.1 The Motive of Justice

3.2 The Convention of Justice

3.2.1 Two Definitions of the Convention of Justice

3.2.2 The Rules of the Convention of Justice

3.3 Hume’s Convention of Justice, Evolution, Utilitarianism, and the Social Contract

3.3.1 The Evolutionary Interpretation

3.3.2 The Rule-Utilitarian Interpretation

3.3.3 The Contractarian Interpretation

3.3.3.1 Original, Explicit, and Tacit Contractarianism

3.3.3.2 Hypothetical Contractarianism

3.3.3.3 The General Scheme of Justice

Chapter 4 - The Rules of Private Property

4.1 The Initial Distribution

4.1.1 Immediate or Present Possession

4.1.2 Original Acquisition

4.1.3 Equality

4.1.4 Desert

4.1.5 Natural Distribution
4.1.6 Utility and Imagination ..............................................189
4.1.7 The Initial Distribution, Evolution, Utilitarianism, and the Social Contract .........................................................192
4.2 The Transference of Property by Consent ............................196
4.3 Utility and Inequality ........................................................198
4.3.1 The Standard Interpretation ............................................198
4.3.2 Social Inequalities and Negative Effects .............................199
4.4 The Redistributive Rule .....................................................205
4.4.1 Hume’s Theory of Property ...........................................205
4.4.2 The Natural Virtue of Equity ...........................................206
4.4.3 Indirect Utility and Rule-Utilitarianism ...............................208
4.4.4 Redistribution and Contractarianism .................................212

Chapter 5 - The Rule of the Obligation of Promises .................221
5.1 Promises and the Natural Law Theory ..................................221
5.2 Hume’s Critique of the Traditional Theories ............................226
5.2.1 The Intelligibility and the Moral Obligation of Promise ........226
5.2.2 Two Objections ..........................................................234
5.3 The Logic of Promises: A Game-Theoretic Analysis ..............238
5.3.1 Coordination Promises and Exchange Promises ...............238
5.3.2 The Farmer’s Dilemma .................................................243
5.3.3 The Infinitely Repeated Farmer’s Dilemma ......................246
5.4 Promises, Evolution, Utilitarianism, and the Social Contract ....253

Chapter 6 - Naturalism and Hume’s Theory of Justice ............265
6.1 Hume’s Naturalism .......................................................265
6.2 The Descriptive and the Normative in Hume .......................270
6.3 The Public and the Private Utility in Hume..............................275
6.4 Degrees of Plausibility and the Interpretation of Hume’s Theory of
Justice........................................................................................................279
6.5 A Final Remark.....................................................................................289

Conclusion..................................................................................................295

References..................................................................................................300
Abstract

This study aims to provide the correct interpretation of Hume’s theory of justice. The study on the nature of justice is of the utmost importance for comprehending Hume’s moral philosophy, for I argue that the virtue of justice is the central concept of his moral theory. In this sense, any plausible interpretation of this moral theory must take into account the nature of the virtue of justice as a priority in Hume. In the tradition, Hume scholars have come up with three main lines of interpretation of Hume’s theory of justice: i) the evolutionary interpretation, ii) the rule-utilitarian interpretation, and iii) the contractarian interpretation. Each of these lines of interpretation emphasizes a particular perspective adopted by Hume to explain the virtue of justice in his theory. And each of the three lines of interpretation makes some plausible points and relevant contributions to a better comprehension of the different parts of Hume’s theory of justice. To settle the interpretative dispute between the three lines of interpretation of Hume’s theory of justice, I assess them by their degrees of plausibility in face of the textual evidence in Hume’s work. In this way, I attempt to develop a correct interpretation of Hume’s theory of justice from the most plausible line of interpretation ascertained by this study.

Keywords: Hume, Theory of Justice, Evolution, Utilitarianism, Contractarianism.

Acknowledgments

I would like to express my sincere thanks to all those who helped me to conclude this doctoral dissertation.

First of all, I would like to thank my professor supervisor, Herr Prof. Peter Stemmer, for all the care and patience with which he guided me and for pushing me as hard as possible to improve. Without his guidance, I would not have been able to complete this doctoral dissertation. I would also like to thank my second supervisor, Herr Prof. Jacob Rosenthal, for the valuable classes and lessons I attended. They have sparked new ideas for my doctoral dissertation. Finally, I would like to thank Herr Prof. Marcelo de Araújo, who has accompanied me since I was an undergraduate in Brazil. I thank him for having read some of my chapters and for helping me a little bit with my English.

Secondly, I would like to thank my close friends who have either been with me personally or with whom I have had the pleasure of exchanging ideas throughout my time in Germany. I would like to thank André Gustavo da Rosa Ribeiro, Noriko Kutsuna, Fabian Michael Stöhr, Kurt Halter, Takayuki Kurihara, Thiago Lucena de Macedo Guedes, Gilmar do Nascimento and Ali.

Third, I would like to thank my family, who has always supported me emotionally during all my life, but especially in this period. Thank you very much, Martiniano Mota de Andrade Neto, Ione Terezinha Fior, Maria do Socorro Araújo do Amaral, and Artemisa Gouvêa Mota de Andrade.

Finally, I would like to thank the Deutsche Akademische Austauschdienst (DAAD) and the Conselho Nacional de Desenvolvimento
Científico e Tecnológico (CNPq) for having financed my studies during this long period. Without the funding, I would never have been able to start this project in the first place.
Introduction

This study was inspired by the interpretative controversy over Hume’s theory of justice. More specifically, it was inspired by a paper published by Gauthier in 1979, which introduced a new line of interpretation of Hume’s theory of justice, the contractarian interpretation. The interpretative debate about Hume’s theory of justice was pretty arid for a long time. A rule-utilitarian interpretation of Hume’s theory of justice has prevailed in the tradition of Hume scholars for many decades. In this context, Gauthier’s paper sparked a whole new interest in the subject. Gauthier popularized this new line of interpretation to the point that many other Hume scholars have supported it in the following years. In this way, Gauthier’s interpretation reintroduced the interpretative controversy over Hume’s theory of justice in the philosophical debate. It invited a recap of this interpretative dispute, and it called for a more careful and systematic study of Hume’s theory of justice once again to define which is the most plausible interpretation of it.

This study frames the interpretative discussion of Hume’s theory of justice in terms of three main lines of interpretation suggested by Hume scholars in the tradition: i) the evolutionary interpretation, ii) the rule-utilitarian interpretation, and iii) the contractarian interpretation. The evolutionary interpretation states that Hume’s theory of justice is best described as explaining the origins and the gradual development of the social and legal institutions over time. The rule-utilitarian interpretation states that Hume’s theory of justice is best described if the social and legal institutions
are comprehended as a set of social rules designed to maximize the overall sum of happiness or welfare of society. The contractarian interpretation states that Hume’s theory of justice is best described if the social and legal institutions are conceived as resulting from a unanimous assent among everyone. While the former interpretation is a descriptive account of Hume’s theory of justice, the latter interpretations are normative accounts of it. So, each line of interpretation, while representing a seemingly plausible description of Hume’s theory of justice, is fundamentally opposed to each other. I intend to show in this study that each of these lines of interpretation emphasizes a particular perspective adopted by Hume to explain justice in his theory.

In this study, I address in detail the set of rules that constitutes Hume’s theory of justice in light of these three lines of interpretation. For Hume, the rules of justice are the following three: i) the rule of the stability of possessions, ii) the rule of the translation of property by consent, and iii) the rule of the obligation of promises. This set of rules is agreed upon by individuals through a convention of justice. At times, Hume refers to these rules as the “three fundamental laws of nature”, which suggests that, for Hume, they are the rules that make up the most basic structure of a cooperative society. And, for Hume, the virtue of justice is the disposition of individuals to behave according to the pattern of behavior defined by this set of rules. Therefore, the main subjects of this study are the rules of justice and the virtue of justice in Hume. In particular, the rules of justice refer to Hume’s theory of property and Hume’s theory of promises, which properly constitute this theory of justice.

The inquiry into the most plausible interpretation of Hume’s theory of justice would, in itself, be a very relevant philosophical task. However, I
believe that this inquiry may be of even greater importance for the comprehension of Hume’s philosophy in general. In this study, I show that the natural virtues are, in a sense, dependent on the artificial virtue justice in Hume’s moral theory. If this interpretative claim is correct, it follows that the virtue of justice should be regarded as the focal point of Hume’s moral theory. Thus, this interpretative controversy must be settled not only because it is necessary to determine the correct interpretation of Hume’s theory of justice, but also because the correct interpretation of the virtue of justice acquires such an importance to the point that it shapes the entire comprehension of Hume’s moral theory. The inquiry into the most plausible interpretation of Hume’s theory of justice may be the most relevant philosophical task in Hume’s moral theory in general.

This study is divided into six chapters. In the first chapter, I provide a brief exhibition of Hume’s moral theory. I also argue for the interpretative claim that the natural virtues are conceptually dependent on the artificial virtue of justice in Hume. This interpretative claim entails that the inquiry into the nature of the virtue of justice is a priority in the discussion of Hume’s moral theory. As such, the virtue of justice in Hume requires a more careful and detailed clarification. For this reason, I introduce the three lines of interpretation of Hume’s theory of justice that have been suggested by Hume scholars throughout history as a guide to orient this study.

In the second chapter, I begin by introducing three philosophical assumptions of Hume’s moral theory: i) the instrumental conception of practical reason, ii) the artificial nature of the virtue of justice, and iii) the circumstances of justice. For Hume, this set of assumptions defines the conditions under which individuals agree to establish a set of rules of justice by an agreement or convention. These assumptions illustrate the structural
properties of the Humean state of nature, which is a social structure of interaction that logically precedes the establishment of the convention of justice. I represent the Humean state of nature by an analogy with game theory, for the choice of individuals to establish the set of rules of justice can be imagined as a solution for a problem of interdependent decisions in this particular social structure of interaction. In particular, I intend to explore the structural similarities between the Humean state of nature and the well-known game theory model of the Stag Hunt. The purpose is to identify the solution for the Stag Hunt and compare it with Hume’s account of justice, which could be interpreted as a solution for the problem posed by the pre-conventional structure of interaction.

In the third chapter, I examine the origins and foundations of Hume’s virtue of justice. Thus, I am concerned with providing a more careful and detailed analysis of Hume’s convention of justice, which is the central concept of Hume’s theory of justice. This analysis requires the clarification of some points. First, how do individuals come to establish a convention of justice? Second, how do individuals come to act according to the pattern of behavior defined by this convention? Third, what is the nature of this convention? Fourth, what are the rules of justice defined by it? Fifth, how should this convention be interpreted in terms of the social advantages that it affords? After clarifying the details of Hume’s convention of justice, I consider this set of rules of justice, in particular, the first rule of justice, the rule of the stability of possessions, in light of the three lines of interpretation of Hume’s theory of justice.

In the fourth chapter, I address the other rules that specify private property rights in Hume. The idea is that, given a rule to stabilize possessions, which ultimately defines private property rights, Hume has to come up with
other rules that specify the distribution and acquisition of private property. Those are the rules that define the initial distribution of possessions, the rules that define the further acquisition of private property, and the rule of the translation of property by consent. This set of rules comprises Hume’s theory of property, which is a central part of Hume’s theory of justice. Consequently, I also examine the set of rules that specify private property rights in Hume in light of the three lines of interpretation.

In the fifth chapter, I conclude the study of Hume’s theory of justice by considering the rule of the obligation of promises. This rule is related to Hume’s theory of promises, which is the other central part of Hume’s theory of justice. I point out that Hume’s theory of promises is very relevant in the history of moral philosophy. I introduce this account in contrast to a more traditional account of promises in Hume’s time, which is formulated by the natural law theory, to illustrate the differences between both theoretical accounts of promises. Finally, I also attempt to interpret the rule of the obligation of promises and, Hume’s theory of promises in general, in light of the three lines of interpretation. This chapter concludes the analysis of Hume’s “three fundamental laws of nature”. In this fifth chapter, the study of the set of rules that comprises Hume’s theory of justice is completed.

In the sixth and last chapter, I summarize the results of the analysis of Hume’s theory of justice in light of the three lines of interpretation. I highlight the conflicting nature of the different descriptions in relation to Hume’s textual source. Each of the three lines of interpretation makes some plausible points and contributions to a better comprehension of Hume. But each one of them fails to provide an entirely accurate description of Hume’s theory of justice as well. Such conclusions raise three questions. Why do the three main lines of interpretation fail to fully explain Hume’s theory of justice? Why
have the three main lines of interpretation been so influential among Hume scholars for so long? How can the interpretative dispute about Hume’s theory of justice be most reasonably settled? I answer these three questions in this last chapter. I consider the three lines of interpretation by surveying their degrees of plausibility in face of the textual evidence in Hume. Therefore, I attempt to develop a correct interpretation of Hume’s theory of justice from the most plausible line of interpretation ascertained by this study. My suggestion is that this interpretative controversy can be settled by finding out which description is able to best harmonize two features of Hume’s theory of justice: his naturalism and his normative character. Therefore, I attempt to develop a correct interpretation of Hume’s theory of justice from the most plausible line of interpretation ascertained by this study.

In this study, I used as reference the editions of L. A. Selby-Bigge from *A Treatise of Human Nature* (1978), *An Enquiry Concerning Human Understanding* (1963), and *An Enquiry Concerning the Principles of Morals* (1963). The version of *The History of England* (1983) is from Liberty Classics. Finally, I used the version of Hume’s *Essays and Treatises on Several Subjects* (2002) introduced by James Fieser in two volumes. The page numbers and passages quoted are all from these editions.
CHAPTER 1

Hume’s Moral Theory and the Virtue of Justice

1.1 Some Preliminary Remarks

Some preliminary remarks concerning this first chapter are necessary to understand how it stands in relation to the whole work. I would like to point out that the main purpose of this study is to engage in the interpretative debate on Hume’s theory of justice. This theory accounts for the origins and reasons for the establishment of the institutions of justice in society. In this sense, the inquiry into the plausibility of the different possible lines of interpretation that have been suggested to accurately describe Hume’s theory of justice is what ultimately drives this work. In particular, there are three main lines of interpretation that concern this study the most: i) the evolutionary interpretation, ii) the rule-utilitarian interpretation, and iii) the contractarian interpretation. As such, I intend to provide a careful and detailed analysis of Hume’s theory of justice to find out which interpretation better suits to describe it.

It is well known that Hume developed a theory of moral sentiments. This kind of moral theory is based on the passions, affections, and sentiments of moral approbation and disapprobation that agents are supposed to entertain from the contemplation of the character traits of others. For the specific purposes of this study, in this first chapter, I intend to introduce Hume’s
theory of justice in the context of his theory of moral sentiments. This theory addresses Hume’s moral psychology and normative ethics. It is concerned with the explanation of the way by which individuals are able to discern moral values, the nature of moral evaluations in general, the sources and standards of moral evaluation, and the proper objects of moral evaluation.

I consider it necessary to give an overview of Hume’s moral theory in this first chapter. This overview will include the exposition of Hume’s main moral concepts, such as that of virtue and vice, moral approbation and disapprobation, utility or usefulness, agreeableness, sympathy, benevolence, and justice. Although Hume’s theory of moral sentiments is not the focus of this study, it is important to point out that this exposition will not be devoid of some critical and interpretative remarks of my own. This is because, as I intend to argue, the conclusions achieved in the course of this study on Hume’s theory of justice depend on the comprehension of Hume’s moral theory and the very nature of his philosophical approach in general. Any alleged interpretation that disregards Hume’s theory of moral sentiments must be deemed as implausible or at least incomplete.

After the exposition of Hume’s main moral concepts, I will introduce his theory of justice by means of the well-known distinction between natural and artificial virtues in Hume’s moral theory. Although this distinction can be called into question based on some textual evidence in Hume’s later work, I intend to argue that it is still meaningful and interpretatively sound to preserve the distinction and to characterize the virtue of justice as artificial, according to Hume’s moral theory. However, I would like to point out that the main purpose of introducing justice as an artificial virtue in this first chapter is to show how it relates to the natural virtues in the context of Hume’s theory of moral sentiments. More specifically, I intend to claim that the
relation between natural virtues and the artificial virtue of justice is characterized by a kind of conceptual dependence of the former on the latter. This interpretation entails that the virtue of justice is the most important concept of the entire Hume’s moral theory. By implication, it enables us to have a better and far more systematic comprehension of the pivotal role of the virtue of justice in Hume’s moral theory. Other interpretations often overlook this point. Consequently, this interpretative claim draws our attention to the importance of correctly understanding the nature of justice in Hume.

1.2 Virtue, Vice, and the General Point of View

1.2.1 Definition of Virtue and Vice

The concepts of virtue and vice are central to Hume’s moral theory. For Hume, discerning moral values is a matter of ascertaining the virtuosity or viciousness of particular objects that can be morally evaluated. So, Hume can be rightly regarded as a kind of virtue ethicist, for he regards virtues and vices as making up the core of morality in general. The analysis should start then by considering Hume’s definitions of virtue and vice in *A Treatise of Human Nature* (1739-1740) and *A Enquiry Concerning the Principles of Morals* (1751). The definitions of virtue and vice are often rephrased in different ways by Hume, but the meaning does not seem to essentially change in the course of his philosophy. So, for the present purposes, I will introduce three definitions. In the *Treatise* Book 3, Part 1, Section 2, Hume defines virtue and vice as follows:

An action, or sentiment, or character is virtuous or vicious; why? because its view causes a pleasure or uneasiness of a particular kind. […] To have the sense of virtue,
is nothing but to feel a satisfaction of a particular kind from the contemplation of a character. (T 3.1.2.3/SBN 471)

In the Treatise Book 3, Part 2, Section 5, Hume rephrases his definition to incorporate additional features. The contemplation is characterized as a “general survey or view”, and moral approbation and disapprobation are identified with the passions of love and hatred:

The pain or pleasure, which arises from the general survey or view of any action or quality of the mind, constitutes its vice or virtue, and gives rise to our approbation or blame, which is nothing but a fainter and more imperceptible love or hatred. (T 3.3.5.1/SBN 614)

In a similar vein, in the Enquiry Appendix I, Hume provides another definition of virtue and vice:

The hypothesis which we embrace is plain. It maintains that morality is determined by sentiment. It defines virtue to be whatever mental action or quality gives to a spectator the pleasing sentiment of approbation; and vice the contrary. (EPM App. 1.10/SBN 289)

What these definitions seem to show is that Hume defines virtue and vice in terms of particular sentiments of pleasure or pain aroused in a given observer from the view, survey, or contemplation of a character trait, an action, or a mental quality. Notice that these definitions bring some questions. Hume’s definition regards virtue and vice as functions of a pleasant sentiment of approbation or an unpleasant sentiment of disapprobation of a spectator from a general view. Thus, for Hume, the very concept of virtue is conceptually connected with a particular stance of moral evaluation, by which a given spectator defines whether it is a case of virtue or not depending on subjective responses. Hume does not define virtue and vice by their essential properties, but by the way in which they are capable of arousing the pleasant or unpleasant subjective responses of spectators in moral evaluations. Thus,
Hume’s definition of virtue seems to conflate a normative concept, such as that of virtue, in the definiendum, with a phenomenon of moral psychology, such as that of moral evaluation, in the definiens. There are other problems that this definition entails, but what makes it problematic at first glance is that, if virtue and vice are defined as things capable of arousing the pleasant or unpleasant sentiments of approbation or disapprobation of spectators in moral evaluations, then one wonders by virtue of what they are capable of arousing such sentiments in the first place. But, Hume simply fails to mention it in the scope of his definition. Had he done that, this would be equivalent to define virtue and vice by at least one of their essential properties and not simply by the subjective response of a moral spectator. It is not the case that Hume does not have anything to say about this. I will address Hume’s account of the sources of moral approbation and disapprobation later on in this chapter. Therefore, it is more useful to turn back to Hume’s definition of virtue and analyze it before considering its problems.

1.2.2 Actions, Mental Qualities, and Character Traits
The definition is composed of three elements. Each of these demands clarification. First, it is specified a set of objects that can be morally evaluated. Second, it is defined a criterion for appropriate moral evaluations. Third, it is established the way by which individuals come to perform moral evaluations. Thus, what is the set of objects of moral evaluation? In the definition, Hume mentions “actions”, “mental qualities”, and “character traits” as the set of objects that can have moral value. But, such objects are very different kinds of things. Actions are essentially temporary. Mental qualities can be either temporary or constant, variable or stable. And character traits are regarded as a kind of long-lasting mental qualities, even if we happen to act contrary to
them occasionally. In face of such candidates for morally evaluable objects, Hume argues that motives or actions should be morally approved or disapproved to the extent that they constitute evidence of a stable disposition of the agents’ character (T 2.3.2.6/SBN 410-411). Hume has two arguments to support this claim. First, Hume brings the case of sudden and unreflected motives or actions. In cases where agents are compelled to make decisions and perform actions abruptly, without the due reflection, such motives or actions do not morally tarnish their character, for they do not reflect the real “principles of the mind” of the agents (T 3.3.1.4/SBN 575). Second, Hume also argues that the agents’ character can still be morally approved or disapproved by others even if the decisions they make and the actions they perform by the influence of these character traits eventually fail to secure their intended ends (T 3.3.1.19/SBN 584).

It could be objected that Hume is wrong in claiming that motives or actions are not the proper objects of moral evaluation. In order to understand this point, it is important to pay closer attention to Hume’s philosophy as a whole. Hume is concerned with a general survey of human nature. This is the guiding theme of his entire philosophy. And Hume suggests a scientific and naturalized approach to conduct this investigation. This survey is referred to

1 In this context, the meaning of terms like “mental quality” or “quality of mind” is supposed to cover a wide range of mental states. At times, Hume refers to transitory mental states such as passions, sentiments, or particular motives that lead individuals to act the way they do. At other times, he refers to certain deep-rooted dispositions that regularly guides individuals to act, such as character traits.

2 Hume’s point with these two arguments is that the performance of hasty and unexpected actions, especially if they are followed by a true regret or a change in the behavior, cannot guarantee a reliable moral evaluation, for they do not express the real motives of the agents, which can only be ascertained by surveying more carefully into their character traits. The same happens if the actions fail to produce their common effects. Even if the consequences of certain actions are not manifest to us, it is still possible to morally evaluate the motives and reasons, which are also a function of the agents’ character.
as the “science of human nature” by Hume. In the Introduction of the Treatise, Hume describes this approach as based on “an attempt to introduce the experimental method of reasoning into moral subjects”. This means that Hume is trying to find out the general causes and effects of human behavior based on the methods of experimental philosophy, namely, by means of empirical investigation. This involves developing mental experiments, observing effects, identifying causes, and tracing causal relations between them (T xx-xxi/SBN 6-7). Thus, Hume, as I interpret him, is not so much concerned with how individuals behave strictly on a case-by-case basis. Instead, Hume is concerned with how actions, character traits, and mental qualities can be articulated in an inquiry into the general regularities and tendencies of human behavior. With regard to the explanation of the moral evaluation, this point is especially relevant for Hume, since the more general causes that explain human behavior produce a more reliable sentimental reaction in the spectators. This seems to be the underlying assumption of Hume’s point, as some passages of the Treatise evidence it:

We are never to consider any single action in our enquiries concerning the origin of morals; but only the quality or character from which the action proceeded. These alone are durable enough to affect our sentiment concerning the person. Actions are, indeed, better indications of a character than words, or even wishes and sentiments; but ‘tis only so far as they are such indications, that they are attended with love or hatred, praise or blame. (T 3.3.1.5/575)

In a footnote of the Enquiry Section V, Hume also calls attention to the relevance of the general tendencies and regularities of human behavior for explaining moral evaluations, as it follows:

For a like reason, the tendencies of actions and characters, not their real accidental consequences, are alone regarded in our moral determinations or general judgments
So, although Hume mentions actions and other mental qualities as objects that could, under the right circumstances, be morally evaluated, since that they are essentially temporary, they should be understood, as a general rule, in relation to the kind of long-lasting mental states that reflect how the agents behave most of the time. And these are, in general, character traits. For Hume, character traits are the most important objects to explain moral evaluation.

1.2.3 The General Point of View

Which criterion must be satisfied for appropriate moral evaluations? One of Hume’s definitions of virtue and vice refers to the idea of a “general survey or view” of actions, mental qualities, and character traits. But, what is this expression supposed to mean? Hume understands that individuals could be subjected to a strong variation of passions and sentiments, which is determined by the degree of relative proximity with the subjects of moral evaluation. This could affect the way by which each spectator feels the sentiments of approbation and disapprobation in each case. Ultimately, this contributes to a kind of partiality of individuals in terms of moral approbation and disapprobation, which is a characteristic that must not feature in any kind of theory of moral evaluation. Hume acknowledges this difficulty and suggests that an appropriate moral evaluation can only be successfully performed when the agent adopts “steady and general point of view”. By adopting this point of view, agents overlook their own interests and imagine themselves in a distanced and more comprehensive perspective of the whole situation. In other words, what Hume means by a “general survey or view” or
a “steady and general point of view” is that individuals must adopt an imagined and impartial perspective for moral evaluations. In the *Treatise* Book 3, Part 1, Section 2, Hume argues that character traits must be morally evaluated in a disinterested way, that is, impartially:

Nor is every sentiment of pleasure and pain, which arises from characters and actions of that *peculiar* kind, which makes us praise or condemn. The qualities of an enemy are hurtful to us; but many still command our esteem and respect. ‘Tis only when a character is considered in general, without reference to our particular interest, that it causes such a feeling or sentiment, as denominates it morally good or evil. (*T* 3.1.2.4/SBN 472)

And, in the *Treatise* Book 3, Part 3, Section 1, Hume mentions the “steady and general point of view” as a way of overcoming the problems related to the partiality of individuals, which result from the particular variation of passions and sentiments of each one of them. The following passage provides examples of this point:

The approbation of moral qualities most certainly is not deriv’d from reason, or any comparison of ideas; but proceeds entirely from a moral taste, and from certain sentiments of pleasure or disgust, which arise upon the contemplation and view of particular qualities or characters. Now ’tis evident, that those sentiments, whencesoever they are deriv’d, must vary according to the distance or contiguity of the objects; nor can I feel the same lively pleasure from the virtues of a person, who liv’d in Greece two thousand years ago, that I feel from the virtues of a familiar friend and acquaintance. Yet I do not say, that I esteem the one more than the other. […] In order, therefore, to prevent those continual *contradictions*, and arrive at a more *stable* judgment of things, we fix on some *steady* and *general* points of view; and always, in our thoughts, place ourselves in them, whatever may be our present situation. (*T* 3.3.1.15/SBN 581-582)

And, when Hume addresses aesthetic evaluation, which he considers to be analogous to moral evaluation, it is described as requiring the adoption of an imagined and unbiased “point of view”, by which any personal consideration is abstracted away, and only the object of survey is taken into account. Spectators must conceive themselves in the unbiased position of a “man in general”. In the absence of this stance, the “sentiments” would be “perverted”
and both aesthetic and moral evaluations would miss the point entirely. It is
in the essay Of the Standard of Taste (1757) that Hume introduces this general
point of view once again:

But to enable a critic the more fully to execute this undertaking, he must preserve
his mind free from all prejudice, and allow nothing to enter into his consideration,
but the very object which is submitted to his examination. We may observe, that
every work of art, in order to produce its due effect on the mind, must be surveyed
in a certain point of view, and cannot be fully relished by persons, whose situation,
real or imaginary, is not conformable to which is required by the performance. […]
In like manner, when any work is addressed to the public, though I should have a
friendship or enmity with the author, I must depart from this situation; and
considering myself as a man in general, forget, if possible, my individual being and
my particular circumstances. A person influenced by prejudice, complies not with
this condition; but obstinately maintains his natural position, without placing
himself in that point of view, which the performance supposes. […] By this means,
his sentiments are perverted; nor have the same beauties and blemishes the same
influence upon him, as if he had imposed a proper violence on his imagination, and
had forgotten himself for a moment. (ST 21/ 239-240)

In the Enquiry Appendix I, Hume brings additional conditions that must be
satisfied for appropriate moral evaluations. Hume argues that, in moral
deliberations, the spectators must be fully informed regarding all the relevant
facts and circumstances in which the agents are involved. They must also be
able to know how the different tendencies of actions and character traits may
affect the agents involved in each particular case. If the spectators are not well
acquainted with all the relevant facts, they are likely to make false moral
evaluations. That is why, in cases of ignorance regarding all the relevant facts,
Hume recommends the suspension of moral judgment. For Hume, it is at this
point that the understanding can operate in cooperation with passions and
sentiments in moral deliberations. Understanding plays a pivotal role in
acquiring knowledge of the relevant matters of fact in each particular case.
This entails that moral evaluations become rationally informed and, thus,
truth-apt as well. It is only after the spectators, by means of reflection, have
set all the relevant facts and circumstances that an appropriate subjective response can be elicited, and moral evaluations can be successful. Hume describes such conditions as follows:

But in moral deliberations we must be acquainted beforehand with all the objects, and all their relations to each other; and from a comparison of the whole, fix our choice or approbation. No new fact to be ascertained; no new relation to be discovered. All the circumstances of the case are supposed to be laid before us, ere we can fix any sentence of blame or approbation. If any material circumstance be yet unknown or doubtful, we must first employ our inquiry or intellectual faculties to assure us of it; and must suspend for a time all moral decision or sentiment. While we are ignorant whether a man were aggressor or not, how can we determine whether the person who killed him be criminal or innocent? But after every circumstance, every relation is known, the understanding has no further room to operate, nor any object on which it could employ itself. […] In moral decisions, all the circumstances and relations must be previously known; and the mind, from the contemplation of the whole, feels some new impression of affection or disgust, esteem or contempt, approbation or blame. (EPM App. 1.11/SBN 290)

Nevertheless, this kind of reflection presented by Hume in the Enquiry is not the only reason why moral evaluations acquire a cognitive status. Still in the Treatise, Hume characterizes the general point of view as a “survey or reflexion”, which indicates that the very activity of conceiving oneself in this impartial stance is already, to some extent, cognitive (T 3.3.1.3/SBN 574-575). So, according to Hume, the successful performance of moral evaluations depends that the spectators adopt the impartial stance of the general point of view and satisfy the criteria observed in the passage aforementioned. First, the spectators must be fully informed about all the relevant facts and circumstances in which the agents are involved. Second, the spectators must be fully informed about how the tendencies of actions and character traits may affect each agent involved. And third, the spectators must remain disinterested in how the tendencies of actions and character traits may come to benefit or harm themselves or those related to them. Therefore, although Hume characterizes this general point of view as a kind of impartial
stance in moral evaluations, it could also be characterized as incorporating the conditions set up by Hume in the Enquiry, as I interpret it. This is the way by which a more distanced, unbiased, and comprehensive perspective for moral evaluations can be achieved.

Notice that my characterization tends to depict Hume as committing himself to the idea that the general point of view must be the perspective of the Ideal Observer, an imagined stance in which spectators need to conceive themselves as a condition for making appropriate moral evaluations. This entails that moral evaluations are not related to the sentiments of actual agents, but rather to the sentiments of agents when they imagine themselves as impartial spectators. This interpretation has been suggested by the majority of philosophers when they interpret Hume. The perspective of Ideal Observers could also be characterized as an imagined stance where spectators are omniscient, equi-sympathetic, and disinterested, as Geoffrey Sayre-McCord puts it. For Sayre-McCord, the distinctive feature of Ideal Observers is that spectators must have full knowledge of all the actual effects of actions and character traits on everyone, which amounts to a kind of omniscience. It is clear that if the perspective is characterized as such, then problems related to the very possibility of the existence of this perspective would follow. However, it is not clear whether Ideal Observers need to be characterized in such extreme terms. The idea that individuals can successfully perform impartial moral evaluations while acknowledging their limits in knowledge and sentimental capacities is completely plausible. In fact, natural constraints in the knowledge of matters of fact and in the scope of the affective responses

---

of individuals is a presupposition of Hume’s philosophy. In Hume’s words, there is a natural propensity of each individual to be more affectively attached to those in their “narrow circle”, namely, relatives and acquaintances (T 3.3.3.2/SBN 602). And yet, despite the natural “degree of selfishness in men”, individuals are still able to abstract from such considerations and adopt an impartial stance for moral evaluations through a suitable reflection (T 3.3.1.17/SBN 582-583).

Under such conditions, it is appropriate to interpret Hume’s general point of view as the perspective of an Ideal Observer. This view is frequently associated with Adam Smith, who first formulated it in The Theory of Moral Sentiments (1759). To assign a similar interpretation to Hume could be deemed as reading him through Smith’s eyes, since the concept was more clearly spelled out not by the former, but by the latter.5 However, I do not think that this is the case, quite the opposite. It is only the very extended logical analysis of Hume’s concept of the general point of view in the course of his work that justifies this interpretative claim. As I have pointed out, Hume starts with the idea of a “survey” or “view” and, in the course of his of work, introduces further criteria that ultimately allow us to characterize it as the stance of an Ideal Observer. It is well known that Smith’s theory of moral sentiments was influenced by Hume and, most certainly, Smith’s theory of

the Ideal Observer should be regarded more as an expression of this influence than an exception of it.

1.2.4 Moral Sentiments

What is the way by which moral evaluations are made? As the definition shows, moral evaluations are made by taking into account how spectators would feel if they adopted the general point of view. More specifically, moral evaluations are related to a kind of sentiment that would be aroused in the spectators when they conceive themselves in the imagined stance of the general point of view. On the one hand, if spectators feel a pleasant sentiment of approbation in considering a specific character trait, it will be regarded as virtuous. On the other hand, if spectators feel an unpleasant sentiment of disapprobation in considering another specific character trait, it will be regarded as vicious.

To better understand the nature of the moral sentiments, it is necessary to understand the relationship between this particular kind of sentiments and the impartial stance of the general point of view. It is important to point out that it is the “general survey or view” that ensures that the particular sentiments of pleasure and pain felt by the spectators acquire their distinctive moral quality (T 3.1.2.4/SBN 472). Hume’s point here is that many different kinds of sensations can be described as pleasant and unpleasant, but their particular qualitative differences can still be recognized. For instance, there is no doubt that anyone would recognize that physical pleasure or pain and psychological pleasure or pain are quite different in kind, even if sometimes they may be causally related. Even when it comes to sensory experiences derived from outer senses, categorical differences are still recognizable. As Hume points out, it is not said of a pleasant music that is has a “good flavour”
or of a pleasant wine that it is “harmonious”. Therefore, one of the things that Hume intends with the concept of the general point of view is to single out the particular moral sentiments from all other kinds of subjective responses that individuals may have. This can be accomplished by defining a standard that specifies necessary and sufficient conditions for the identification of such sentiments whenever they arise. In the particular case of the moral sentiments, they can be identified through their arousal in impartial spectators, who must take as the object of their contemplation the character traits of agents and their tendencies to affect other individuals. In this way, it is possible to draw a categorical distinction between pleasant and unpleasant sensations of moral approbation and disapprobation and other sensory experiences of pleasure and pain.

As I have pointed out, moral sentiments are appropriately aroused in impartial spectators only when moral evaluations are rationally informed, namely, when they are performed from a point of view that satisfies the three criteria mentioned in the last section. This entails that agents, when conceiving themselves in the impartial stance of the general point of view, must engage in a cognitive activity in order to acquire the necessary amount of factual knowledge for the arousal of the appropriate subjective responses. At times, Hume uses the analogy of the visual sensory perception to illustrate the effects of the general point of view. Just as an episode of visual sensory perception must occur under standard observational conditions for it to be considered reliable, moral evaluations must be performed from the general point of view for them to be true. Hume thinks that the general point of view works as a standard for moral evaluation that enables agents to avoid the influence of variable and biased feelings and to develop constant and adequate moral sentiments. It operates by changing their perception of the
current state of affairs and by correcting their affective partiality towards others (T 3.3.1.15/SBN 581-582; EPM 5.42/SBN 227-228).

There is a further point regarding the relationship between moral sentiments and the impartial stance of the general point of view. It concerns the communication of moral sentiments. By invoking the general point of view, Hume is concerned with settling a linguistic issue as well. For Hume, individuals need to establish a certain standard without which it would be impossible for them to be able to communicate their moral sentiments to others. In other words, for Hume, there would be no language of morality, strictly speaking. Without the general point of view, moral discussions and controversies would not come to any conclusive result, but only yield contradictions and disappointments, for individuals would use their own conceptual framework to refer to their own biased sentiments (T 3.3.1.15/SBN 581; EPM 5.42/SBN 229). This is because these sentiments would be associated with character traits that would be represented as they appear to individuals in their particular perspectives at a particular time. Hume’s suggestion seems to be as follows. Among all the rules for the correct use of our common concepts, the general point of view might work for the correct use of moral concepts. In Hume’s particular case, such concepts are those of moral “blame” or “praise” (T 3.3.1.16/SBN 582). A public standard is required in order for us to convey our sentiments to others intelligibly through our ordinary framework of moral concepts. This entails that moral sentiments can only be properly conveyed when individuals adopt this public standard for moral evaluations. For Hume, the general point of view works as this public and intersubjective standard that makes it possible for us to morally approve and disapprove character traits as they really are, rather than just as they appear to each one in their purely subjective and biased points of
view. This is how we can safely expect mutual comprehension regarding moral sentiments in regular conversations and debates.

It seems to me that Hume is right in assuming that a public and intersubjective standard for communicating moral sentiments must entail an impartial point of view. In fact, any conception of morality, regardless of endorsing moral sentiments or not, must assume an impartial point of view as a public standard for moral evaluations. And the way in which Hume characterizes the nature of the general point of view is quite plausible. Although the reflective activity of conceiving oneself in this impartial stance is depicted as a subjective performance, the emergence of it depends on the social context. As Hume puts it, the general point of view emerges with the “converse with mankind” and “greater social intercourse” (EPM 5.42/SBN 228). The only problem with this picture is that it is not clear whether the general point of view, which is supposed to be the public and intersubjective rule for the correct use of moral concepts, is supposed to correct the sentiments of the agents for moral evaluations, as Hume believes. In fact, it may be the case that the rule for the correct use of moral concepts should not refer to the sentiments of the agents at all. The only safe conclusion is that a rule for the correct use of moral concepts should assume, as a necessary condition, an impartial stance. Therefore, the claim that relates our framework of moral concepts with the general point of view for correcting the moral sentiments of the agents should not be simply taken for granted, for it clearly depends on the plausibility of Hume’s moral theory as a whole.

1.3 Sources of Moral Approbation and Disapprobation
The analysis of the definition of virtue and vice covered its three main elements. They are the character traits, the general point of view, and the
moral sentiments. This definition entails some problems. Part of them seems to derive from the fact that virtue is defined in terms of the moral evaluation of a spectator. For instance, as Harrison and Rawls have pointed out, there are problems related to circularity, for virtue is defined as that kind of character trait that is morally approved by the impartial spectator, while the impartial spectator is defined as that kind of agent that morally approves virtuous character traits.\(^6\) This circularity problem is related to the point that I have mentioned earlier, namely, that Hume does not define virtue by its essential properties. In the scope of the definition, Hume does not say by virtue of what virtuous character traits are capable of arousing pleasant sentiments of approbation in spectators in moral evaluations. So, Hume has to supplement his definition somehow. Moral approbation and disapprobation must be defined by more than just the subjective responses of the impartial spectator.

In the *Enquiry* Section IX, Part I, Hume makes the following suggestion based on his survey and “enumeration” of the different kinds of virtues and vices:

> Yet, I must confess, that this enumeration puts the matter in so strong a light, that I cannot, *at present*, be more assured of any truth, which I learn from reasoning and argument, than that personal merit consists entirely in the usefulness of agreeableness of qualities to the person himself possessed of them, or to others, who have any intercourse with him. (*EPM* 9.13/SBN 278)

In the *Treatise* Book 3, Part 3, Section 1, Hume states this claim:

> Thus, to take a general review of the present hypothesis: Every quality of the mind is denominated virtuous, which gives pleasure by the mere survey; as every quality, which produces pain, is call’d vicious. This pleasure and this pain may arise from four different sources. For we reap a pleasure from the view of a character, which is

---

naturally fitted to be useful to others, or to the person himself, or which is agreeable
to others, or to the person himself. (T 3.3.1.30/SBN 590-591)

As the aforementioned passage shows, Hume suggests a “review” of his
“hypothesis”. This hypothesis refers to what Hume states next in the passage,
namely, the definition of virtue and vice considered so far. And the review is
meant to be a supplement of the definition. Hume argues that the pleasant
sentiment of moral approbation is determined by two basic moral categories:
utility or usefulness and agreeableness. In turn, unpleasant sentiments of
moral disapprobation are determined by the opposite categories: disutility or
uselessness and disagreeableness. Hume considers that there are as many
sources of moral approbation and disapprobation as there are senses of the
concepts of utility and agreeableness. For Hume, there are four different
sources of utility and agreeableness and, therefore, of moral approbation and
disapprobation: i) utility of the character trait to others, ii) utility to the person
to whom the character trait belongs, iii) agreeableness of the character trait to
others, and iv) agreeableness to the person to whom the character trait
belongs. This entails that any character trait that can be morally evaluated as
virtuous by spectators in the impartial stance of the general point of view must
either have the quality of being useful or the quality of being agreeable either
to the agents themselves or to others or society. Every virtuous character trait

7 It is important to point out that Hume does not draw any sharp or precise conceptual
distinction between “utility” or “usefulness”. At times, it seems that Hume makes subtle
distinctions when he characterizes “utility” as a moral general idea or a moral standard
(EPM 4.20/SBN 211; EPM 6.25/SBN 244), whereas “usefulness” is regarded as an
instance of this more general “utility” in particular actions, mental qualities, and
character traits (EPM App. 1.2/SBN 285). However, this distinction proves itself to be
blurry, for, in other moments, Hume also emphasizes the “utility” of several virtues and
character traits, such as, for instance, that of courage and benevolence (EPM 7.11-
19/SBN 254-257). Therefore, it is safe to assume that Hume regards both of them as
synonyms.
must be possible to be predicated as being useful or agreeable to the agents themselves or to others. At this point, one can understand in virtue of what certain character traits are capable of arousing pleasant sentiments of approbation in spectators in moral evaluations and others not. For, as Hume describes, there are some essential properties that define virtuous and vicious character traits. Impartial spectators feel pleasant sentiments of approbation from the contemplation of a character trait insofar as it is useful or agreeable to the agents themselves or to others. Therefore, Hume’s definition of virtue and vice could be rephrased as such: virtues and vices are the kinds of character traits that induce pleasant and unpleasant sentiments of approbation and disapprobation from a survey of impartial spectators, for they are either useful or agreeable to the agents themselves or others. In the Treatise Book 3, Part 3, Section 5, almost at the end of the work, Hume makes sure to introduce these further elements to the reformulated definition of virtue and vice in order to “justify more fully that hypothesis”:

> The pain or pleasure, which arises from the general survey or view of any action or quality of the mind, constitutes its vice or virtue, and gives rise to our approbation or blame, which is nothing but a fainter and more imperceptible love or hatred. We have assign’d four different sources of this pain and pleasure; in order to justify more fully that hypothesis, it may here be proper to observe, that the advantages or disadvantages of the body and of fortune, produce a pain or pleasure from the very same principles. The tendency of any object to be useful to the person possess’d of it, or to others; to convey pleasure to him or to others; all these circumstances convey an immediate pleasure to the person, who considers the object, and command his love and approbation. (T 3.3.5.1/614)

Now, the question is whether Hume is capable of articulating a normative account from the introduction of such properties. In other words, the question is whether Hume is capable of justifying the moral standards that define virtue and vice. I believe that Hume can be interpreted as formulating a normative moral theory. This claim is surely controversial. Some philosophers would
not go as far as to argue that Hume has ever developed a normative ethical theory. They argue that Hume’s moral theory is only a description of the collective moral psychology of human beings when they engage in moral evaluations, instead of a proper justification of the standards that enables us to distinguish certain character traits as morally good and others as not.

So, to understand the point for which I intend to argue, it is necessary to start with the essential properties of virtuous and vicious character traits introduced by Hume. Of course, it is important to be careful with my characterization. By “essential properties” is not meant an overly metaphysical or essentialist conception of virtue and vice. As I have pointed out, Hume’s method consists in the empirical investigation of human nature. His account is heavily based on observation and inductive generalizations. And, the way by which Hume undertakes the task of figuring out the standards of morality is just an extension of this kind of empirical survey that attempts to identify which character traits tend to be morally approved and disapproved by human beings in general. Thus, Hume’s method of searching for “general” and “universal principles” upon which moral approbation and disapprobation are based is nothing more than the Aristotelian strategy of enumerating character traits in a catalog of virtues and vices after a suitable empirical survey (T 3.1.2.6/SBN 473; EPM 1.10/SBN 173-174). In this way, Hume singles out which properties are common to each one of the character traits that tend to be regarded as virtuous. Thus, the properties of being useful or agreeable to the individuals themselves or to others are “essential” to virtues only insofar as they are discovered by empirical inquiry. To the category of virtues useful to the individuals themselves belong the virtues of prudence,

---

temperance, frugality, discretion, good sense, discernment, industry, assiduity, enterprise, dexterity, and many others. To the category of virtues useful to others or to society belong the virtues of generosity, humanity, benevolence, honesty, allegiance, fidelity, truth, chastity, and justice. To the category of virtues agreeable to the individuals themselves belong the virtues of good humor, the greatness of mind or spirit, courage, tranquility, the delicacy of taste, and natural abilities, such as beauty and fortune. And to the category of virtues agreeable to others or to society belong the virtues of wit, ingenuity, eloquence, and good manners, such as politeness, modesty, and cleanliness.

Once again, it still could be pressed the point that this kind of account does not explain why certain character traits are morally good and others not. In fact, it reinforces the argument that Hume did not develop a normative moral theory, since this empirical survey can only show which character traits tend to be morally approved by human beings in general. This is nothing more than an analysis of the collective moral psychology of human beings. Virtues are just the kind of character traits that most individuals happen to morally approve most of the time. And the alleged general and universal principles of morality are deduced from an empirical survey that ascertains how human beings commonly behave and not how they should behave. This is a pressing

9 It is important to make a brief remark here. Benevolence is a very important concept for Hume. Both in the *Treatise* and the *Enquiry*, benevolence is conceived as a kind of virtue. However, it is a slightly special one. It is a broader disposition of character as it covers many other virtuous character traits, such as, generosity, good-nature, friendship, gratitude, beneficence, compassion, and so on. For Hume, this set of derivative virtues make up a benevolent character, as he points out in the *Treatise*, whereas, in the *Enquiry*, theses virtues are characterized as “benevolent affections”. In the *Enquiry*, benevolence is also characterized by Hume as a kind of universal sentiment that plays the role of rendering individuals sensitive to any possible object of moral approbation or disapprobation. This point will be addressed in the next section.
objection. Normative ethical theories are concerned with introducing and justifying a standard according to which individuals should behave. And this is not what Hume seems to be doing in the course of his exposition. Hume merely provides us with a catalog of virtues and vices after observing how individuals effectively behave.

I think it is important to draw a distinction here between the moral standards and the method used to discover them. Hume’s moral theory is part of the science of man. It is intended to explain a particular social phenomenon by the methods of empirical investigation, namely, the phenomenon of morality. For this reason, it is by a comprehensive empirical inquiry that ascertains the set of human inclinations, attitudes, and sentiments that Hume is capable of coming up with the standards of moral approbation and disapprobation. But it does not follow from this that such moral standards are totally devoid of justification. Instead, moral standards can be justified because the character traits that conform to them either please human beings due to their natural psychological makeup or satisfy the common necessities of human beings, provided, of course, that the set of human inclinations, attitudes, and sentiments is well defined by Hume’s empirical inquiry. And human beings have a reason to morally approve the kind of behavior that enables them to satisfy their most different needs, whatever they may be. This is the reasoning that allows Hume to draw a distinction between character traits in a catalog of virtues and vices and criticize the “monkish virtues” prescribed by the theological ethics of his time, such as celibacy, fasting, penance, mortification, self-denial, humility, silence, and solitude. For Hume, such character traits, although highly praised and recommended by religious
people in society, should be considered vices, as they are neither able to please individuals nor satisfy their common necessities (EPM 9.3/SBN 270).

Here, the justification strategy is similar to a kind of reflective endorsement, just like Korsgaard has suggested in her interpretation of Hume.\(^\text{10}\) This interpretation claims that moral normativity is based on the endorsement given to the moral standards after a reflection on the empirical survey that reveals the origins, operations, and benefits that practices defined by these standards bring to our lives. So, to the extent that this empirical survey can be reflectively evaluated, the moral standards could be regarded as morally approved based on their desirable effects. Therefore, it is not necessarily the case that the moral standards suggested by Hume, albeit identified by purely empirical inquiry, are utterly devoid of justification. Or, in other words, it is not necessarily the case that a moral theory based on a general description of which character traits tend to be morally approved by human beings prevents us from having a reason to morally approve at least some of them. One only fails to understand the normative character of Hume’s moral theory when one overlooks or rejects this method of justification.

However, this interpretative claim still lacks textual evidence. It is important to show where this view can be supported in Hume’s text. At the very end of the Treatise, when Hume’s theory is fully explained, a passage is commonly referred to as supporting the idea that Hume’s entire moral investigation is, in fact, an activity of reflective endorsement:

> All lovers of virtue (and such we all are in speculation, however we may degenerate in practice) must certainly be pleas’d to see moral distinctions deriv’d from so a noble source, which gives us a just notion both of the generosity and capacity of human nature. It requires but very little knowledge of human affairs to perceive, that a sense of morals is a principle inherent in the soul, and one of the most powerful

\(^{10}\) See Korsgaard (1996: 63).
that enters into the composition. But this sense must certainly acquire new force, when reflecting on itself, it approves of those principles, from whence it is deriv’d, and finds nothing but what is great and good in its rise and origin […] According to their system, not only virtue must be approv’d of, but also the sense of virtue: And not only that sense, but also the principles, from whence it is deriv’d. (T 3.3.6.3/SBN 619)

As I interpret it, what Hume seems to suggest in this passage is that it does not require much effort to realize that a sense of virtue is part of human nature. A mere observation of reality is sufficient evidence of this. However, when one undertakes a reflective survey of this sense of virtue to find out its underlying principles in human nature, one ends up endorsing it by morally approving the actual workings of the sentiments, dispositions, and standards in light of the needs of human life. In Hume’s words, when one “reflects” on this “sense of virtue”, one does not only “approve” it, but one also approves the “principles” from “whence it is deriv’d”. A reflection on this description “finds nothing but great and good” in the “rise” and “origin” of this “sense of virtue”. So, if it is taken for granted that Hume’s empirical survey is accurate in the description of the psychological makeup of human beings, then moral sentiments and standards will be morally approved upon the reflection of their operations.

It would be helpful to consider an example of how Hume should be read in light of this interpretation. Consider, for instance, the concept of utility in Hume. In the Enquiry Section V, Hume puts forward the thesis that “utility” is commonly regarded as the primary moral standard. This can be inferred by Hume’s statement that utility is a standard addressed in “all moral decisions”, it is “inseparable” from all other “social virtues”, and is a “foundation” of the “chief part of morals”:

It appears to be a matter of fact, that the circumstance of utility, in all subjects, is a source of approbation: That it is constantly appealed to in all moral decisions
concerning the merit and demerit of actions: That the sole source of that high regard paid to justice, fidelity, honour, allegiance, and chastity: That it is inseparable from all the other social virtues, humanity, generosity, charity, affability, lenity, mercy, and moderation: And, in a word, that it is a foundation of the chief part of morals, which has a reference to mankind and our fellow-creatures. (EPM 5.44/SBN 231)

Hume seems to be arguing that the quality of a character trait of being useful, either for oneself or for others, provides the most important standard for moral approbation and disapprobation. Since useful character traits are suitable for either solving individual and collective problems, or achieving individual and collective ends, or even satisfying the common necessities of human beings, then it is also possible to claim that there is a reason for them to be morally approved of by individuals. And when individuals behave in this way, they enjoy the moral approbation of others, for human beings are the kind of beings who are disposed to behave in ways to solve individual or collective problems. Obviously, when it comes to the concept of utility, it must be borne in mind that it might refer to many different things. It might refer to my own utility, to the utility of others, and to the utility of society or mankind in general. And conflicts are expected to happen to the extent that some character traits may be useful to the interests of some to the detriment of that of others in particular cases. It all comes down to how Hume can provide an account that harmonizes the different utility claims. But this will be analyzed in the course of this study. So, back to the original point, this view is similar to the way Sayre-McCord interprets Hume. Sayre-McCord regards Hume’s moral theory as advancing the thesis that moral evaluations are determined by how well suited a given behavior is to solve the specific kind of problems that individuals typically face.\(^\text{11}\) And, as I understand it, Sayre-McCord also

goes as far as to recognize that Hume’s justification strategy of his moral theory consists in a kind of reflective endorsement, in that he points out that the moral sense “enjoys” a kind of “self-ratification”.12 As I have already stated, the point is that this pattern of behavior that individuals tend to morally approval is itself approved by individuals upon the reflection of its fitness to satisfy their individual and collective necessities.

In the context of the general discussion, this thesis comes down to this fact. Useful and agreeable character traits to the individuals themselves or to others evoke and command moral approbation by spectators in the impartial stance of the general point of view. Such character traits convey pleasant feelings of moral approbation to the impartial spectator because they have the qualities of being useful and agreeable. As far as the moral standard of utility is concerned, spectators in the impartial stance should morally approve character traits well suited for solving problems, achieving ends, and satisfying necessities, whereas they should disapprove the opposite, as a matter of rationality. Consider, as last, the following passage, for instance: “common interest and utility beget infallibly a standard of right and wrong” (EPM 4.20/SBN 211). Hume is arguing here that utility generates a moral standard. But how should this be interpreted? Under the interpretation here considered, it seems to be clear that the notion of utility plays a relevant causal role in moral evaluation. For Hume, spectators are disposed to morally approve useful character traits when they adopt the impartial stance of the general point of view. Nevertheless, I interpret Hume’s point as stronger. Useful character traits command moral approbation of impartial spectators too. According to this interpretation, Hume believes that both casual and

rational accounts come together and reinforce his moral theory in these two different fronts. Utility must not only be conceived as having a strong psychological influence in our patterns of behavior of approbation but also as being rationally mandatory for us.

1.4 Sympathy and Benevolence

As I have pointed out, Hume has supplemented his definition of virtue to incorporate two basic moral standards: utility and agreeableness. So, the definition of virtue can be rephrased as the kinds of character traits that induce a pleasant sentiment of approbation or disapprobation from the contemplation of impartial spectators in the general point of view, because they are either useful or agreeable to the agents themselves or to others who may be related to them. Therefore, as far as the analysis goes, it seems that Hume’s moral theory is based on two theses:

(1) Virtues and vices are defined by the arousal of pleasant and unpleasant sentiments of impartial spectators in moral evaluations.

(2) The arousal of pleasant and unpleasant sentiments of impartial spectators is derived from the qualities of character traits of being useful or agreeable either to the individuals themselves or to others.

This moral theory, as it is presented, still cannot fully explain how moral evaluations are possible. It is plausible to assume that impartial spectators might be naturally disposed to entertain pleasant sentiments of approbation with regard to the character traits of an agent which are useful or agreeable to others only to the extent that the spectators themselves are somehow concerned. And they might be naturally disposed to feel unpleasant sentiments of disapprobation in the opposite case. However, the question that can still be posed against this moral theory is the following one. How can
impartial spectators entertain pleasant and unpleasant sentiments of approbation and disapprobation of character traits of agents totally unrelated to them? Recall that I have already alluded to the fact that Hume believes that the scope of affective responses of individuals is quite restricted. The affective bonds of individuals do not tend to naturally extend beyond that of what is found in their “narrow circle”, the group of friends and family members. This poses a challenge for a moral theory that relies on a sentimentalist basis, such as Hume’s. Recall also that, in the last section, I have argued in favor of the interpretation that claims that the nature of the requirement of moral approbation by impartial spectators of useful and agreeable character traits to the individuals themselves or to others is a normative one. But this presupposes that impartial spectators are able to entertain pleasant and unpleasant sentiments of approbation and disapprobation with regard to character traits of agents unrelated to them in the first place. Therefore, Hume needs a way of bridging the distance between impartial spectators and other agents from the affective point of view.

This challenge can be met when Hume’s concept of sympathy is considered in this picture. Sympathy is a technical term in Hume’s philosophy. It is not supposed to mean the same as the common usage of the term. Consider, for instance, Smith’s concept of sympathy, which, to some extent, corresponds to that of the common sense. For Smith, sympathy refers to a psychological capacity of having a fellow-feeling with any sentiment that agents may entertain. To sympathize with someone, one must be able to imagine what it would be like to be in someone else’s situation in order to feel what the other claims to feel. In other words, sympathy requires that one
should place oneself in someone else’s position. This notion of sympathy as a kind of shared feeling by attempting to assume the others’ point of view is very similar to that of the common sense. Hume’s concept of sympathy is slightly different. It also refers to a psychological capacity of communication of sentiments and opinions among individuals, but it does not operate by imagining oneself in the position of others in order to feel what the other claims to feel. Rather, it starts with an inference from the overt behavior of others, which enables us to form a corresponding idea or thought. This idea is strengthened and enlivened to the point that it is converted to a passion similar to the original one felt by others. In this way, it is possible that others’ sentiments and opinions are communicated and felt by us as our own (T 2.1.11.2-3/SBN 316-317). Páll Árdal thinks that this psychological mechanism can work only if two conditions are satisfied in Hume’s theory: i) the idea of the “self” must mediate sympathetic relations, for it is the most intimately present and strong idea produced by us, and ii) there must be a relation of similarity between both communicating agents, so that they can be affected at all. It is worth noticing that, for Hume, animals are also endowed with this capacity. The relation of similarity is conceived so broadly that human beings can very well hold relations of sympathy to them and vice versa (T 2.2.12.6/SBN 398). Considering that sympathy is a psychological mechanism, it should be understood as a mechanical and causal process of communication of sentiments and opinions. At any rate, the principles of association of ideas such as, resemblance, contiguity in time and space, and

14 See Árdal (1966: 43-44).
causation, are expected to exert variable influence on the agents throughout their whole psychological process (T 2.1.11.4/SBN 317-318).

Sympathy plays a major role in Hume’s moral theory. It enables the communication of sentiments and opinions among individuals. More specifically, sympathy operates as the psychological mechanism by which spectators can become affectively aware of others’ sentiments and opinions in moral evaluations. Spectators come to have access to the pleasant and unpleasant sentiments of others and the opinions about what they consider useful or agreeable to them. This is possible because sympathy enables spectators to receive pleasant sentiments felt by the others in virtue of the useful and agreeable character traits that they cultivate. In this way, to the extent that spectators are able to establish sympathetic relations with others, it is, in principle, possible for them to feel pleasant sentiments of approbation from the contemplation of character traits useful or agreeable to the agents themselves and to others unrelated to them. And, in the opposite case, they are also able to feel unpleasant sentiments of disapprobation. Hume thinks that contiguity relations determine more strongly relations of sympathy. This partly explains the natural partiality that individuals hold to their “narrow circle”. Nevertheless, it is the similarity that all human beings share that allows them, among other things, to communicate moral sentiments and opinions corresponding to virtues and vices in moral evaluations (T 2.1.11.5/SBN 318). So, in principle, individuals who would take up the impartial stance of the general point of view will be able to morally evaluate character traits of agents unrelated to them, provided that they are endowed with the psychological capacity of sympathizing with them.

Of course, the role of sympathy is not limited to the moral evaluation of character traits not directly related to us. Sympathy is closely related to the
wide array of indirect passions, which implies that it permeates the whole process of moral evaluation. For instance, it plays a major role in the moral self-evaluation of agents, in that they can sympathize with the indirect passions of love and hate of others regarding their character traits and feel pride and humility accordingly (T 2.1.11.9/SBN 320). Sympathy is also conducive to the emergence of an intersubjective social perspective, in that it enables a “concurrence” of “sentiments” among individuals (T 3.3.3.2/SBN 602-603). In other words, the communication of sentiments and opinions caused by the interaction of many sympathies is a necessary condition for the emergence of the intersubjective and impartial standard for moral evaluation or the general point of view. It does not follow from this that sympathy, by itself, is conducive to impartial responses. Instead, it promotes the social harmonization of several different sentiments and opinions through their communication. That is why sympathy is sometimes regarded as the general principle of morals for Hume.15

This claim can only be regarded as justified if it is taken into account the concept of sympathy as formulated in the Treatise. The concept of sympathy is extensively treated and strongly emphasized in this earlier work. However, in later works, such as the Enquiry of 1751 and even in A Dissertation on the Passions of 1757, Hume seems to reserve a much less important position for this concept in his moral theory. Sympathy is not as much mentioned in these later works as before. In the Enquiry, the concept of sympathy is still characterized as a kind of communication or, as Hume puts it, a “contagion” of sentiments among agents. But, at times, Hume suggests that this communication presupposes a kind of natural disposition of having

a fellow-feeling with human happiness or misery. And sympathy is featured as this natural disposition \((EPM\ 5.45/\text{SBN\ 231};\ EPM\ 7.2/\text{SBN\ 251};\ EPM\ 7.29/\text{SBN\ 260})\).

As I have pointed out, in the *Enquiry*, the concept of sympathy is not that much highlighted. Instead, it is the concept of benevolence that is particularly emphasized. Benevolence is a natural and basic disposition of human beings that causes them to develop a wide array of sentiments. In particular, the sentiment of humanity, caused by a benevolent disposition, makes individuals capable of entertaining sentiments of moral approbation and disapprobation to the same kinds of objects of evaluation. Given that, according to Hume, benevolence and affection for humanity are fairly “universal” and “comprehensive” dispositions of human beings, it follows that the large majority of individuals would be disposed to morally approve and disapprove the same kinds of character traits “even of the persons most remote” \((EPM\ 9.4-5/\text{SBN\ 271-272})\). It is important to notice, however, that the communication of sentiments through relations of sympathy still plays a major role in Hume’s moral theory, as it is presented in the *Enquiry*. The communication of sentiments among individuals enables the emergence of an intersubjective social perspective. Sympathy, then, is still a necessary but not sufficient condition for suitable moral evaluations \((EPM\ 5.42/\text{SBN\ 229})\).

But, once again, what Hume still needs to explain so that his moral theory can be regarded as plausible is how impartial spectators could feel pleasant and unpleasant sentiments of moral approbation and disapprobation with regard to agents who are totally unrelated to them. In the *Treatise*, Hume argues that sympathy, as a psychological capacity of communication of sentiments and opinions among the members of the same or similar species, is what enables spectators to be aware of each others’ sentiments and opinions
in moral evaluations. In the *Enquiry*, sympathy also plays the role of communicating sentiments and opinions, but Hume feels that only sympathy is not a sufficient condition for suitable moral evaluations. Apparently, just being aware of others’ sentiments and opinions through a psychological mechanism of communication does not ensure convergence in moral evaluations. Thus, Hume has to postulate another concept that refers to a more basic disposition. And this disposition is benevolence. More than just a collective and mutual transmission of sentiments and opinions, benevolence ensures that agents are not wholly indifferent in their presence. It makes the large majority of human beings disposed to morally approve and disapprove of the same kinds of character traits. That is why, in the *Enquiry*, benevolence could be regarded as much of a general principle of morals as sympathy.¹⁶ Hume’s explanations to complement the two theses that make up his moral theory, both earlier in the *Treatise* and later in the *Enquiry*, go as in the following:

(3a) Sympathy is what enables impartial spectators to feel pleasant and unpleasant sentiments of approbation and disapprobation with regard to character traits of agents unrelated to them through a psychological mechanism of communication.

(3b) Sympathy is what enables sentiments to be communicated to each other, but it is benevolence that induces impartial spectators to feel pleasant and unpleasant sentiments of approbation and

---

disapprobation with regard to the same character traits of agents unrelated to them.

So, Hume provides two slightly different explanations, albeit involving similar concepts, to deal with the problem of affective bonding among individuals. This is a fundamental problem for any moral theory based on sentimentalist foundations, such as Hume’s. After all, in this framework, moral evaluations in an intersubjective social perspective presuppose that individuals have some form of access to the sentiments of others. Hume’s claim that individuals tend to be affectively sensitive to those who are similar to them is quite a reasonable one. A kind of capacity that can play the role of enabling us to communicate sentiments and opinions, such as Hume’s sympathy does, is empirically plausible. However, the additional claim that individuals are endowed with a natural benevolent disposition is much more controversial. A sufficiently comprehensive disposition of individuals to regularly approve and disapprove of the same kinds of character traits, regardless of the agents who cultivate them, seems to be a far too distant picture of reality. A simple empirical observation would show that, even under the assumption of ideal conditions of sociability, some individuals seem to show signs of having a benevolent disposition, while several others seem to be quite indifferent to other agents. In addition, the adequate manifestation of moral sentiments with regard to the display of character traits of unrelated agents seems to be very dependent on different individual or social factors. So, even if it is assumed that this disposition exists universally, it does not necessarily follow that it will properly induce adequate moral sentiments. Hume’s most plausible claim is not the one put forward in the Enquiry, his later work. It still the former one in the Treatise that is more convincing. Sympathy bridges the sentimentalist gap between spectators and
other agents in moral evaluations. However, for individuals to be able to sympathize with unrelated agents for the purpose of moral evaluations, it is necessary to understand the conditions upon which an intersubjective social perspective can emerge. And the virtue of justice, which is essential to understand it, will be examined in the next section.

1.5 The Role of the Virtue of Justice

To understand the virtue of justice, it is important to take a look at a specific distinction drawn by Hume in the *Treatise*. In this earlier work, Hume divides his theory of virtues into two kinds of virtues, the natural virtues and the artificial virtues. On the one hand, the natural virtues are defined as a set of common and innate dispositions of human beings that motivate them to behave in a way that is likely to induce a pleasant sentiment of approbation to individuals who are able to recognize them in the stance of impartial spectators. As examples of natural virtues, Hume mentions benevolence, generosity, prudence, temperance, moderation, charity, and so on. On the other hand, the artificial virtues, as opposed to the natural virtues, are defined as the dispositions that convey a pleasant sentiment of approbation to individuals without being motivated by any kind of common and innate disposition of human beings. Artificial virtues are established through rules, inventions, conventions, contrivances, incentives, and other sorts of artifices, to favor the collective interests of society. As examples of artificial virtues, Hume mentions justice, honesty, allegiance, fidelity, truth, chastity, and so on. For Hume, the natural virtues are the character traits commonly regarded as useful or agreeable to the individuals themselves who cultivate them,
whereas the artificial virtues are the character traits useful to others or to society.

In this context, it is reasonable to ask how the different kinds of virtue can be articulated in Hume’s moral theory. For this reason, I intend to suggest an interpretation to solve this issue. My interpretative claim is that, for Hume, the natural virtues are, in a certain way, conceptually dependent on the artificial virtue of justice. So, my interpretative claim could be summarized as such:

(I.C.) The natural virtues are conceptually dependent on the artificial virtue of justice in Hume’s moral theory.

To understand and evaluate the plausibility of this claim, we have to start by examining the textual evidence that supports it. There are two passages that, when understood according to the particular interpretation that I suggest, can pave the way for a better understanding of this articulation. Both in the Treatise and the Enquiry, Hume states, with exactly the same words, the first passage:

The intercourse of sentiments, therefore, in society and conversation, makes us form some general unalterable standard, by which we may approve or disapprove of characters and manners. (T 3.3.3.3SBN /603; EPM 5.42/SBN 229)

Hume’s point with this first passage is to show that “some general unalterable standard” suitable for moral evaluations can be reached in a social context of exchange of sentiments. Now, it is clear that this “general unalterable standard by which we may approve or disapprove characters and manners” refers to the general point of view, since it is this point of view that serves as the proper standard for moral evaluations. This standard of moral evaluation can only emerge in the context of an intersubjective social perspective. And it is only through a social coexistence that the emergence of this standard is
possible. But, for Hume, the coexistence of individuals in society presupposes that they are willing to join themselves together under rules of justice. This point is made by Hume in the *Treatise*, in particular, in the second passage that supports my interpretative claim:

But tho’ it be possible for men to maintain a small uncultivated society without government, ’tis impossible they shou’d maintain a society of any kind without justice, and the observance of those three fundamental laws concerning the stability of possession, its translation by consent, and the performance of promises. (*T* 3.2.8.2/SBN 541)

This passage shows that the very social coexistence upon which the “general unalterable standard” for moral evaluations is based depends on the establishment of a set of rules, the rules of justice. Hume’s thesis is that society can only exist within a framework of rules of justice. As Hume observes, this set of rules consists in: i) the rule of the stability of possessions, ii) the rule of the translation of property by consent, and iii) the rule of the obligation of promises. It is important to point out that, for Hume, rules of justice are established by means of a convention or agreement among individuals. This entails that individuals are not naturally disposed to comply with such rules. Justice is not a natural virtue for Hume. Instead, this set of rules must be first created through a kind of convention or agreement, so that individuals may develop the disposition to behave accordingly. Hume defines the convention of justice as a general “sense of common interest” expressed by each individual restrains their behavior according to certain rules, for they believe that, when they coordinate their actions in this way, it will be conducive to socially advantageous outcomes. In the *Treatise*, Hume understands the convention of justice as it follows:

This convention is not of the nature of a *promise*: For even promises themselves, as we shall see afterwards, arise from human conventions. It is only a general sense of common interest; which sense all the members of the society express to one another,
and which induces them to regulate their conduct by certain rules. I observe, that it will be for my interest to leave another in the possession of his goods, provided he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express’d, and it is known to both, it produces a suitable resolution or behavior. (T 3.2.2.10/SBN 490)

This “sense of common interest” means that individuals, in consulting their interests, manifest to each other a shared perception that the disposition to regulate their behavior by common rules of justice, provided that the others also behave in this way, turns out to be socially advantageous. And, for Hume, this disposition to comply with this set of socially advantageous rules is what defines the virtue of justice. Therefore, the very association of individuals in a collective body depends on a kind of convention or agreement for the establishment of a set of rules of justice. And, for Hume, this set of rules of justice reflects the interest of each individual by ensuring socially advantageous outcomes.

At this point, the interpretative claim can be properly comprehended. The concept of the general point of view is implied in the very concept of virtue and vice. And the general point of view can only emerge in the context of the intersubjective social perspective. This intersubjective social perspective requires, among other things, the interaction of many sympathies in a suitable social coexistence. In turn, this social coexistence requires that individuals join themselves together under certain basic rules of social organization. For Hume, the basic rules of social organization are the rules established by the convention of justice. It follows that it is necessary to presuppose the existence of the convention of justice and the virtue of justice, which is a regular disposition of each individual to comply with the pattern of behavior defined by the rules of justice. This is a necessary assumption for
a general point of view to emerge. Thus, moral evaluations of character traits as virtues and vices, in general, depend on the previous existence of the convention and the virtue of justice.

Although a degree of dependence of the natural virtues on the artificial virtue of justice can be ascertained, there are two points that still demand clarification. First, it is necessary to clarify in which conceptual sense there is a dependence between the two concepts. Second, it is necessary to explain the possible circularity claim implied in the idea that the virtue of justice is presupposed for the emergence of a general point of view that makes it possible to morally evaluate virtues in general. As for the first point, it is important to recall that the rules of justice enable individuals to join themselves together in a socially advantageous system of actions from which they can expect to benefit. This social structure of interaction, in which the social coexistence among individuals extends far beyond the initial boundaries of each one’s “narrow circle”, causes individuals to correct their naturally partial sentiments by taking an interest in the sentiments of others. They are conceived as partners sharing common interests in society. This change can only be generated by a psychological association of ideas operated by the mechanism of sympathy.

Recall that sympathy, when confined to the “narrow circle”, might operate to reinforce partial or biased sentiments of individuals. But sympathy is not particularly conducive to either partial or impartial subjective responses. It is only a mechanism for communicating sentiments and opinions. In a social structure of interaction that affords socially advantageous outcomes, individuals are increasingly more prone to change their sympathetic responses towards others. And, by taking into account the sentiments of others in deliberations, individuals develop and enrich their
own sentimental framework. In contrast to the operation of sympathy confined only to the “narrow circle”, Hume sometimes refers to this correction as “an extensive sympathy”, on which the “sentiments of virtue depend” (T 3.3.1.23/SBN 586). As a result, the intercourse of sentiments in society provides an intersubjective social perspective that enables a general point of view to be adopted for moral evaluations. In this way, individuals who adopt the impartial stance of the general point of view are able to recognize as virtues character traits of agents, whoever they may be, that are useful and agreeable to themselves, because, at this time, they are capable of being sensitive to the sentiments of others. This is the thesis of the progress of sentiments in Hume’s moral theory, as I understand it. It comes down to the idea that individuals tend to increasingly expand their reciprocal subjective responses as the social coexistence also widens. The thesis of the progress of sentiments is featured in the *Enquiry* as such:

Suppose that several families unite together into one society, which is totally disjoined from all others, the rules, which preserve peace and order, enlarge themselves to the utmost extent in society; but becoming then entirely useless, lose their force when carried one step farther. But again suppose, that several distinct societies maintain a kind of intercourse for mutual convenience and advantage, the boundaries of justice still grow larger, in proportion to the largeness of men’s views, and the force of their mutual connexions. History, experience, reason sufficiently instruct us in this natural progress of human sentiments, and in the gradual enlargement of our regards to justice, in proportion as we become acquainted with the extensive utility of that virtue. (*EPM* 3.21/SBN 192)

The nature of the conceptual dependence of the natural virtues on the artificial virtue of justice lies in the fact that the impartial stance of the general point of view, implied in the very concept of virtue and vice in general, depends on an intersubjective social perspective. The social coexistence upon which this perspective is based can only exist under rules of justice. The exchange of sentiments among individuals in a society is what makes this intersubjective
social perspective possible. In this way, extensive sympathy, which operates as to render individuals sensitive to the sentiments of unrelated agents is only possible in the context of a socially advantageous social structure of interaction. Therefore, this particular kind of social interaction and the process of getting acquainted with the sentiments of others provides individuals with an intersubjective social perspective. And the impartial stance of the general point of view that comes with it changes the otherwise biased perception of individuals of the virtuous and vicious character traits in an ever-increasing comprehensive system of moral evaluation. This interpretation strikes as even more plausible in face of the following passage, in which Hume suggests that a moral sense “follows upon” justice and injustice:

We partake of their uneasiness by sympathy; and as every thing, which gives uneasiness in human action, upon the general survey, is call’d Vice, and whatever produces satisfaction, in the same manner, is denominated Virtue; this is the reason why the sense of moral good and evil follows upon justice and injustice. And tho’ this sense, in the present case, be deriv’d to extend it even to our own actions. The general rule reaches beyond those instances, from which it arose; while at the same time we naturally sympathize with others in the sentiments they entertain of us. (T 3.2.2.24/SBN 499)

This passage is part of a broader context regarding the moral evaluation of just and unjust behavior in society. For Hume, it is through sympathetic responses that individuals come to morally approve and disapprove just and unjust character traits. But the point is that, in this part of the passage, Hume extends the range of morally evaluable objects to include “every thing” which produces “uneasiness” and “satisfaction” upon “the general survey”, in order to point out that character traits, in general, are morally approved and disapproved in a similar way. For Hume, this similar process of moral evaluation proves that a distinctive moral sense “follows upon”, succeeds, or
results from justice and injustice. And, as Hume argues, there is a natural tendency to sympathize with whom we interact, which, alongside further generalizations, allows for moral evaluations in general.

Still, with regard to the relation between this conceptual dependence and the thesis of the extensive sympathy, it is important to understand the implications of this interpretative claim to the performance of moral evaluations. When I claim that natural virtues are conceptually dependent on the artificial virtue of justice, I do not mean that individuals are completely unable to recognize and evaluate character traits that can be characterized as natural virtues in the absence of the virtue of justice. In some cases, in particular, when moral evaluations concern individuals of the agent’s “narrow circle”, such as relatives or acquaintances, moral evaluations could, to some degree, be successfully performed, for the agent would be able to sympathize with them. Isolated cases of success in the performance of moral evaluation are possible. What this interpretative claim states is that, in the absence of the virtue of justice, individuals lack the resources to fully form an appropriate sense of virtue, which is the disposition of individuals to behave and recognize as virtuous or vicious certain character traits of agents, whoever they may be. This is because, in the absence of an intersubjective social perspective, that comes in the context of a social order, they would be unable to fully conceive themselves in the position of a spectator that adopts the general point of view for the purpose of moral evaluations.

As for the second point, the interpretative claim that the natural virtues are conceptually dependent on the artificial virtue of justice could be understood as entailing a circular reasoning, insofar as the artificial virtue of justice is presupposed for the moral evaluations of virtues and vices in general. The interpretative claim does not entail circularity for the following
reason. For Hume, a social coexistence among individuals is only possible under rules of justice. The convention of justice is the agreement that establishes the set of rules of justice and, thus, defines the pattern of behavior of compliance with this set of rules that may be properly referred to as the virtue of justice. However, it should be pointed out that the convention of justice is established on grounds of self-interest. On similar grounds is the maintenance of the pattern of behavior of compliance with this set of rules. They are only established in view of possible socially advantageous outcomes. For Hume, they are entirely devoid of any moral value at this point. It is only afterwards that a sympathetic response to the public interest is produced in impartial spectators. This behavioral pattern is regarded as a virtue only after its consolidation as a regular practice, as Hume claims in the *Treatise*:

Thus, self-interest is the original motive to the establishment of justice: but a sympathy with public interest is the source of the moral approbation, which attends that virtue. *(T 3.2.2.24/SBN 500-501)*

Here, it is necessary to distinguish between the establishment of the rules of justice and the moral approbation of the virtue of justice. Since an extensive sympathy with the public interest is necessary for the identification of this particular behavioral pattern as the virtue of justice, then a social coexistence based on the pattern of behavior of compliance with the rules of justice must be presupposed. But this does not imply that this kind of behavior is regarded as a virtue yet. At least, this is definitely not the case for Hume. The moral approbation of the virtue of justice comes only with the sympathy with the public interest, which presupposes that individuals are sufficiently sensitive
to others’ sentiments in a social coexistence under rules of justice. So, strictly speaking, there is no circularity implied in the interpretative claim.

The thesis that virtues and vices are, in general, dependent on artificial and conventional elements is not a new one among Hume scholars. Some of them have already hinted that the emergence of the general point of view requires artifices and conventions, but they differ about the implications of such a thesis for Hume’s moral theory. The controversy lies in the nature of this dependence and in the role played by the artifices and conventions in the recognition of virtues and vices from a general point of view. John L. Mackie suggests that the very general point of view should be characterized as an artifice. It is a social construction that operates as an intersubjective and impartial standard of a system of moral evaluation of virtues and vices in society. In this way, Mackie claims that the general point of view is quite similar to the artificial virtues, in that both are designed to promote social cooperation, mutual aid, and peaceful coexistence. Therefore, the identification of character traits as virtues depends on the artifice of the general point of view. Annette Baier also agrees with Mackie that artifices are needed for the recognition of virtues and vices from the general point of view, although she does not seem to agree with the strong identification of natural and artificial virtues that Mackie’s interpretation ultimately seems to entail. William Davie stresses that the relationship between the general point of view and linguistic conventions. For Davie, the general point of view of moral evaluations is a kind of artificial device similar to linguistic conventions. Lorraine Besser-Jones argues that the role of justice in moral evaluations is to redirect the otherwise misdirected indirect passion of pride through the
development of a sentiment of concern with one’s reputation in a social context established by the convention of justice.\textsuperscript{17}

I argue that the textual evidence in Hume allows me to infer that the interpretative claim that virtues and vices are dependent on artifices and conventions is correct. But my interpretative claim differs from those of other scholars with regard to the nature of this dependence. Not only I claim that there is a conceptual dependence of natural virtues on artifices and conventions, but I also point out that this dependence is due to the artificial virtue of justice. It is a conceptual dependence because, for Hume, the concept of virtue and vice is defined in terms of moral evaluations of character traits that involve the arousal of the suitable subjective responses in impartial spectators. And both the standard for moral evaluation and the suitable subjective responses can only arise in a society established by a set of socially advantageous rules. This dependence is due to the artificial virtue of justice, as it requires not only that this set of rules of justice be established by a convention, but also that there is a pattern of behavior of compliance with the rules that preserves it. This pattern of behavior is precisely the virtue of justice. My interpretative claim is very closely related to that of Jacqueline Taylor. Her interpretation also stresses the importance of the socially advantageous nature of the convention of justice in the establishment of society. She also credits this particular structure of interaction with the redirection of the partial affections towards an extensive sympathy. And, finally, she also claims that the general point of view, which is the standard for moral evaluations in general, is a result of this extensive sympathy.\textsuperscript{18}

Taylor’s interpretation captures the intuition of my interpretative claim almost entirely. The only point I would reinforce is that the nature of the relationship between the sentiments of moral evaluation and the convention of justice is a conceptual one. This stronger claim is justified by the fact that both a convention and a regular practice of justice are necessary conditions for the very existence of a general point of view, which is a concept entailed in the very definition of virtue and vice. But this point could be read into Taylor’s interpretation as well.

This interpretative claim is important in repositioning the virtue of justice to the focal point in Hume’s moral theory. It is true that Hume has a legitimate concern for the virtue of justice. This can be inferred from the fact that both in the *Treatise* and the *Enquiry* Hume begins his account of the natural virtues just after an extensive presentation of the artificial virtue of justice. However, a stronger conceptual relation between both parts of his moral theory had not been established yet. From an interpretative perspective, this claim entails two consequences. First, the role of the virtue of justice in Hume’s moral theory is clarified in a more comprehensive and systematic level. Second, it draws attention to the fact that the analysis of the nature of the virtue of justice is a priority for any possible interpretation of Hume’s moral theory. According to the interpretative claim, the pattern of behavior of compliance with the rules the justice by most of the individuals leads to a psychological development of each one through an enlargement of the scope of sympathy towards others and all passions influenced by this capacity. Ultimately, this sentimental transformation allows the emergence of a standard of moral evaluation in an intersubjective social perspective. Thus, a correct interpretation of the virtue of justice leads to a better understanding of Hume’s moral theory, given the wide influence of this virtue in Hume’s
theory of social psychology and, by implication, in his theory of moral evaluation. And it is exactly with the analysis of the nature of the virtue of justice that what I will be concerned in the course of this study.

Notice that this interpretative claim articulating the natural virtues and the artificial virtue of justice presupposes that Hume’s moral theory can be divided into natural and artificial virtues in the first place. In fact, even the supposed methodological distinction that I have assumed between Hume’s theory of natural virtues and his theory of justice and government would also fail in this case. This concern is justified by the fact that, in the *Enquiry*, Hume drops the distinction between natural and artificial virtues altogether. There is not one single mention of the expression “artificial virtues” in his later work. Instead, Hume characterizes the virtue of justice as “social virtue”, just as the virtue of benevolence. Presumably, Hume thought that he had overcome the old controversy between nature and artifice, or, in other words, *physis* and *nomos*, by his naturalistic approach to moral philosophy. In this way, Hume is led to conclude that there is no substantial distinction between natural and artificial virtues when they are conceived in a broader naturalistic perspective.\(^{19}\) After all, in a broader naturalistic perspective, it is in the very nature of individuals to create artifices to advance their collective purposes. Another evidence for this concern is Mackie’s interpretation of Hume’s moral theory. Mackie thinks that Hume fails to distinguish the natural virtues from the artificial virtues. Both virtues overlap in Hume’s moral theory. This entails that, according to Mackie’s interpretation, Hume’s natural virtues are a subset of the set of artificial virtues. Ultimately, Hume would be committed to the thesis that the entire morality is an artificial construct.\(^{20}\) However, I

---

\(^{19}\) See Haakonsen (1981: 23).

think that the distinction between natural and artificial virtues is still relevant. More specifically, I think that there are very relevant differences among the virtues which allow them to be distinguished, at least in one sense, in natural and artificial ones.

When Mackie claims that Hume’s theory of virtues eventually leads him to commit himself to the thesis that all virtues are artificial, he is right in one sense, but wrong in another. It is clear that virtues can only be identified as such in a context of social coexistence under rules of justice. In other words, a social structure of interaction must be presupposed. Thus, in a sense, Mackie is right when he claims that virtues and artifices are intrinsically connected. However, in another sense, Mackie is mistaken, for some virtues are quite different from each other in terms of their moral evaluations, essential properties, and motivational source. In the *Enquiry*, Hume exemplifies the distinctions between virtues as such:

The social virtues of humanity and benevolence exert their influence immediately by a direct tendency or instinct, which chiefly keeps in view the simple object, moving the affections, and comprehends not any scheme or system, nor the consequences resulting from the concurrence, imitation, or example of others. […] A generous man cheerfully embraces an opportunity of serving his friend; because he then feels himself under the dominion of the beneficent affections, nor is he ever concerned whether any other person in the universe were ever before actuated by such noble motives, or will ever afterwards prove their influence. […] And as the good, resulting from their benign influence, is in itself complete and entire, it also excites the moral sentiment of approbation, without any reflection on farther consequences, and without any more enlarged views of the concurrence or imitation of the other members in society. […] The case is not the same with the social virtues of justice and fidelity. They are highly useful, or indeed absolutely necessary to the well-being of mankind: but the benefit resulting from them is not the consequence

---

21 A distinction between natural and artificial virtues can be constructed on the basis of two aspects of Hume’s theory of virtues, which are elucidated by the “evaluative interpretation” and the “motivational interpretation” by Ken O’Day. See O’Day, K. (1994: 121-141).
This passage is very instructive in highlighting the particular differences in the moral evaluations, essential properties, and motivational sources of the virtues. For instance, Hume thinks that the virtues of humanity and benevolence induce sentiments of moral approbation by a “direct tendency or instinct”. In such cases, moral evaluations take place “immediately”. This means that the moral evaluation of this set of virtues is explained by common and innate dispositions of human beings to identify a set of character traits as virtuous without the mediation of a specific kind of reflective activity on the consequences of this kind of behavior. This is because of the nature of their essential properties. The “good” which results from this set of virtues is “complete” and “entire”. It stems from every single action of humanity and benevolence, for such virtues are agreeable to individuals themselves and to others and also useful to individuals themselves. This also explains the motivational sources of this set of virtues. They are motivated by common and innate dispositions of individuals that causes them to behave in a way that conveys pleasant sentiments of moral approbation in impartial spectators. This can be illustrated by Hume’s example of the “generous man”, who “under the dominion of the beneficent affections”, always attempts to serve his friends.

The virtues of justice and fidelity are quite different. They induce sentiments of moral approbation only through a specific kind of reflective activity on the consequences of this kind of behavior. This set of virtues should be morally evaluated with reference to the whole scheme or system of actions in which they are inserted and, ultimately, constitute. This is also because of the nature of their essential properties. The “benefit” which results
from this set of virtues stems not from every single action, but from the advantageous consequences of the whole system of actions, for such virtues are useful to others or to society as a whole. This is the reason why justice and fidelity make reference to social rules, while humanity and benevolence do not. Some character traits can only be regarded as virtues when comprehended as coordinated with all actions in a cooperative system designed to promote socially advantageous outcomes. And the moral approbation of such virtues requires a reflection on how this kind of behavior can promote the interests of society. And given that this kind of character traits requires a scheme of coordinated actions guided by rules to be regarded as virtues, it can be concluded that not only their motivational source is not a common and innate disposition of human beings, but, in at least one sense, Hume has to admit that some virtues are artificial. For some virtues, such as the virtue of justice, are the result of the establishment of social rules of behavior, while others are not. So, even in the *Enquiry*, where Hume never mentioned the expression “artificial virtues”, this particular distinction between the several kinds of virtues should be admitted.

1.6 Three Lines of Interpretation
In the last section, I have argued for the interpretative claim that the natural virtues are conceptually dependent on the artificial virtue of justice. As a consequence, any plausible interpretation of Hume’s moral theory must take into account the nature of the virtue of justice as a priority in the context of this kind of study. The analysis of the nature of the virtue of justice should start by spelling out the three main lines of interpretation of Hume’s theory of justice with which I will be concerned in this study. They are: i) the
evolutionary interpretation, ii) the rule-utilitarian interpretation, and iii) the contractarian interpretation.

(1) Evolutionary Interpretation (E.I.) - The evolutionary interpretation is the view that Hume’s purpose with his theory of justice would be to account for the origin and the development dynamics of the social and legal institutions related to justice over time. In other words, Hume would claim that these social and legal institutions should be understood as practices that gradually grow and evolve as natural responses of mankind in the light of the given empirical conditions. This interpretation of Hume is supported by philosophers such as Brian Skyrms, F. A. Hayek, and Knud Haakonssen.22

(2) Rule-Utilitarian Interpretation (RU.I.) - The rule-utilitarian interpretation is the view that Hume’s theory of justice would be best described if the social and legal institutions related to justice are understood as a set of social rules designed to maximize the overall sum of happiness or welfare of society. In this sense, these social and legal institutions would maximize the sum of happiness or welfare of society, if each individual, in similar circumstances, performs actions in accordance with this set of institutional rules. This interpretation of Hume is the most accepted one among moral philosophers in general. In fact, many of them have characterized Hume’s moral theory as one form or another of utilitarianism.23 But I intend to restrict the analysis

to the rule-utilitarian interpretation of Hume, supported by philosophers such as Alistair Macleod, Jonathan Harrison, and John Rawls.24

(3) Contractarian Interpretation (C.I.) - The contractarian interpretation is the view that Hume’s theory of justice would be best described if they are conceived as resulting from a hypothetical agreement. The suggestion is that the social and legal institutions related to justice, insofar as they favor the interests of each individual concerned and assure that no one of them ends worse off, could be thought of as earning unanimous assent among everyone. This implies that this set of rules could be considered as morally justified in order to regulate the structure of mutual interaction among each and every individual. This interpretation of Hume was started by David Gauthier and is also supported by philosophers such as Robert Sugden, and Frederick Whelan.25

Each of these lines of interpretation emphasizes a particular perspective adopted by Hume to explain the virtue of justice in his theory. (E.I.) reflects a purely causal or descriptive perspective adopted by Hume in the course of his exposition. (RU.I.) reflects the perspective of the global advantages afforded by the scheme of justice itself, which is sometimes addressed by Hume. (C.I.) reflects the perspective of the individual choice, for Hume sometimes frames the discussion about justice by the rational choices that individuals would make given the circumstances in which they find themselves.

It is worth pointing out that, although these three main lines of interpretation are fundamentally inconsistent with each other, each one of them makes some plausible points and important contributions for an accurate reading of Hume. This must be taken into account in the context of this analysis. For this reason, my purpose is to show precisely at which points these interpretations succeed or fail to correctly account for Hume’s theory of justice. As a result, I intend to assess which one of them is the most plausible interpretation of Hume. It is by surveying these interpretations that I intend to carry out the analysis of Hume’s theory of justice starting from the next chapter.
CHAPTER 2

The Background Conditions for the Convention of Justice

2.1 The Background Conditions for the Convention of Justice

2.1.1 Instrumental Rationality

The interpretative analysis of Hume’s moral and political philosophy should start with his theory of justice or, as he also puts it, the origins of justice. And the examination of Hume’s theory of justice should take into account some of his assumptions that explain his approach of justice as a kind of convention. These assumptions, which are, per se, already important elements of Hume’s moral theory, are three: i) the instrumental conception of practical reason, ii) the artificial nature of the virtue of justice, and iii) the circumstances of justice. As I intend to argue in this chapter, a systematic interpretation of these assumptions will set up a suitable basis for a game-theoretic analysis of Hume’s convention of justice in terms of an agreement arising from a problem of social interaction and interdependent rational decisions. Thus, it is necessary to begin by introducing and explaining the first assumption, namely, Hume’s conception of practical reason. It is in the Treatise Book 2, Part 3, Section 3 that Hume develops his instrumental conception of practical...
reason. Hume’s famous assertion of the thesis of instrumental rationality goes as such:

Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them. (T 2.3.3.3/SBN 415)

It is important to point out two nuances between the standard thesis of instrumental rationality and Hume’s particular conception of it. Instrumental rationality is a thesis about the nature of practical rationality. It states that practical reason should be understood as instructing us to adopt the suitable means to satisfy our given ends. And just as any thesis about the nature of practical rationality, the thesis of instrumental rationality entails the normative requirement that we should adopt the suitable means to satisfy our given ends as a matter of rationality. When individuals violate the instrumental norm, they can be deemed as irrational in a practical sense. Now, of course, this normative requirement can be understood in stronger or weaker ways. According to this conception, rationality is a matter of degree related to the abilities of individuals to form a coherent system of desires, predict and calculate the probability of the different means to bring about the given desires, and make decisions accordingly. All this is implied in the instrumental norm. Hume’s instrumental conception of practical reason seems to imply a rather weaker normative requirement, because such elements were not fully developed by Hume in his account. In addition, according to this thesis, reason is unable to define or evaluate, through mere a priori reflection, the ultimate ends that individuals may have. Rather, the ends that individuals may have are determined solely by their subjective attitudes, such as passions or desires, but not by reason. In Hume’s instrumental conception of practical reason, the ends are determined by passions and desires. And desires are a kind of passions for Hume, namely,
direct passions (T 2.1.1.4/SBN 277). Therefore, for Hume, instrumental rationality relates the suitable means to our given passions.

But how does Hume actually support this instrumental conception of practical reason? The argument that Hume puts forward in favor of the instrumental rationality involves addressing the problem of practical motivation. Hume attempts to specify the different functions performed by reason and passions in the context of practical deliberation. Hume characterizes reason as “the discovery of truth or falsehood” (T 3.1.1.9/SBN 458). This means that the function of reason is to operate so that we can figure out the standards of truth and falsehood. But it is evident that the standards of truth and falsehood can be valid only for the kind of objects that can be true or false in the first place. Beliefs, which, according to Hume, are kinds of ideas, are cognitive states that can be true or false, while passions, which are reflective impressions, are conative states that can be neither true nor false. Then, Hume claims that the truth or falsity of our beliefs is unable to move our will towards action if the content of these beliefs is wholly indifferent to us. It is necessary that the content of these beliefs be somehow related to some kind of passion that we actually hold. The conclusion is that reason, the capacity of inquiring into the truth or falsity of our beliefs, cannot alone motivate us to action. Reason is not, only by itself, capable of being practical. Hume states this point in these terms:

But 'tis evident in this case, that the impulse arises not from reason, but is only directed by it. 'Tis from the prospect of pain or pleasure that the aversion or propensity arises towards any object: And these emotions extend themselves to the causes and effects of that object, as they are pointed out to us by reason and experience. It can never in least concern us to know, that such objects are causes, and such others are effects, if both causes and effects be indifferent to us. Where the objects themselves do not affect us, their connexion can never give them any
influence; and ‘tis plain, that reason is nothing but the discovery of this connexion, it cannot be by its means that objects are able to affect us. (T 2.3.3.3/SBN 414)

However, as this statement stands, it is still not completely clear the grounds on which Hume thinks that reason is inert in terms of motivation. It is necessary to show in what consists this particular motivational quality that reason lacks and passions possess. For Hume, the crucial point here is that, on the one hand, beliefs have “representative quality”, while, passions, on the other hand, do not have any “representative quality”. The most distinctive feature of this “representative quality” is that it renders certain mental states fit to represent certain objects. In Hume’s terms, these mental states are “copies” of the objects they purport to represent. The distinction is illustrated by Hume as such:

A passion is an original existence, or, if you will, modification of existence, and contains not any representative quality, which renders it a copy of any other existence or modification. When I am angry, I am actually possess’d with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high. ‘Tis impossible, therefore, that this passion can be oppos’d by, or be contradictory to truth and reason; since this contradiction consists in the disagreement of ideas, consider’d as copies, with those objects, which they represent. (T 2.3.3.5/SBN 415)

In order to better understand what Hume means with the expression “representative quality”, the concept of “direction of fit” might be helpful. The concept of “direction of fit” is meant to explain the distinction between the mental states of beliefs and desires by making explicit their different functional roles. Elizabeth Anscombe first introduces this distinction to make explicit the different intentions expressed by each mental state, but it is Michael Smith who uses this concept to specify the different functions of beliefs and desires in the standard Humean theory of motivation.26 So, what

is the difference in the functional roles of beliefs and desires? It seems that individuals intend to correctly describe the states of affairs in the world through their beliefs. They are constantly attempting to adjust their beliefs in order to fit them with the world, to make their propositional content match with reality. As a result, as soon as it is ascertained that a belief does not fit with the world, this belief is taken as false, and it should be abandoned. The performance of this kind of role assigns to beliefs a representative function of reality. In the terms of the concept, beliefs have a mind-to-world direction of fit, in that they aim to fit reality. Thus, to say that beliefs are mental states with representative quality means to say that they can depict the world as state of affairs that can be true. Desires or, in Hume’s terms, passions, on the other hand, are mental states without representative quality. Individuals do not intend to correctly describe the states of affairs in the world through their passions. Rather, passions purport to bring about some state of affairs in the world, to make reality fit with their propositional content. As a result, it follows that, if the propositional content of a passion has not yet been satisfied, it does not necessarily imply that this passion must be abandoned. Instead, it compels individuals to make the world to fit their passions. In the terms of the concept, passions have a world-to-mind direction of fit, in that they aim that reality “fit” them. For this reason, namely, by performing this kind of role, that passions lack a representative quality, but they have this particular motivational quality. They could also be described as goals or ends. And Hume is well aware of this functional difference between both mental states. In Hume’s terms, passions are “original existences”. They are not
meant to be a “copy of any other existence”, while beliefs are “ideas” which are meant to be “copies” of the “objects which they represent”.

Hume takes this argument as a justification to conclude that it is not up to reason to set the ultimate practical purposes of our lives. Reason can only be the guide of our passions, which, in turn, specify the ultimate practical purposes of our lives. Therefore, it can be concluded that Hume supports an instrumental conception of practical reason by means of a motivation argument. It seems to me that this motivation argument, which supports the thesis of the inertial nature of reason, is right. It is not clear, however, whether this argument can, by itself, make a case for the thesis of the instrumental rationality once and for all. Part of the plausibility of this conception of practical rationality is the rather simple and coherent way in which it deals with the problem of practical motivation. However, there is also a long tradition of philosophers who have rejected this conception of practical reason by arguing, in many different ways, that reason can, only by itself, be practical.  

At any rate, what is important to stress here is that Hume’s motivation argument is quite a plausible one, and it provides a challenge to the opponents of the instrumental conception of practical reason to come up with a good answer to the problem of practical motivation. Thus, the thesis of instrumental rationality seems to stand on a higher level in terms of plausibility.

But Hume’s main point in the whole discussion should not be overlooked. It should be borne in mind that, in postulating an instrumental conception of practical reason, Hume has in view its implications for morality. Hume is deeply concerned with rejecting a conception of practical

---

27 See Kant (G 4:401), McDowell (1979: 345-346), Scanlon (2014: 54-55).
reason related to the tradition of moral rationalism. In the Treatise Book 2, Part 3, Section 3, Hume characterizes this conception of practical reason as based on the “pre-eminence of reason over passion”, as it claims that individuals are “only so far virtuous” as they give “preference to reason” against passion and “conform themselves to its dictates”. More generally, according to this conception, “every rational creature” ought to “regulate his actions by reason” over “any motive” that may challenge it. Any motive ought to be “brought to conformity with that superior principle” (T 2.3.3.1/SBN 413). So, in the conception of practical rationality supported by moral rationalists, reason makes substantial demands on the passions or desires of individuals. In fact, it defines the ends one ought to have in a purely rational way, regardless of any subjective attitudes. Now, in the context of Hume’s theory of justice, the implication of this conception of practical rationality is clear. It entails that the reason individuals would have for cultivating a disposition to behave according to the rules of justice would be given by a purely rational norm, regardless of any passions or desires that individuals may actually hold regarding this social practice. Individuals would be bound to comply with the rules of justice due to the demands of reason alone. They would be required to comply with the rules of justice simply because it is the right thing to do. This is something that Hume rejects on grounds of the a priori motivational deficit on the part of reason. But, the rejection of this conception of practical reason demands Hume to be capable of providing an alternative account of how a theory of justice, based on an instrumental
conception of practical reason and a set of subjective attitudes, can be convincing.

2.1.2 Artificiality

Given Hume’s instrumental conception of practical reason, it is then necessary to inquire into the nature of the subjective attitude that disposes individuals to behave fair and honestly, according to rules of justice, such as, respecting private property rules and keeping promises. Thus, Hume’s second assumption consists of an analysis of the subjective attitude that grounds the virtue of justice and the reasons why its artificial nature should be accepted.

In the last chapter, I have already addressed the artificial nature of the virtue of justice, as this behavior is not the result of a common and innate disposition of human beings, but it refers to a set of previously established social rules of behavior. However, in this section, I intend to delve deeper into the reasons that justify this thesis. As a virtue ethicist, Hume claims that, in his moral theory, the set of objects of moral evaluation are mainly character traits. Therefore, Hume’s task comes down to proving that the disposition to behave fair and honestly, according to rules of justice, is not a natural character trait in human psychology.

Since Hume does not address this point in the Enquiry, it is in the Treatise where this discussion takes place. In Treatise Book 3, Part 2, Section 1, Hume attempts to prove the hypothesis of the artificiality of the virtue of justice by an argument structured as a disjunctive syllogism coupled with a kind of inductive reasoning as premise. Hume’s argument frames the question of the nature of the virtue of justice in two hypotheses: either the virtue of justice is a natural one or an artificial one. If it is assumed that the virtue of justice is a natural one, then there must be a natural subjective attitude in
human psychology suitable to ground the virtue of justice. Hume proceeds then to survey and consider four possible alternatives for an allegedly natural subjective attitude suitable to ground the virtue of justice and, by reducing each one of them to absurd, he infers that there is no natural subjective attitude suitable to ground the virtue of justice. Yet, there is still a pattern of behavior that can adequately be described as the virtue of justice. Therefore, it can be concluded that the virtue of justice cannot be a natural virtue, but only an artificial one. This is the structure of Hume’s argument in the Treatise Book 3, Part 2, Section 1.

Now, of course, the strength of an argument that resorts to inductive reasoning depend heavily on whether the survey of instances is exhaustive enough for the intended generalizing conclusion. And, unlike the argument developed to support the instrumental conception of practical reason, this one can yield only a probable conclusion in favor of the artificiality thesis at best, for it does not make any conceptual point or is based on deductive reasoning. Thus, one could always cast doubt on the plausibility of this kind of argument by arguing that the survey is not exhaustive enough. However, this kind of argument suits Hume’s philosophical approach just fine. As I have pointed out in the first chapter, empirical observations and generalizations are no less legitimate ways to prove philosophical points than conceptual analysis or deductive arguments in the context of Hume’s philosophy.28 And I tend to

---

28 It is probable that, if social anthropology and ethnography were already consolidated parts of science in the eighteenth century, and not only in the late nineteenth and early twentieth centuries, Hume would have at his disposal sufficient empirical evidence to support his conclusion. For instance, he could bring the evidence of the ignorance of some Tongan populations of the concept of promise, which is a fundamental rule of justice for Hume. See Korn and Korn (1983: 447).
agree with him at this point. At any rate, Hume is perfectly consistent with the tenets of his philosophy.

Hume begins this discussion by clarifying the relation between actions, mental qualities, and character traits. He claims that the moral value of actions is derived from the moral value of the mental qualities that cause them (T 3.2.1.4/SBN 478). In order to ascertain the moral value of any given action, it is necessary to search for the mental quality or the motive of the agent’s character that causes the action. Hume proceeds to argue that a virtuous motive cannot, on pains of circularity, be a regard to the virtuosity of the action itself. Instead, a virtuous motive must be some other kind of sentiment (T 3.2.1.4/SBN 478). Hume’s reasoning goes as follows. Any moral action must have an underlying motive that causes it. This motive must be morally evaluated to bestow a moral value to the action. If a moral evaluation that regards the action as virtuous is always subsequent to the motivation of this action, it presupposes a prior moral evaluation of its motive. It follows that the motive must not be a regard to the virtuosity of the subsequent action. For a regard to the virtuosity of an action to exist, it is necessary that the action can be morally evaluated as virtuous, which, in turn, requires another kind of virtuous motive (T 3.2.1.4/SBN 478).

Hume’s rather trivial point here seems to be that no sense of duty or virtuosity with regard to an action can alone be the first moral motive of it. It makes no sense to argue that an action is virtuous simply because we have a particular regard to its virtuosity or a sense of duty to perform it. It is precisely the virtuosity of actions that Hume wants to explain here. It is necessary to point out the moral quality on which the virtuosity of an action or the sense of duty is based. As a presupposition of the very sense of virtuosity or duty with regard to an action, there must be a natural motive, independent of the
very sense of morality, that causes it and grants it with a moral value. Of course, there are individuals motivated solely by a sense of virtuosity or duty with regard to an action to perform it, but this must presuppose a natural motive in the background, which must have already been morally evaluated as virtuous. Hume concludes this point by stating his “undoubted maxim” in the Treatise Book 3, Section 2, Part 1:

In short, it may be establish’d as an undoubted maxim, that no action can be virtuous, or morally good, unless there be in human nature some motive to produce it, distinct from the sense of its morality (T 3.2.1.7/SBN 479)

As this passage shows, this maxim is fundamental to Hume’s moral theory, since it applies to virtues in general. The best way of summarizing this maxim is through Rachel Cohon’s view of it. She interprets Hume as claiming that any virtue, regardless of being natural or artificial, is caused by some non-moral motive, which is distinct from the mere sense of virtue or duty and, when morally approved under careful contemplation, renders actions virtuous.²⁹

It is only after Hume establishes this moral maxim that he proceeds to investigate the nature of the subjective attitude suitable to ground the virtue of justice to ascertain whether it is a natural or an artificial virtue. However, one conclusion can be firmly established. A first hypothesis that suggests that a natural motive of the virtue of justice could be a sense of virtuosity or duty with regard to justice is wrong. It can be concluded that the natural motive that induces individuals to behave according to the rules of justice cannot be

A sense of virtuosity or duty with regard to this kind of behavior (T 3.2.1.9/SBN 479-480).

A second hypothesis would be that a natural motive of justice could be our self-interest or, as Hume puts it, the passion of “self-love”. This natural motive refers to “a concern to our private interest or reputation”. However, Hume dismisses this hypothesis by arguing that the passion of self-love, in its most natural and unrestrained expression, does not induce us to behave fairly and honestly. Quite the opposite, Hume believes that our natural self-love, when involving individuals in a large scale, leads to widespread conflict. Hume states: “But ’tis certain that, self-love, when it acts at its liberty, instead of engaging us to honest actions, is the source of all injustice and violence; nor can a man ever correct those vices, without correcting and restraining the natural movements of that appetite.” (T 3.2.1.10/SBN 480). So, for Hume, it can be concluded that a disposition to behave according to rules of justice cannot be based on the pure and unrestrained self-interest of individuals.

A third hypothesis would be that a natural motive of justice could be a regard to the public interest or, as Hume puts it, a passion of “public benevolence” or “love of mankind”. Hume is skeptical that a regard for the overall well-being of society can be the natural motive of justice. Indeed, some individuals are sensitive to the public interest and might even have this motive in mind when they respect others’ private property or keep their promises in the context of social interactions. But, as Hume points out, there is simply no such widespread and universal natural sense of regard for humanity, merely as such, regardless of all other particular considerations that might affect each individual. Hume characterizes this motive as “too remote” or “too sublime” to have sufficient influence in the “generality of mankind” (T 3.2.1.11-12/SBN 481). It can be concluded that the natural motive that
induces individuals to behave according to the rules of justice cannot be a regard for the public interest.

A fourth hypothesis would be that a natural motive of justice could be a regard for the interests of others or, as Hume also calls, a “private benevolence”. This alternative strikes Hume as even more implausible than the motive of public benevolence, for, although the virtue of justice requires individuals to behave fairly and honestly, there is no widespread natural motive among human beings to care for the interests of all others. Consider, for instance, the case of individuals who are admittedly dishonest or cruel. The virtue of justice would require a fair and honest behavior towards them too, but this disposition would certainly not be induced in individuals by private benevolence. (T 3.2.1.13/SBN 482). So, while the virtue of justice, on the one hand, involves a disposition to comply with rules in every possible case, private benevolence, on the other, is a very variable and unstable motive. Some people may display it, others may not. Typically, it is manifested more strongly in some people and more weakly in others (T 3.2.1.16/SBN 483). For Hume, it is impossible that the virtue of justice could be grounded on a motive of this nature. Thus, for all purposes, it can be concluded that the natural motive of justice cannot be a regard for the interests of others or private benevolence.

At this point, Hume’s argument comes to a conclusion. By surveying four possible alternatives for a natural subjective attitude suitable to induce individuals to behave fairly and honestly in general, the conclusion is that none of them could be considered a natural motive of the virtue of justice. Individuals might well be occasionally, or even frequently, motivated by a sense of duty for justice, natural self-interest, public benevolence, or private benevolence to behave fairly and honestly, but none of these motives are
necessarily related to the practice of justice. They are only contingently related to it. Such motives cannot be ascribed to each individual in all cases of compliance with the rules of justice. None of them can ground the virtue of justice. Therefore, it can be concluded that the virtue of justice, a disposition to behave fairly and honestly, according to a particular set of social rules, is not a common and innate disposition of human beings. Hume has to identify and present a more plausible motive to serve as the proper ground of the virtue of justice. And Hume suggests that the motive of the virtue of justice must be artificially created as a result of human conventions (T 3.2.1.17/SBN 483).

2.1.3 Circumstances of Justice
Hume’s first assumption of instrumental rationality establishes that reason cannot, by itself alone, prompt us to behave according to rules of justice. The virtue of justice depends on the subjective attitudes of individuals, such as their passions and desires. Hume’s second assumption of artificiality guarantees that none of the subjective attitudes that might motivate the virtue of justice is a common and innate disposition of human beings. Individuals are not naturally disposed to behave fair and honestly, according to rules of justice. Now, after considering the assumptions about the rationality and the psychological capacity of individuals in general, it is appropriate to consider the assumption that defines the conditions under which individuals engage in social interaction. Hume’s third assumption refers to the description of the so-called “circumstances of justice”. The circumstances of justice are the set of empirical conditions that specify the kind of social interaction among
individuals under which rules of justice can be possible. Without really providing a precise definition of the circumstances of justice, Hume suggests the relation between the rules of justice and a specific set of empirical conditions as such:

Thus, the rules of equity of justice depend entirely on the particular state and condition in which men are placed, and owe their origin and existence to that utility, which results to the public from their strict and regular observance. (*EPM 3.12/SBN 188*)

In the *Treatise* Book 3, Part 2, Section 2 and in the *Enquiry* Section III, Part I, Hume introduces the three circumstances of justice: i) the “selfishness” or “confined generosity” of human beings; ii) the “moderate scarcity of provisions or goods” available to satisfy the needs and desires of human beings; and iii) the approximate equality in physical and mental abilities among human beings. For Hume, selfishness or confined generosity refers to the psychological propensity of human beings, in general, of loving themselves more than anyone else. This explains their particular partiality

---

30 This definition of the circumstances of justice is loosely based on John Rawls’ take on the subject in *A Theory of Justice* (1971: 126-129). Since Rawls claims that his approach of the circumstances of justice does not essentially differ from Hume’s, and since Hume himself does not provide us with a precise definition of this concept, this definition satisfies our present purposes of clarifying Hume’s own account of it.

31 The concept of public utility, which features in this quote, will be addressed later in this study. For now, it is more important to consider “the particular state and condition in which men are placed”.

32 As a fourth circumstance of justice, Clayton Hubin suggests the condition of interdependence as a requirement that individuals are not entirely self-sufficient, but that they should be dependent on others for several activities (1979: 7). Although Hume is not clear about this particular condition as a circumstance of justice, it could still be inferred from the interpretation of passages in the *Treatise* (*T 3.2.2.2-3/SBN 485*) and in the *Enquiry* (*EPM 3.20/SBN 291*). However, this condition seems to be too general. Not only is interdependence a requirement for a social interaction under rules of justice, but virtually for any kind of social interaction. This condition seems to be almost trivial, which could explain why Hume did not consider it as worth mentioning as a circumstances of justice.
towards themselves and the individuals in their “narrow circle” of relationships, namely, friends and family members. However, Hume does not mean that human beings are utterly devoid of any trace of altruism, benevolence, or concern for the well-being of others by this condition. Rather, Hume wants to emphasize the tendency of the passion of self-love of each individual to make them more affectively concerned with themselves and their acquaintances than with others (T 3.2.2.5-6/SBN 487).

The moderate scarcity of provisions or goods refers to the general state of affairs of the external world in which there is neither an absolute abundance nor an absolute scarcity of resources to satisfy the needs and desires of human beings. Hume argues that there are three different kinds of goods: i) the mental abilities, ii) the physical attributes, and iii) the external possessions acquired by labor and fortune. For Hume, it is only the last kind of goods that this circumstance of justice applies to (T 3.2.2.7/SBN 488). This interpretation of the second circumstance of justice brings some particular problems to Hume’s theory of justice, as I will show in the course of this study.

The approximate equality in physical and intellectual abilities refers to the fact that, while physical and intellectual differences expressed in a wide array of different talents still exist among individuals, there must be a baseline of roughly equal strength and intelligence among them without which a social interaction under rules of justice would not be possible. Therefore, it follows that, when human beings engage with others of such a feeble species in terms of physical and intellectual abilities to the extent that they are unable to physically resist to the advances of human beings or make them “feel the effects of their resentment”, then this circumstance of justice would not apply to this kind of social interaction. Such is the case of the social interaction
between humans and animals, for instance, where is inappropriate to speak of rules of justice. Thus, as Hume points out, a “degree of equality” is assumed in order for human beings to form a “society” with others (EPM 3.18-19/SBN 190-191).

In the Treatise, Hume characterizes the two first circumstances of justice as arising from our “natural temper” and “outward circumstances” (T 3.2.2.5/SBN 486). On the one hand, confined generosity is derived from the “natural temper” or the psychological makeup of human beings, on the other, the moderate scarcity of provisions or goods available for the satisfaction of their needs and desires is an “outward circumstance” or a state of affairs that characterizes how the external world is. As for the third circumstance of justice, it is assumed that it would fall under the category of outward circumstances, however, it is only in later texts, such as the Enquiry and Of the Original Contract (1758), that Hume addresses this condition and, unfortunately, without a truly systematic approach of it as a circumstance of justice. Hume’s characterization inspired John Rawls’ own account of the circumstances of justice and also anticipates his distinction between the subjective and objective circumstances of justice. Of course, Rawls introduces additional conditions, such as the coexistence at the same space and time and the rough similarity in needs and interests, to refine his account of the circumstances of justice. But, as Rawls himself acknowledges, the additional conditions hardly add much more to the analysis than the empirical conditions already considered by Hume.33 Thus, Hume’s description of the

circumstances of justice is still considered the standard account of the general conditions under which problems of justice make sense at all.

Now, the question that can be asked is how Hume is able to single out this specific set of empirical conditions, over many others, as the circumstances under which rules of justice can be possible and justified. In other words, Hume must come up with an argument to prove that this set of empirical conditions are the circumstances of justice. In the Enquiry Section III, Part I, Hume’s argumentative strategy consists in a thought experiment depicting five hypothetical scenarios characterized by the opposite empirical conditions from those above mentioned. The purpose of this thought experiment is to analyze whether individuals could behave according to rules of justice under such conditions.

In the first scenario, Hume assumes that mankind would be provided with such an extensive amount of goods by the natural environment that it would be sufficient to satisfy whatever needs and desires human beings might have. The standard empirical circumstances would be characterized by the overall plenitude of material goods. Unfortunately, Hume is not entirely consistent at this point. He assumes that at least some physical attributes and mental abilities are provided in abundance as well in this scenario, which partially contradicts what he has previously stated in the Treatise. Recall that Hume rules out mental abilities and physical attributes as the kind of goods which the circumstance of justice applies to. And yet, Hume still brings them into the discussion of the circumstances of justice. This inconsistency is implied by Hume’s depiction of the “natural beauty” of human beings as surpassing that of all other things and that “music and contemplation” are their “sole business”, which suggests that they would be endowed with so much intelligence that it would be superfluous to seek for knowledge (EPM
But, aside from this minor inconsistency, we face a hypothetical scenario of plenitude of external goods. And given this scenario of abundance, one would never need to worry about occasional shortages of goods that could compromise one’s own well-being. No labor or engagement in productive activities would be required in order to secure one’s livelihood. One would not need to establish the division and stabilization of goods among individuals through private property rights and related obligations, for resources would be so plentiful that everyone would have their needs and desires promptly satisfied by such a generous nature (T 3.2.2.15/SBN 493-494; EPM 3.3/SBN 183-184). For Hume, the virtue of justice, which is, in this case, expressed by the disposition to respect rules of private property that stabilize possessions among individuals, would not be of any use to individuals.

In the second scenario, Hume assumes that the amount of goods available for the satisfaction of the needs and desires of human beings remains scarce, but each one would be suddenly endowed with such an extended sentiment of generosity towards others that the concern with one’s own interests would not essentially differ from the concern with the interests of others. In this case, the standard empirical circumstances would be characterized by an extended sympathy and a strong mutual interest of each individual in one another’s concerns (EPM 3.6/SBN 184-185). In this second hypothetical situation, there would be no distrust among individuals in societies. Each individual would be able to count on the goodwill of their fellows regarding the satisfaction of their own interests. As Hume points out, social relationships would bear a strong resemblance to family relationships (EPM 3.7/SBN 185). Hume claims that, under such circumstances, there would be no use in establishing promises, contracts, pacts, agreements, or any
other kind of similar obligation, for every individual would already be disposed to assist anyone in need (T 3.2.2.15/SBN 493-494. EPM 3.6/SBN 185). In this scenario of enlarged affection and extended benevolence among individuals, the virtue of justice, which is, in this case, expressed by the disposition to keep promises, contracts, pacts, and agreements, would not of any use to individuals.

The third scenario is represented as the opposite of the first one. Hume assumes that the empirical circumstances would be characterized by such an extreme shortage of goods so that even the basic needs and desires of each individual would not be fully satisfied given the present amount of available provisions. The standard empirical circumstances would be characterized as a state of extreme misery and deprivation (EPM 3.8/SBN 186). Given a scenario of extreme shortage of goods, Hume assumes that our natural instincts of self-love and self-preservation would surpass any consideration for complying with rules of justice. From the perspective of each individual, private property rights and related obligations would be useless for the purpose of distribution and stabilization of possessions, since they would be already so scarce that their possession and enjoyment can barely satisfy the basic needs and desires of each one. Under such conditions, the logic of interaction among individuals would not be that of compliance with rules of justice, but of fierce and violent competition for the remaining goods (EPM 3.8/SBN 186-187). As a result, Hume concludes that the virtue of justice, which is, in this case, the disposition to respect rules of private property, would not be of any use to individuals in this scenario of extreme scarcity of material goods.

The fourth scenario is represented as the opposite of the second one. Hume assumes that each human being is suddenly endowed with sentiments
of animosity and disregard towards each other. Not only would each individual show no signs of concern for others, but they would also be disposed to exploit or harm them whenever they had a chance for it (EPM 3.9/SBN 187). Notice that there is a subtle point here. Hume is not merely arguing that individuals are just predominantly self-interested, as Hobbes, for instance, does. Rather, Hume’s point here is that individuals must be characterized by such animosity and hostility to each other to the point that coexistence becomes, in principle, impossible. So, while Hobbes’ depiction of the interaction of self-interested individuals allows for a cooperative coexistence among them, albeit under the rule of a common authority, Hume’s depiction of the fourth scenario is far more extreme and seems to preclude this alternative entirely. Thus, the standard empirical circumstances would be characterized by a generalized mutual distrust. In this scenario, just as in the third one, individuals would be prompted to act solely according to the principles of self-preservation. Promises, contracts, pacts, and agreements would not be honored or enforced within this community, given the disposition of every individual to behave aggressively towards the others (EPM 3.9/SBN 187). Therefore, in this scenario of widespread distrust among individuals, the virtue of justice, which is, in this case, the disposition to keep promises, contracts, pacts, and agreements would not be of any use to individuals.

The fifth and last scenario is characterized by the interaction of human beings and other kinds of beings that are vastly different from us in terms of physical and intellectual abilities. As I have already pointed out, human beings cannot be expected to interact with feeble creatures, such as animals,
under rules of justice. Since the difference in physical and intellectual abilities is so large, it is only through “compassion” and “kindness” that the terms of this interaction can be established in favor of such creatures (*EPM* 3.18-19/SBN 190-191). The standard empirical circumstances would be characterized by a great imbalance in power and intelligence between the interacting parties. As a result, in this scenario, the stronger and more intelligent beings would have no incentives to interact with the more feeble ones under rules of justice. Rather, they would either be generous with them or exploit them. As Hume stresses, the “intercourse” between the interacting parties “could not be called society”. Therefore, in this scenario of drastic imbalance in power and intelligence between the interacting parties, the virtue of justice, which is, in this case, the disposition to comply with the rules of justice would be of no use for at least the group of stronger or intelligent beings.

As I have pointed out, Hume’s point with this series of thought experiments is to provide an argument to prove how the rules of justice can only be possible under a particular social structure of interaction defined by a specific set of empirical conditions, the circumstances of justice. It should be noticed that the analysis of the circumstances of justice is carried out by Hume from the perspective of the individual choices. It addresses the empirical conditions under which justice can be useful to individuals involved in the first place. Under specific empirical conditions, justice would be of no use for them, which, as a matter of rational choice, would make them dispense with the rules of justice altogether. So, Hume’s argument is intended to show that, given this specific set of empirical conditions, the rules of justice are possible because of the utility that they are able to afford to the individuals involved. Should any of these circumstances change, this set of rules becomes useless
to them and, as a consequence, the whole practice of justice is compromised, as Hume points out:

Reverse, in any considerable circumstance, the condition of men: Produce extreme abundance or extreme necessity: Implant in the human breast perfect moderation and humanity, or perfect rapacious and malice: By rendering justice totally useless, you thereby totally destroy its essence, and suspend its obligation upon mankind (EPM 3.12/SBN 188)

It is a fact that part of any attempt to justify the practice of justice must involve considerations about the utility of this practice to the individuals involved. But Hume goes as far as to argue that justice is entirely grounded on utility. More specifically, in Hume’s words, the public utility is “the sole origin” of justice and “the sole foundation” of its moral value (EPM 3.1/SBN 183). The discussion about how to properly interpret this thesis and its consequences will be addressed later on in the course of this study. But a brief glimpse of this point helps us to understand how Hume’s argument regarding the circumstances of justice is intended to work.

But just making the circumstances of justice explicit does not directly explain how rules of justice can be possible. Nor does it explain how the virtue of justice can be fabricated and cultivated in the minds of each one of them.

In fact, Hume is aware that these empirical conditions could be very inconvenient to the establishment of justice at first sight, as he points out:

For while each person loves himself better than any other single person, and in his love to others bears the great affection to his relations and acquaintances, this must necessarily produce an opposition of passions; which cannot but be dangerous to the new-establish’d union. […] 'Tis however worth while to remark, that this contrariety of passions wou’d be attended with but small danger, did it not concur with a peculiarity in our outward circumstances, which affords it an opportunity of exerting itself. (T 3.2.6-7/SBN 487)

Confined generosity is characterized by the tendency of the passion of self-love or the “selfish affections” to influence behavior more strongly than the
“kind affections”. This self-love, in concurrence with the circumstance of moderate scarcity and easy transition of the possessions of some to others, takes the form of an intense and “insatiable avidity” for acquiring the scarce supply of external goods for oneself, their relatives, and their acquaintances. The concurrence of both circumstances of justice, when involving individuals on a sufficiently large scale, is likely to engage them into conflicts in a way that is “directly destructive of society” (T 3.2.2.12/SBN 491-492).

However, Hume has also noticed that, within the very structure of social interaction defined by the circumstances of justice, individuals are able to realize that they can secure advantageous outcomes by engaging in a collective effort under rules of justice. Under circumstances of confined generosity, moderate scarcity of provisions, and rough equality in physical and intellectual abilities, which are the standard conditions in social interactions, rules of justice prove to be very useful to them, for, by acting in accordance with them, their material conditions are likely to improve. In fact, Hume’s claim is surprisingly strong here. Under the standard conditions of social interaction, which are characterized by the circumstances of justice or, the “medium amidst all these extremes”, rules of justice are “necessary” in society, as Hume points out in the Enquiry:

The common situation in society is a medium amidst all these extremes. We are mutually partial to ourselves, and to our friends; but are capable or learning the advantage resulting from a more equitable conduct. Few enjoyments are given us from the open and liberal hand of nature; but by art, labour, and industry, we can extract them in great abundance. Hence the ideas of property become necessary in all civil society: Hence justice derives its usefulness to the public: And hence alone arises its merit and moral obligation. (EPM 3.13/SBN 188)

This allusion to a kind of necessity, which is particularly unusual in Hume’s philosophy, must be properly interpreted. I will address this particular necessity of the rules of justice in a society in the next chapter. Now, the point
is to understand how individuals could still be able to engage in a collective effort under rules of justice, albeit the seemingly unfavorable empirical conditions for this purpose. So, on the one hand, a moderate scarcity of provisions yields competition regarding the distribution of the available goods for the satisfaction of the needs and desires of each individual. But, the moderate scarcity of provisions is variable as well. The total amount of goods is not fixed in supply. This supply can be increased through the cooperation of individuals. In this way, advantages can be reaped by the participants of this collective effort. And this cooperation is possible when individuals get together under rules of justice. This is what Hume implies in the aforementioned passage. He argues that “few enjoyments” are provided by the “liberal hand of nature”, but “by art, labour, and industry” individuals can produce them in “great abundance”. This is where “ideas of property” and “justice” afford their utility “to the public”.35

On the other hand, a confined generosity produces a tendency of partiality towards oneself with regard to others. Hume points out that the condition of human beings is that of being “mutually partial to ourselves, and to our friends”. In virtue of this confined generosity, individuals may be occasionally disposed to act in hostile ways against others outside their “narrow circle” of relationships. But, despite this partiality, individuals can acknowledge the advantages of a more “equitable” behavior towards others.

35 This point is particularly stressed by David Gauthier. Gauthier points out that Hume’s condition of moderate scarcity of provisions, by itself, is not sufficient for the establishment of rules of justice. For if this scarcity could not be alleviated by a cooperative and collective effort, then justice would have a very limited scope and, probably, would not last as a stable and enduring institution. Thus, this scarcity must be, in some way, relative or variable. Although Hume has not explicitly stated this objective circumstance of justice as variable scarcity of provisions, it can be inferred, from the passage above, that he believed that the total supply of goods could be increased through a joint activity. See Gauthier (1986: 114).
when they regulate their conduct by general rules of justice. Thus, it can be inferred from this passage that Hume thinks that the analysis of the circumstances of justice already suggests the idea of cooperative behavior among individuals under rules of justice with the purposes of attaining advantageous outcomes.

Recall that the purpose here is to provide a systematic analysis of how Hume’s assumptions explain his approach of justice as a kind of convention. The assumptions are the instrumental conception of practical reason, the artificial nature of the virtue of justice, and the circumstances of justice. As I understand it, Hume’s assumptions of instrumental rationality and the inexistence of a natural disposition for the virtue of justice are mainly, though not exclusively, meant to reject many traditional approaches to justice. They exclude rationalistic and theological ethical approaches to the theory of justice. And, in the background of the assumptions of rationality and psychology capacity of individuals, I understand the assumption of the circumstances of justice as framing the convention of justice as a problem of game theory. Some other scholars have also argued in this way.\footnote{See Hardin (2007: 148), Baier (1991: 233; 2010: 36, 112), Mackie (1980: 88-91).} I believe that the perspective adopted by Hume to describe the circumstances of justice reinforces this intuition. The assumption of the circumstances of justice, in particular, the confined generosity and the moderate scarcity of provisions, poses the kind of practical problem to each individual that might give rise to questions of justice. Each one of them has to decide how to behave regarding the acquisition of shares of external goods, given the behavior of others. On the one hand, each individual is inclined to compete for the acquisition of increasing shares of scarce external goods for satisfying their needs and
desires, which is likely to lead them to conflicts. On the other hand, individuals are able to realize that there is also the possibility for the improvement of their material conditions through a collective effort, which is possible by means of a convention establishing a set of rules of justice, in order to organize the distribution of the shares of scarce external goods.

The only problem is to explain how such a convention can be secured among the kind of rational and self-interested individuals described by Hume. Since each one of them intends to make a rational decision regarding the problem of acquiring shares of scarce external goods, this decision should be addressed by the theory of rational choice. And, since this decision must be made in a context of social interaction, given the rational decisions of all other individuals, it must be rational in face of all other rational decisions. Thus, this decision should be addressed by game theory. This is what I would like to argue for in the next section.

2.2 Analysis of the Pre-Conventional Structure of Interaction

2.2.1 The Humean State of Nature

The assumptions define some of the general features of individuals and the conditions under which they engage in social interactions. They define the structural elements of the pre-conventional structure of interaction. Now, before considering the game-theoretic analysis of it, it is necessary to address some further elements that make up its full description. In other words, it is necessary to address the Humean state of nature. This discussion is strongly influenced by Hobbes, for one of Hume’s concerns with the portrayal of the state of nature was clearly Hobbes’ interpretation of it. It is well known that Hume was a reader of Hobbes. In fact, it is safe to say that Hobbes features as one of the most important influences of Hume’s moral and political
philosophy. References to Hobbes often occur throughout Hume’s work. In addition, Hume’s critique of the portrayal of the state of nature both in the *Treatise* and the *Enquiry* is frequently referred to Hobbes’ view of it. Hume believes that the problem with Hobbes’ portrayal of the state of nature is that it does not properly reflect the principles of human nature. It is not the case that individuals are entirely selfish and utterly devoid of any “kind affections”, as Hobbes seems to assume. For Hume, this kind of metaphysics of human nature implies that the state of nature must be portrayed as consisting of a mere collection of individuals, who are conceived as the basic social unit in that state. They are represented, in principle, as solitary beings, whose behaviors are exclusively guided by self-interested reasons. And the result of the social interaction of such individuals is a “perpetual war of all against all”, given each one’s disposition of “untamed selfishness and barbarity”. Hobbes might resort to this “mere philosophical fiction” in order to account for his political theory, but it is implausible that this state “could ever exist” (*T* 3.2.2.5/SBN 486-487; *T* 3.2.2.14/SBN 493; *EPM* 3.15-16/SBN 189-190).

Now, Hume’s reading of Hobbes is certainly not the most accurate one. Not only Hume seems to suggest that Hobbes’ portrayal of the state of nature is a factual description of a past moment in history, but also that the state of perpetual war of all against all that characterizes the Hobbesian state of nature is more the result of the blind selfishness of human beings than anything else. But, more charitable interpretations of Hobbes are possible as well. They understand the state of nature as a pure thought experiment,

37 Further evidences of this influence can be noticed from similarities between the structure of Hume’s *Treatise* and Hobbes’ *Elements of Law* and *Leviathan*. See Russell (2008: 63-65).
without any trace of factual existence. They also understand the state of perpetual war of all against all more as the result of a structure of social interaction that generates distrust among rational and self-interested individuals than of the mere “untamed selfishness” of human beings. Although it is not really the point here to evaluate the plausibility of Hume’s reading of Hobbes, it is still by contrast to Hume’s reading or, rather, misreading of Hobbes that a Humean state of nature can be properly portrayed. And there are two major critiques of the Hobbesian state of nature explored by Hume that I would like to address here.

First, for Hume, the state of nature, which is a scenario that logically precedes a social state ruled by a political authority, must be properly conceived not as being primarily composed of individuals, but of families. In the *Enquiry*, Hume argues that families are the basic social unit in this state, for individuals are “necessarily born” in “family-societies”, where they are first “trained up” to “rules of conduct and behavior”. And this conception of the state of nature is presented by Hume in clear contrast to that of Hobbes:

> Whether such a condition of human nature could ever exist, or if it did, could continue so long as to merit the appellation of a *state*, may be justly doubted. Men are necessarily born in a family-society, at least; and are trained up by their parents to some rule of conduct and behavior. But this must be admitted, that, if such a state of mutual war and violence was ever real, the suspension of all laws of justice, from their absolute inutility, is a necessary and infallible consequence. (*EPM* 3.16/SBN 190)

This point deserves some attention. Unlike in the first paragraphs of the *Treatise* Book 3, Part 2, Section 2, where Hume is particularly ambiguous regarding the term “society” in his account, and which ultimately leads him to some inconsistencies in confusing family units with larger societies, in the *Enquiry*, Hume is careful enough to characterize the kind of social unit prevailing in the state of nature as “family-societies”. This is not an irrelevant
point. Recall the interpretative claim that I have raised in the first chapter. It claims that, for Hume, societies can only exist within a framework of rules of justice. And there is a substantial difference between family societies and large-scale societies. Hume argues that it is within family societies that individuals are first introduced into morality and manners in general, as they are “trained up” to “rules of conduct and behavior”. They are first instructed both in character traits and in rules of family organization, which, in the *Treatise*, Hume calls “rudiments of justice”, in order to suggest that these rules might serve as inspiration for the actual rules of justice (*T* 3.2.2.13/SBN 493). This point explains why Hume believes that individuals are able to be aware of moral distinctions even in the absence of a social coexistence under rules of justice. Since individuals are more strongly disposed to be affectively attached to those of their “narrow circle” of relationships, they are able to engage with them in a moral stance. However, in order for them to be able to engage with unrelated individuals in a moral stance, it is necessary for them to join themselves together under rules of justice. This implies that it is only when a convention of justice is agreed upon by individuals that several family units can be incorporated into a single and comprehensive society organized by rules of justice.

The idea that family units may play a relevant role in the state of nature is not unique to Hume’s account. James Buchanan and John Rawls have also pointed that out. Buchanan claims that when a “natural distribution” emerges in the state of nature, a mutual agreement for the establishment and assignment of rights might be imagined to take place in a *n*-person model, in which groups, subgroups, or coalitions of individuals, such as, for instance, families or tribes, engage in this contracting process. For Buchanan, as long as several separate coalitions gradually come to mutual agreements for the
establishment and assignment of rights, this process tends to converge to a contracting dynamics in a two-person model, but involving many persons.\textsuperscript{38} On the other hand, Rawls argues that in the selection of the two principles of social justice, the contracting parties, who should imagine themselves in the original position and behind a veil of ignorance, are to be conceived as family heads and not as individuals. This is because each contracting party has the interest that each of their family members can also be contemplated by the principles of social justice.\textsuperscript{39} Otherwise, a narrowly individualistic portrayal of the contracting parties, which completely disregards their family ties, might overlook the interests of relevant individuals in the deliberation process for the selection of the two principles of social justice. Such are some examples of how the representation of the state of nature structured by individuals as members of family units provides a more accurate description for the purposes of imagining a maximal comprehensive mutual agreement in the establishment of rules of justice.

Second, Hume’s emphasis differs from that of Hobbes with regard to the reasons that individuals would have to reach a mutual agreement. And these reasons are intrinsically related to the different interpretations of the state of nature. For Hobbes, the state of nature is a condition of “a war of every man against every man”. Under such circumstances, the life of man is “solitary, poor, nasty, brutish, and short”. There are many incentives for individuals to come up with a mutual agreement for the establishment of common rules of social coexistence, such as a “desire of such things” necessary to “commodious living” and a “hope” to obtain them “by their industry”. However, the most pressing incentive is the “fear of death”. For

\textsuperscript{38} See Buchanan (1975: 33-34, 41-42).
\textsuperscript{39} See Rawls (1968: 52, 70).
Hobbes, it is “the continual fear, and danger of violent death”, which is “the worst of all” things, that prompt individuals to seek a peace agreement. Therefore, it should be remarked that Hobbes’ social contract is mainly a peace agreement. Hume does not disagree with Hobbes on this point. For Hume, the state of nature is a “wild”, “primitive”, “wretched”, and “savage” condition in which violence and misery prevail (T 3.2.7.1/SBN 534; RC 16/396; CP 5/504). Under such circumstances, individuals are not able to properly satisfy their basic needs and desires. However, they can still come up with a solution for this problem through a greater “conjunction of forces”, “partition of employment”, and “mutual succour” by forming a society. In the Treatise, Hume presents this possibility as follows:

When every individual person labours a-part, and only for himself, his force is small to execute any considerable work; his labour being employ’d in supplying all his different necessities, he never attains a perfection in any particular art; and as his force and success are not at all times equal, the least failure in either of these particulars must be attended with inevitable ruin and misery. Society provides a remedy for these three inconveniences. By the conjunction of forces, our power is augmented: By the partition of employment, our ability increases: And by mutual succour we are less expos’d to fortune and accidents. ‘Tis by this additional force, ability, and security, that society becomes advantageous (T 3.2.2.3/SBN 485).

Here it is important to reinforce one point. Hume shares Hobbes’ fundamental concern for peace. In arguing for the “strict observance” of the rules of justice, which are established by the convention of justice, Hume is concerned with the “peace and security of human society” (T 3.2.6.1/SBN 526). Given the structural conditions of confined generosity and variable scarcity of provisions, the dynamics of social interaction in the state of nature is characterized by a state of potential conflict due to the distrust and the competition over the scarce supply of external goods. Thus, it is implied that

the convention of justice allows for peaceful social coexistence by departing from this “wild” and “savage” state of nature. But, Hume is also concerned with the fruits of social cooperation. Given the structural condition of variable scarcity of provisions, the dynamics of social interaction could, in principle, be that of a collective effort to increase the overall supply of goods for all, should individuals come to a mutual agreement to do so. Hume alludes to the fact that collective efforts may bring greater social benefits than individual and isolated ones. And a collective effort in a social coexistence under rules of justice provides increasingly greater social benefits as society expands, which, in the state of nature, under a collective effort of limited scale, would not be possible. This is Hume’s point when he argues, in the Treatise, that “society provides a remedy” for the “inconveniences” by the “conjunction of forces”, “partition of employment” or division of labor, and “mutual succour”, which strengthens and improves the “force”, “ability”, and “security” of individuals. Or when he argues, in the Enquiry, that “justice becomes necessary” for individuals to “extract greater enjoyments” by “labour” and “industry”. So, for Hume, individuals are not only seeking for peace through a mutual agreement. They are also aiming at the increase of the social benefits that they could reap from an ever-increasing collective effort through a mutual agreement. This is the reason why Hume’s convention of

41 Here, Hume’s concern is with what is sometimes referred to as the “cooperative surplus”, as Gauthier puts it (1986: 130), or the “social gains” or “benefits” produced by the “collaboration” or “cooperative venture”, as Rawls characterizes it (1971: 4). But, at this point, Hume is not addressing the distributive shares of the fruits of social cooperation yet.

42 Many things could be comprehended under the idea of social benefit or, in Hume’s terms, “enjoyments” or “advantages”. For instance, if one comprehends peace and security as social benefits, Hume is not substantially different from Hobbes. However, I doubt that Hume intended to mean abstract things such as a social benefit, such as peace, security, education, healthcare, etc. Social benefits are supposed to mean very concrete things like the resources or goods that are collectively produced.
justice establishes rules for the organization and distribution of goods and services. Therefore, it should be remarked that Hume’s convention is mainly an agreement on the rules of justice.

It is clear that the terms of the mutual agreement that establishes a society are defined by the conditions in which individuals find themselves in the logically previous social interaction, namely, the social interaction in the state of nature. The interesting point, however, is to notice how differences in the interpretation of the state of nature by each philosopher shape their conception of the mutual agreement. For instance, Hume differs from Hobbes in that, for him, individuals, although still strongly concerned with peace and security, aim at greater social benefits that result from a collective effort under rules of justice. They are able to conceive others as potential cooperators in this collective effort. So, for Hume, the reasons that individuals would have for coming to a mutual agreement in order to depart from the state of nature are twofold: i) the desire for peace and security, and ii) the desire for improvement of their material conditions. The Humean state of nature could be fully described as a social interaction represented by the following features: i) it consists of individuals conceived as members of family units; ii) such individuals are already aware of moral distinctions and manners to some limited extent by the instruction in family units; iii) such individuals are instrumentally rational and devoid of a natural disposition to behave fair and honestly towards others; iv) such individuals are moderately generous and naturally partial towards their families and friends; v) it consists of a scenario of variable scarcity of goods to satisfy the needs and desires of the individuals; and vi) it consists of a dynamics of social interaction of potential conflict due to the distrust and the competition over the scarce supply of goods. Under such circumstances, individuals are willing to come to a mutual agreement in
order to secure peace and security and improve their material conditions. However, the question still stands as to how exactly this mutual agreement can be reached by rational and self-interested individuals.

2.2.2 The Stag Hunt

In a game-theoretic analysis of the pre-conventional structure of interaction, there are two important questions that Hume’s theory of justice must answer. First, why is it rational for individuals to establish the convention of justice? Second, why is it rational for individuals to act according to the specific pattern of behavior defined by the convention of justice? In the context of game theory, these same questions amount to the problem of proving what counts as a solution for a specific non-cooperative game.\textsuperscript{43} The concept of a solution for a specific non-cooperative game is the equilibrium point or a Nash equilibrium, which is the outcome of a game defined by the set of the players’ strategies, such that each maximizes the payoff of the respective player given the maximizing strategies of all other players involved. In other words, in the equilibrium point, each player’s strategy is optimal against those of others, so that no player has a reason to deviate from the adopted strategy.\textsuperscript{44}

It is then unsurprising that game theory has been widely used by moral and political philosophers as a useful tool for studying models of social interactions and coming up with solutions that express the rational choices of the participants. My purpose is to interpret Hume’s pre-conventional structure of interaction, that is, the Humean state of nature, as a game, a scenario characterized by problems of interaction and interdependent decisions, and its participants as players, in order to show, through this analogy, that each

\textsuperscript{43} See Skyrms and Vanderschraaf (1999: 404-405).
\textsuperscript{44} See Nash (1951: 286-287), Skyrms and Vanderschraaf (1999: 405).
individual has a decisive reason not only for agreeing to establish the convention of justice but also for acting according to the pattern of behavior defined by the rules of justice. The point here is not to delve deeply into game theory. I will not present theorems or proofs, I will not represent games in their normal, extensive, or strategic forms, I will not even numerically specify the values of the players’ payoffs. Rather, I would just like to bring some very general concepts of game theory into the discussion in order to illustrate the solution for a game analogous to Hume’s pre-conventional structure of interaction.

In the last section, I have argued that Hume’s pre-conventional structure of interaction, the Humean state of nature, can be correctly represented from the proper interpretation of Hume’s philosophical assumptions. First, instrumental rationality ensures that individuals are concerned with the adequacy between the available means and the satisfaction of their set of individual preferences, or their utility functions. By reflecting on the adequacy of the available means for satisfying their desired goals, it is implied that they are concerned with increasing the degree of satisfaction of their utility function. They are, in the terminology of game theory, rational utility maximizers. Second, artificiality ensures that individuals are not typically endowed with a strong individual preference that induces them to behave fairly or honestly with each other, regardless of any other preferences, as an element of their utility functions. Individuals are endowed with moderate generosity. They are naturally selfish and partial towards themselves. They will only act in favor of others if they expect some reciprocal advantage in return (T 3.2.5.8/SBN 519). Thus, individuals are not naturally concerned with maximizing the degree of satisfaction of a supposed preference to comply with rules of justice in order to sustain a fair and
equitable coexistence with others. This entails that the decision to establish a
convention of justice and act according to the pattern of behavior defined by
its rules do not emerge spontaneously or unreflexively among individuals,
but they must be somehow agreed upon by them. Third, the circumstances of
justice characterized by confined generosity, variable scarcity of provisions,
and rough equality in physical and mental abilities ensure that individuals do
effectively have something to lose by interacting with each other, in that they
can be exploited or deprived of their possessions by others who aim at utility
maximization. But Hume also observes that individuals are sensitive to the
increasing expected advantages that they can reap from a possible collective
effort under rules of justice (T3.2.5.8/SBN 519-520). This state of affairs can
be described as Hume’s pre-conventional structure of interaction, or the
Humean state of nature.

This state of affairs could be comprehended by an analogy with the
classical game theory model of the Stag Hunt. The Stag Hunt is an example
of a non-zero-sum game. This type of game can be explained in contrast to
zero-sum games. Zero-sum games are those in which the sum of the players’
payoffs at the different possible outcomes equals to zero. In these games, the
gain of one player is the loss of the other. Given the players’ diametrically
opposed interests in these cases, zero-sum games are also called games of
pure or total conflict.\textsuperscript{45} In contrast, non-zero-sum games are those in which
the sum of the players’ payoffs at the different possible outcomes is variable,
that is, either the sum of their gains and losses is more than zero or less than
zero. In this situation, the gain or the loss of one player does not necessarily
result in the gain or the loss of the other. This entails that, in non-zero-sum

\textsuperscript{45} See Skyrms and Vanderschraaf (1999: 393).
games, both players could gain from mutual cooperative behavior. For this reason, non-zero-sum games are also called games of \textit{partial conflict or coordination}.\textsuperscript{46}

So how is exactly the Stag Hunt? Jean-Jacques Rousseau is credited as the first one to have formulated a variation of this game in \textit{A Discourse on Inequality} (1755):

In this manner, men may have insensibly acquired some gross idea of mutual undertakings, and of the advantages of fulfilling them: that is, just so far as their present and apparent interest was concerned: for they were perfect strangers to foresight, and were so far from troubling themselves about the distant future, that they hardly thought of the morrow. If a deer was to be taken, everyone saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs. (1992: 45)

So, suppose that two newly met hunters have gathered to hunt. Both of them are anonymous to each other so that they have no information about each other’s past behavior in order to form reliable expectations. Each hunter must individually choose what to hunt without knowing the choice of the other. Each hunter has the choice of hunting a stag or hunting a hare. The chances of an individual hunter successfully hunting a hare are independent of what the other hunters do. But there is no chance for an individual hunter to successfully hunt a stag. A successful stag hunt requires that both hunters cooperate. However, while hunters, when hunting individually, are able to get

\textsuperscript{46} Lewis suggests, based on Thomas Schelling’s work, that games should be better comprehended as instances of three categories: i) games of pure conflict, ii) games of pure coordination, and iii) games of partial conflict or coordination. In games of pure conflict, the players’ interests are completely opposed and the sum of their payoffs is zero. In games of pure coordination, the player’s interests coincide completely and their payoffs have equal value. These two categories set the opposing limits of the scope of possible games, while games of partial conflict or coordination cover the range in between the former categories varying in proportions of conflict and coordination. See Schelling (1960: 83-87) and Lewis (1968: 13-14).
hares by themselves, a stag is much more valuable than a hare. If each one chooses to hunt a hare, this amounts to a pair of competitive strategies, since it means that both attempt to hunt individually in order to get the less risky, but the less valuable reward. It implies that they give up hunting together to try to get the most valuable reward from the hunt. If each one chooses to hunt a stag, this amounts to a pair of cooperative strategies, since it means that both attempt to work together in order to get the most valuable reward from the hunt. If a hunter chooses to hunt cooperatively, while the other chooses to hunt individually or competitively, without providing any assistance to the other, the first one ends up with nothing from the hunt, while the second one has the prospects of securing a less valuable reward. If both hunters decide to hunt individually or competitively, each will have the prospects of securing a less valuable reward. If both hunters decide to hunt cooperatively, each will have the prospects of getting the most valuable reward.

Since in this general description of the game each hunter is a rational utility maximizer and is not really concerned with the other’s interests, it follows that there are four possible outcomes determined by the combination of the player’s strategies, which can be ranked as such in terms of their expected utility function payoffs: i) the best outcome is to hunt cooperatively, ii) the second best is to hunt individually while the other hunts cooperatively, iii) the second worst is to hunt individually while the other hunts in the same way, and iv) the worst is to hunt cooperatively while the other hunts individually.47 What would be then the rational choice for each hunter under

47 Strictly speaking, there is no difference between the utilities afforded by the second best and the second worst outcomes. This particular preference-ranking needs to make further assumptions for this rank order to make sense, such as, a competition for a scarce supply of hares, for instance. If this property is added to the Stag Hunt base game, then we have an “Assurance Game”.
such circumstances? It depends on the beliefs of each hunter about how the other will choose. Unlike in the Prisoner’s Dilemma, in which individual rationality prescribes non-cooperative behavior as a strictly dominant strategy for the players in this game, and, in turn, makes the outcome defined by the pair of competitive strategies of the two players the equilibrium point for this game, in the Stag Hunt, there are two equilibrium points. Both the outcome in which the hunters hunt a stag cooperatively and the outcome in which each hunter hunts hares individually are solutions for the Stag Hunt. The only difference is that while the first outcome affords the riskiest, but also a greater expected utility, the second outcome affords the least risky, but also a lower expected utility. In other words, there is a payoff dominant equilibrium and a risk dominant equilibrium. Thus, if the players are more payoff-oriented, they would choose to hunt a stag, but if they are more risk averse-oriented, they would choose to hunt hares. This is the reason why the outcome may vary depending on the beliefs of the players. Thus, in the Stag Hunt, the conflict between personal risk and mutual benefit leads players to choose either one of the outcomes in equilibrium. However, the outcome in which both hunters hunt a stag is Pareto dominant with respect to the outcome in which both hunters hunt hares. Pareto dominance refers to the property of an outcome to be better than others in making at least one player better off without making any other player worst off. In other words, the outcome in which both hunters hunt a stag affords a greater utility to them than the outcome in which both hunters hunt hares. So, the crucial problem in this game is how to create a mechanism of trust between the hunters capable of changing their expectations and beliefs, so that they can improve their

48 This account is based on the extensive analysis of Brian Skyrms. See Skyrms (2001: 32-33), Skyrms (1996: 1-4).
conditions by departing from the baseline equilibrium to the Pareto superior equilibrium.

Hume’s pre-conventional structure of interaction represents a strategic problem similar to the one depicted by the Stag Hunt. In both accounts, on the one hand, players could expect a far greater utility from the outcome defined by a set of cooperative strategies, which is, in Hume’s particular case, made possible by the establishment and compliance with the rules of justice. And yet, on the other, players might have stronger incentives not to cooperate, because they would bear in mind the risk of exploitation from possible competitive behavior of others in the acquisition of external possessions. However, the outcome defined by a set of competitive strategies cannot be an acceptable result for Hume’s theory of justice, as it would render the convention of justice and the pattern of behavior according to its rules impossible. Thus, the question is how the players can select the Pareto superior outcome for this game along the lines of what Hume does.

### 2.2.3 Solutions for the Stag Hunt

Some solutions have been suggested for the Stag Hunt game. From the perspective of the theory of rational choice, it is addressed the question of which choice self-interested players should rationally make in a Stag Hunt game. On the one hand, if this game is conceived in a static model, then it is implied that it is played only once, that is, a one-shot game. So, which is the rational choice that self-interested players are supposed to make in face of everyone else’s rational choice? Recall that, depending on the beliefs of each individual, the solution to this game may be either an outcome defined by a pair of competitive strategies or an outcome defined by a pair of cooperative strategies. However, given that the outcome in which hunters hunt
cooperatively a stag is the riskier outcome for each player and, given that there is no mechanism of trust established between them to coordinate their actions, it should be assumed that players would have fewer incentives to choose a cooperative strategy in a Stag Hunt one-shot game, as they stand to bear the highest loss in this social interaction. Both of them would rather choose a competitive strategy in order to secure the less risky payoff in a Stag Hunt one-shot game. Thus, the solution for a Stag-Hunt one-shot game could either be a competitive or a cooperative outcome, but it is far more likely that it settles on a competitive outcome than otherwise.

On the other hand, if this game is conceived in a dynamic model, then it is implied that it is played repeatedly, that is, a finitely or infinitely repeated game. If the social interaction described by the Stag Hunt is interpreted as embedded in a series of finitely repeated interactions played over time, then we might expect different strategy choices from the players and different outcomes for the games. This is because, by considering the Stag Hunt game as a round of a more comprehensive and repeated game consisting of identical base games, the players do not choose strategies based on single plays, but on patterns of plays over the time. Their strategy selection comprises a plan composed of several choice strategies for the whole game. This enables the players to adjust their expectations, change their beliefs, and select new strategies for the game. As a result, the outcome in equilibrium that arises in repeated games could be a set of strategies distinct from that which constitutes an equilibrium in the Stag Hunt one-shot base game. This allows us to explain how a cooperative outcome can become a solution for players in this game, as they are able to realize that, by considering their long-run gain prospects,

it is more beneficial to choose a cooperative strategy, provided that the others do the same. For this reason, it is necessary to conceive the Stag Hunt as a repeated one.

However, is the mere introduction of the property of repetition into the Stag Hunt game model enough to ensure a cooperative outcome among players? If it is further assumed that players have complete information regarding the structure of the game and this fact is common knowledge, it is reasonable to suppose that they know that the series of interactions will have a last round. And in this last round of social interaction, we have, once again, a Stag Hunt one-shot game-type of situation, for players will select strategies aware of the fact that there will be no future opportunity to maximize their utility. For this reason, players will choose strategies that maximize their utilities, which, in the case of the last round of a finitely repeated game, will likely to amount to a competitive strategy. In the last round of a finitely repeated game, a cooperative outcome is unlikely.

The problem, however, goes even deeper. In the perspective of the theory of rational choice, given the property of perfect information that characterizes the repeated Stag Hunt, players’ strategy selection is not repeatedly made in every single round of the game, but is defined as a plan for the whole game, for they are interested in their long-run average payoff. This implies that, in the context of this game, players select a strategy only once. In addition, each choice is dependent on the previous choice in the context of the overall strategy choice. Now, as players are capable of predicting the high likelihood of defection in the last round of the game, they infer that, in the round before the last, they should behave in the very same way, for they are facing, once again, a situation in which there will be no future opportunity to maximize their utility by selecting a cooperative
strategy. Therefore, given their utility-maximizing behavior, players will most likely not engage in cooperation in the round before the last either. By now, it is easy to see that this reasoning, when applied to a plan in the context of a finitely repeated game, entails the unwanted conclusion of the high likelihood of generalized defection, in exactly the same way as in the Stag Hunt one-shot base game. This argument, which is based on the reasoning of backward induction, proves that the mere introduction of the property of repetition in the Stag Hunt is not enough to justify the emergence of a cooperative outcome.

The Stag Hunt game model must be supplemented by at least one additional property. It must be either assumed: i) a property of infinite repetition of the same base game in a supergame covering the actual players’ entire lifespan and extending itself through generations, or ii) a property of imperfect information in cases of finitely repeated or iterated games. When either property is introduced, the players either: i) know that there will not be a last round of social interactions with which they should be concerned, or ii) do not know whether there will be a last round of social interactions or not. In either case, a cooperative outcome seems to be permanently on the players’ horizon of expectations, which could strengthen the incentives for the choice of cooperative strategies. This implies that the argument that correlates the possibility of cooperative outcomes and the repeated Stag Hunt game could, in principle, work for either case. It seems then that the selection of a Pareto superior outcome could be possible as long as the Stag Hunt is conceived as an infinitely repeated game, as it would increase the likelihood that players will change their expectations and beliefs in order to choose cooperative strategies.

50 See Skyrms and Vanderschraaf (1999: 430).
strategies for this game. Still, the mere description of the Stag Hunt as an infinitely repeated game is insufficient and overlooks important points. In fact, as it is, cooperative outcomes could be undermined by factors such as the size of the population of players or random variations in their behavior by chance of mistake in the game. Such factors often feature and play a relevant role in social interactions, so they should be taken into account. And, under suitable circumstances, they can contribute to diminishing the incentives for the choice of cooperative strategies. Players may then deviate from the Pareto superior outcome and select competitive strategies in face of the uncertainty with regard to the behavior of others under suitable circumstances. So, it can be concluded that, in the perspective of the theory of rational choice, cooperative outcomes, although possible in principle, are less likely to be selected, as some factors diminish the incentives for cooperative outcomes.

The Stag Hunt, conceived as an infinitely repeated game, bears a resemblance to another kind of approach to game theory. For instance, in the perspective of the evolutionary dynamics, it is addressed the question of which pattern of behavior self-interested players should rationally converge on in an extended Stag Hunt game. This implies that this game must also be conceived in a dynamic model. More specifically, this game must be an infinitely repeated game. However, unlike in the theory of rational choice, in which players choose a strategy only once as a plan for the whole game, in the evolutionary dynamics, players are supposed to make successive rational choices in face of everyone else’s successive rational choices. Ultimately, the successive rational choices of players should make them converge on a pattern of behavior characterized as a stable equilibrium, which is an outcome from which no player has a reason to deviate in extended games. The analysis of the dynamics of social interaction shows that, where all or most hunters
start by hunting hares, all tend to end up hunting hares, for this outcome affords the less risky payoff in a repeated Stag Hunt game. Hunting hares becomes then a stable equilibrium. But, if it is assumed that a sufficient number of players happen to deviate from the competitive strategy by chance or mistake, it is possible that a convergence towards stag hunting can be built over the long run, provided that all or most hunters are able to realize the far greater utility afforded by the stag hunt. In this way, hunting stags becomes a stable equilibrium as well. However, this equilibrium might become unstable if a sufficient number of players happen to deviate back from the cooperative strategy to the competitive strategy. In fact, since the stable equilibrium defined by the outcome in which hunters hunt hares is the least risky, the convergence leans more strongly towards this particular stable equilibrium in random variations in players’ behavior. Therefore, it should be noticed that, in the evolutionary dynamics, suitable random variations in players’ behavior can drastically shift the pattern of behavior from one stable equilibrium to the other and vice versa. This particular dynamic is made possible by the fact that players must make successive rational choices in each round of the game. So, from the perspective of the evolutionary dynamics, the solution for the Stag Hunt game is still, in most cases, the same as that of the perspective of the theory of rational choice.

However, the analysis of the solutions for the Stag Hunt game shows that, either in the perspective of the theory of rational choice or in the perspective of the evolutionary dynamics, a cooperative outcome in which both hunters hunt a stag might be possible, albeit it is a less likely and only temporary outcome, given the factors aforementioned. For a cooperative outcome to be a standard stable solution, it is necessary to supplement this
picture with additional properties that ultimately make the Pareto superior outcome a solution for this game.

First, anonymity between players, as portrayed in Rousseau’s version of the Stag Hunt game, strengthens their risk aversion, as they lack information about how others are likely to behave in the future. The size of the population of players matters in terms of anonymity, which is a factor that diminishes or strengthens the incentives for cooperation. If it is assumed that the Stag Hunt is represented as a game to be played by a large number of players, random encounters between anonymous players are more likely to happen. But, if the Stag Hunt is designed as a game to be played by a small number of players, random encounters become quickly non-existent. Anonymity is then ruled out, as players gradually get to know each other. This makes it possible for players to know how others are likely to behave and to form expectations accordingly. In this way, players will adjust their strategies each turn of this extended game.

Second, if it is assumed that players are able to engage in a social learning process, by which they can become aware of others’ reputation based on the past experience and communicate this information to others, random variations in players’ behavior are not as disruptive as they frequently are. As a result of this social learning process, players are able to identify and distinguish potential cooperators from non-cooperators. And as players will be willing to behave cooperatively only with those who behave in the same manner, non-cooperators will be treated competitively and will be punished with the exclusion of this particular collective effort. In other words, it is a rational choice for stag hunters to interact only with those they identify as stag hunters, while avoiding interaction with those they identify as hare hunters.
This identification is enhanced by collective communication involved in this social learning process.\textsuperscript{51}

When the infinitely repeated Stag Hunt game model, supplemented by these additional properties, is imagined by rational players, it seems that convergence towards the cooperative stable equilibrium presupposes that they choose a particular kind of strategy as a rational choice in this game. This strategy is that of a disposition of conditional cooperation. In other words, players should adopt a trigger, a tit-for-tat, or a conditional strategy regarding cooperative behavior. Or still, players should sanction non-cooperative behavior.\textsuperscript{52} Conditional cooperation is equivalent to a trigger, a tit-for-tat, or a conditional strategy regarding the cooperative behavior of other players. This strategy prescribes that, as a general rule, each player should cooperate, provided that the other players also behave in the same way. If, however, the other players choose not to cooperate, then each one should not cooperate either. Under such conditions, as a general rule, stag hunters interact exclusively with stag hunters, while hare hunters interact exclusively with hare hunters. And, given that the stag hunt yields a far greater utility for each player, it is expected that hare hunters weigh better the consequences of adopting specific strategies, adjust their expectations, change their beliefs, and select new strategies for the game. Ultimately, it is expected that players would join stag hunters in order to reap the greatest expected advantage from the collective effort. And a well-established social learning process prevents

\textsuperscript{51} Strictly speaking, it is not necessary to assume communication between players so that cooperation can succeed. Only anonymity and social learning are required for that. But, it is safe to argue that, provided that a pattern of cooperative behavior emerges, communication frequently takes place and speeds up the process of convergence on a cooperative stable equilibrium. See Sugden (2005: 43).

that the pattern of behavior keeps bouncing back and forth from a stable equilibrium to the other, as players can distinguish cooperators from non-cooperators more accurately based on their reputation, which becomes firmly consolidate from past experience.53

Both the theory of rational choice and evolutionary dynamics have provided clues to understand how stag hunting, the Pareto superior outcome, can become a stable equilibrium. But, notice that the Pareto superior outcome can only be a solution for the Stag Hunt if the game is represented as embodying some properties. Thus, if the Stag Hunt is imagined in this way, there is still a rational strategy that comprises the whole game. First, it must be designed in a dynamic model, as an infinitely repeated game. Second, it must be allowed sufficient random variations in players’ behaviors, so that a pattern of behavior of stag hunting can emerge in the first place. Notice that it is not necessary that this pattern is defined by the behavior of the majority of players in order to stabilize, given further assumptions of anonymity and social learning process. Third, it must be designed as played by a small population of players, so that random encounters between players are excluded and, by implication, the element of anonymity is ruled out. And, fourth, it must include players capable of learning from past experiences in this kind of social interaction. It is through the social learning process that they can tell stag hunters apart from hare hunters and behave accordingly. Once players rationally consider the structure of social interaction, they can, in principle, agree that the strategy to develop a disposition of conditional cooperation is the rational choice in each round of this extended game. And, given that the outcome in which both hunters hunt stags is the Pareto superior

outcome, this particular pattern of behavior tends to stabilize as an equilibrium point, as stag hunters thrive by receiving the greater payoff under such circumstances.

Hume’s pre-conventional structure of interaction bears some structural resemblance to this kind of game. The choice that individuals must face could be framed as such. On the one hand, each individual risks less by avoiding to take part in a convention of justice and following the pattern of behavior defined by its rules in order to attempt to acquire external goods individually in the “wild” and “savage” state of nature, as there is no mechanism that ensures that this convention will be enforced. On the other hand, the convention of justice and the pattern of behavior defined by its rules yield far greater and ever-increasing expected advantages in comparison to the life in the “wild” and “savage” state of nature. The structure of social interaction could be interpreted as infinitely repeated as well, as situations of choice between acts of justice and injustice related to the convention could be expected to arise indefinitely throughout their lives before other individuals. When the social interaction is imagined in such terms by each individual, they
must choose whether the conformity to the convention of justice is rationally preferred to non-conformity or not.\textsuperscript{54}

In this chapter, I introduced three philosophical assumptions of Hume’s moral theory: i) the instrumental conception of practical reason, ii) the artificial nature of the virtue of justice, and iii) the circumstances of justice. I argued that this set of assumptions should be interpreted as setting the circumstances under which individuals would have to agree to establish a set of rules of justice. In other words, they define the structural properties that characterize a pre-conventional structure of interaction. I have argued that Hume’s pre-conventional structure of interaction, the Humean state of nature, could be adequately represented by an analogy with game theory. More specifically, I have argued that Hume’s pre-conventional structure of interaction could be adequately represented as bearing structural similarities to the Stag Hunt game model. Now, it is necessary to examine whether Hume’s theory of justice, namely, his account of the convention of justice and the specific pattern of behavior defined by its rules, could be adequately interpreted as a “solution”, in the terms here developed, to this problem of interdependent decisions arising from this particular structure of social interaction. For this reason, it is necessary to examine, in Hume’s work, the nature of the convention of justice and the virtue of justice. This is what I will address more carefully in the next chapter.
CHAPTER 3

Hume’s Convention of Justice

3.1 The Motive of Justice

In the last chapter, I have attempted to show that Hume’s pre-conventional structure of interaction, the Humean state of nature, bears structural similarities to the game theory model of the Stag Hunt. After the analysis of the conditions under which a cooperative stable equilibrium can emerge in the Stag Hunt game, it is necessary to examine whether Hume’s account of the convention and the virtue of justice corresponds to this solution. Thus, there are some questions that need to be addressed in the analysis of the nature of the convention of justice. For instance, how do individuals come to establish a convention of justice? How do individuals come to act according to the pattern of behavior defined by this convention? What is the nature of this convention? What are the rules of justice defined by it? How should this convention be interpreted in terms of the social advantages that it affords? The answers to such questions require a more careful and detailed analysis of the central concept of Hume’s theory of justice, the convention of justice. This analysis will be carried out with the purpose of understanding Hume’s convention of justice in light of the three main lines of interpretation suggested in the first chapter: i) the evolutionary interpretation, ii) the rule-utilitarian interpretation, and iii) the contractarian interpretation. In the way
Hume’s approach has been so far characterized, the analysis should start by identifying the motives that induce individuals to agree upon a convention of justice and behave accordingly.

Recall that, for Hume, the inquiry into the nature of the virtue of justice is a matter of ascertaining the motive or mental quality that induces individuals to behave fairly and honestly, according to the rules of justice. And, as Hume attempts to prove, there is no natural motive that induces human beings to behave fairly and honestly in general. As a result, since human beings are not naturally disposed to behave according to rules of justice, Hume’s alternative is to ground this virtue on an artificial motive, which results from human conventions. Thus, it must be explained how individuals come into a convention in the first place, so that this motive can be artificially fabricated and cultivated in the mind of human beings. In the Treatise Book 3, Section 2, Part 2, Hume provides a general description of how individuals come to agree upon a convention of justice as follows:

The remedy, then, is not deriv’d from nature, but from artifice; or more properly speaking, nature provides a remedy in the judgment and understanding, for what is irregular and incommodious in the affections. For when men, from their early education in society, have become sensible of the infinite advantages that result from it, and have besides acquir’d a new affection to company and conversation; and when they have observ’d, that the principal disturbance in society arises from those goods, which we call external, and from their looseness and easy transition from one person to another; […] This can be done after no other manner, than by a convention enter’d into by all the members of the society to bestow stability on the possession of those external goods, and leave every one in the peaceable enjoyment of what he may acquire by his fortune and industry. (T 3.2.5.9/SBN 489)

In this passage, Hume restates the thesis that it is the conjunction of the “irregular” and “incommodious” passions and affections of human beings with the looseness and easy transition of scarce external goods from one person to another that constitutes the “principal disturbance in society”.

The
The concurrence of these two circumstances generates distrust and competition over the scarce supply of goods. In fact, for Hume, the unstable passions and affections of human beings are hardly harmful by themselves alone. They become inconvenient and dangerous only in concurrence with the looseness and easy transition of scarce external goods in social interaction, which suggests that Hume considered this external circumstance to be very relevant in his analysis. It is only through a broader reflective activity, or in Hume’s terms, a change resulting from “judgment” and “understanding”, that individuals come to realize that this external circumstance is the primary source of conflict among them and begin to consider the “infinitive advantages” that come with society. This “disturbance” can only be sorted out that “all members of the society” “enter into” a “convention” to introduce the rule to stabilize the possession of external goods among individuals. So, it is through this kind of reflection that they are able to understand the nature of the main disturbance to social coexistence and come up with the solution for it. However, what is most important for my current purposes is in the second part of the passage:

By this means, every one knows what he may safely possess; and the passions are restrain’d in their partial and contradictory motions. Nor is such a restraint contrary to these passions; for if so, it cou’d never be enter’d into, nor maintain’d; but it is only contrary to their heedless and impetuous movement. Instead of departing from our own interest, or from that of our nearest friends, by abstaining from the possessions of others, we cannot better consult both these interests, than by such a convention; because it is by that means that we maintain society, which is so necessary to their well-being and subsistence, as well as to our own. (T 3.2.5.9/SBN 489)

At this point, Hume considers how the passions are relevant in this discussion. There is one natural motive previously considered and dismissed by Hume as a suitable motive for the virtue of justice that is particularly relevant here. This natural motive is the passion of self-love or self-interest. In the last
chapter, I have alluded to the fact that, for Hume, when individuals interact with each other motivated by a pure and unrestrained sense of self-interest, a convention of justice is unlikely to emerge. A social interaction between individuals based on a sense of pure and unrestrained self-interest cannot be expected to yield cooperative outcomes. Instead, it is the source of potential conflict among them. Pure and unrestrained self-interest, in concurrence with moderate scarcity and easy transition of goods from one person to another, takes the form of another passion. This is the passion of “love of gain”, which is an avidity or sense of urgency for acquiring the scarce supply of external goods for oneself, their relatives, and their acquaintances. But Hume also points out that passions need not be conceived only in their pure and natural manifestation. They can be, in his words, “restrain’d in their partial and contradictory motions”. More specifically, the passions to which Hume refers here are the “self-love”, “interested passions”, or “interests” of individuals and that of their “nearest friends”. In this context, they all stand for the same kind of passion. This passion must be “adjusted” so that the collective arrangement of the several different interests may give rise to the convention of justice, as Hume points out in another passage:

[…] So that the laws of justice arise from natural principles in a manner still more oblique and artificial. ‘Tis self-love which is their real origin; and as the self-love of one person is naturally contrary to that of another, these several interested passions are oblig’d to adjust themselves after such a manner as to concur in some system of conduct and behavior. (T 3.2.2.6/SBN 529)

There are three relevant observations that should be made about the role of the passions in Hume’s theory of justice, as it is presented in the aforementioned passages. First, self-interest is the motive for the establishment of the convention of justice and maintenance of the pattern of behavior defined by its rules, namely, the virtue of justice. However, as
Hume points out, a restraint on the “interested passions” of individuals is required, for, in the absence of this control, the convention “cou’d not be enter’d into, nor maintain’d”. At any rate, it should be observed that individuals do not give up their own self-interest or that of persons of their narrow circle of relationships by establishing the convention of justice and complying with the pattern of behavior defined its rules. Rather, as Hume says, individuals could not “better consult both these interests” when a convention of justice is agreed upon and preserved. Hume seems to be arguing here that the very self-interest of individuals can be more efficiently satisfied through the convention and the virtue of justice than otherwise. This is made explicit by Hume when he argues that “self-love” is the “real origin” of “the laws of justice”. Thus, the motive that grounds the convention and the virtue of justice is still the self-interest of individuals.

Second, the motive that grounds the convention and the virtue of justice is neither the pure and unrestrained sense of self-interest nor the absolute restraint on this passion, but the “restrain’d” self-interest in its “partial and contradictory” or “heedless and impetuous” motions. The passion of self-love should neither be left completely free nor be completely suppressed. It needs only to be restrained or controlled so that it does not induce individuals to engage in potential conflicts. This restraint or control should redirect the passion of self-interest towards a pattern of behavior that can best satisfy their interests. Now, the question is how this passion can be restrained or controlled. In the causal and mechanistic view held by Hume in his theory of passions, this restraint involves an alteration in the direction of this passion. And it can only be made by “the very affection itself”, that is, by self-interest, and not by any other passion. This is because a passion can only be counteracted or redirected by another of similar strength, and, for Hume,
none is as strong as the passion of self-love associated with a desire to acquire external goods, or, the “love of gain”. Thus, if the love of gain is to be altered in its direction, it must be through a way in which this very same passion can be more efficiently satisfied. The redirection of this passion can be performed at the “least reflection” on the very nature of the passion of self-interest, given the empirical circumstances in which individuals face themselves in, through their judgment and understanding, which shows new forms of conduct by which self-interest may be more efficiently satisfied. In the Treatise Book 3, Section 2, Part 2, Hume illustrates this process as such:

‘Tis certain, that no affection of the human mind has both a sufficient force, and a proper direction to counter-balance the love of gain, and render men fit members of society, by making them abstain from the possessions of others. [...] There is no passion, therefore, capable of controlling the interested affection, but the very affection itself, by an alteration of its direction. Now this alteration must necessarily take place upon the least reflection; since ‘tis evident, that the passion is much better satisfy’d by its restraint, than by its liberty, and that in preserving society, we make much greater advances in the acquiring of possessions, than in the solitary and forlorn condition, which must follow upon violence and an universal licence. (T 3.2.2.13/SBN 492)

This reflection on the very nature of the passion of self-interest shows that a pure and unrestrained sense of self-interest must be regulated by the establishment of a set of rules, according to which individuals should behave, in order to afford socially advantageous outcomes in general. In particular, of all rules of justice, it is the rule of the stability of possessions that is strongly emphasized here, as it is the problem of the violent acquisition of possessions, the “principal disturbance in society”, that one seeks to settle at this point. Therefore, it is not merely the passion of self-interest that grounds the convention and the virtue of justice, but a restrained and controlled sense of
self-interest, which is redirected towards a more socially advantageous kind of behavior.

Third, restraints on self-interest constitute the artificial nature of justice. When individuals, oriented by this reflection, realize that the pure and unrestrained sense of self-interest, along with the looseness and easy transition of scarce external goods from one person to another, constitute the main cause of potential conflict in society, they understand that this sense of self-interest should be restrained and redirected towards a more beneficial pattern of behavior. For Hume, the alternative that individuals come up with to solve this social inconvenience is to establish a set of rules to regulate their behavior. This is the convention of justice. This set of rules, which should be agreed by each individual upon reflection, operates as restraints on the passion of self-interest. In this way, individuals are able to create an artificial “system of conduct and behavior” in which their “several interested passions” must somehow “adjust themselves” and “concur”, which yields socially advantageous outcomes.

Self-interest, albeit restrained or controlled in its natural motions, is the suitable motive that grounds both the convention and the virtue of justice, as I interpret it. The passion of self-interest, when restrained and redirected by reflection, induces individuals to agree upon a convention of justice and behave accordingly. This thesis is supported by some of the aforementioned passages. The idea that if the passion of self-interest is not somehow restrained, a convention of justice “cou’d never be enter’d into, nor maintain’d”, and that “self-love”, in a “more oblique and artificial” way, is the “real origin” of “the laws of justice”. Therefore, the interpretative claim is that the restrained self-interest is not only a necessary but also a sufficient
condition both for a convention and the virtue of justice. Self-interest grounds the convention and the virtue of justice.

First, individuals undertake a reflective restraint and redirection of the passion of self-interest towards the establishment of a set of rules designed to regulate their behavior, for, in this way, it could foster a mutual coexistence that affords advantageous outcomes. Second, it is cultivated a disposition to act according to this set of rules, a virtue of justice, based solely on this restrained self-interest. This is because, by behaving in this manner, individuals can expect to sustain a pattern of behavior well-suited to “maintain society” and secure “their well-being and subsistence”. In other words, it is the sense of self-interest, albeit restrained, redirected, and regulated by the rules of justice, that prompts individuals to behave according to them, for it is more advantageous for them to act in this way than to behave based on a pure and blind self-interested passion. For Hume, the motive of justice is a restrained self-interested disposition, which induces individuals to establish a set of rules of justice. And this motive is cultivated in the minds of individuals by the steady compliance with the pattern of behavior defined by this set of rules. When the passage in the Treatise Book 3, Section 2, Part 2 is considered as a whole, it is possible to understand how the passion of self-interest can be redirected by a suitable reflection towards the establishment of the rule of the stability of possessions, which is supposed to regulate their behavior and restrain their passions onwards:

The remedy, then, is not deriv’d from nature, but from artifice; or more properly speaking, nature provides a remedy in the judgment and understanding, for what is irregular and incommodious in the affections. For when men, from their early education in society, have become sensible of the infinite advantages that result from it, and have besides acquir’d a new affection to company and conversation; and when they have observ’d, that the principal disturbance in society arises from those goods, which we call external, and from their looseness and easy transition from one person to another; […] This can be done after no other manner,
Given the emphasis on the passion of self-interest in this description, the motive of justice in Hume is sometimes called the “enlightened self-interest”. The thesis of the enlightened self-interest could be spelled out as such: the restrained self-interest of individuals is the motive that grounds both the convention and the virtue of justice.

The interpretative claim that the enlightened self-interest is the suitable motive of justice seems to correspond to the description developed so far. This interpretation takes into account the fact that, for Hume, individuals are rational utility maximizers. They are mainly self-interested and are not particularly concerned with the interests of individuals unrelated

---

55 The interpretation of the enlightened self-interest is supported by scholars such as Baier and Gauthier. See See Baier (1991: 242-243; 2010: 107), Gauthier (1992: 408, 410, 424). There is a relevant interpretative controversy over the nature of the motive of justice in Hume’s theory of justice. Some scholars such as Cohn, Darwall, Baron, and Haakonsen disagree with the interpretative claim that the enlightened self-interest qualifies as the suitable motive of justice. They claim that the restrained self-interest is the motive that grounds the convention of justice, but the virtue of justice can only be grounded by a kind of moral motive. They disagree that the passion of self-interest, no matter how “enlightened” it may be, can alone sustain the pattern of behavior defined by the rules of justice. Clearly, this interpretation casts doubt on the description of Hume’s theory of justice that I have provided and relied on so far. However, I believe that the textual evidence that supports is underwhelmingly weak in comparison to the interpretative controversy here because I believe that it deserves a study on its own. See Cohn (1997: 105-107; 2008: 182), Darwall (1996: 297-298), Baron (1982: 541), Haakonsen (1982: 33-35).
to them. This entails that individuals are inclined to behave in ways that best satisfy their interests. And, after a proper reflection on the possible effects of the self-interest of human beings, along with the empirical circumstances that characterize their social interaction, individuals come to realize that restraints on self-interest could be the most advantageous choice in the context of social interaction. Such restraints come in the form of a set of rules to regulate collective behavior, which allows for social coexistence and ongoing organization of a collective effort.

3.2 The Convention of Justice

3.2.1 Two Definitions of the Convention of Justice

After having established the conclusion that it is a sense of a restrained self-interest that induces individuals to establish a convention of justice and behave according to the pattern of behavior defined by its rules, it is necessary to address the nature of this convention. And a detailed analysis of the definition of the convention of justice in Hume is required for this purpose.

Hume provides two definitions of the convention of justice, one in the Treatise and one in the Enquiry. In the Treatise Book 3, Section 2, Part 2, the definition is formulated as such:

This convention is not of the nature of a promise: For even promises themselves, as we shall see afterwards, arise from human conventions. It is only a general sense of common interest; which sense all the members of the society express to one another, and which induces them to regulate their conduct by certain rules. I observe, that it will be for my interest to leave another in the possession of his goods, provided he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express’d, and is known to both, it produces a suitable resolution and behavior. And this may properly enough be call’d a convention or agreement betwixt us, tho’ without the interposition of a promise; since the actions of each of us have a reference to those of the other, and are perform’d upon the supposition, that something is to be perform’d on the other part. Two men, who pull the oars of a boat, do it by an agreement or convention, tho’ they have never given promises to each other. Nor is the rule concerning the stability of possessions the less deriv’d from human
conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. (T 3.2.2.10/SBN 490)

In the Enquiry Appendix III, the definition is quite similar, but Hume makes an additional point. Hume points out that the convention of justice yields a “general plan or system of action” that “tends to the public utility”. The definition is formulated as such:

It has been asserted by some, that justice arises from Human Conventions, and proceeds from the voluntary choices, consent, or combination of mankind. If by convention be here meant a promise (which is the most usual sense of the word) nothing can be more absurd than this position. The observance of promises is itself one of the most considerable parts of justice, and we are not surely bound to keep our word because we have given our word to keep it. But if by convention be meant a sense of common interest; which sense each man feels in his own breast, which he remarks in his fellows, and which carries him, in concurrence with others, into a general plan or system of actions, which tends to public utility; it must be owned, that, in this sense, justice arises from human conventions. For if it be allowed (what is, indeed evident) that the particular consequences of a particular act of justice may be hurtful to the public as well as to individuals; it follows that every man, in embracing that virtue, must have an eye to the whole plan or system, and must expect the concurrence of his fellows in the same conduct and behavior […] Thus, two men pull the oars of a boat by common convention for common interest, without any promise or contract: thus gold and silver are made the measures of exchange; thus speech and words and language are fixed by human conventions. Whatever is advantageous to two or more persons, if all perform their part; but what loses all advantage if only one perform, can arise from no other principle. There would otherwise be no motive for any one of them to enter into that scheme of conduct. (EPM App. 3.7-8/SBN 306-307)

Both definitions bring some elements that must be unpacked and clarified. In explaining the elements that constitute both definitions, it is possible to assess whether Hume’s convention of justice could be characterized as a solution in the same terms defined by the analogy of the Stag Hunt game model
suggested in the second chapter. Therefore, it is important to address the elements of the definitions right now.

First, Hume starts to formulate both definitions by drawing a distinction between promises and the convention of justice. For Hume, the convention of justice is “not of the nature” of a “promise”. Hume characterizes the convention of justice as a “sense of common interest”, which is the mutual expression of a shared perception of individuals on how they should behave. Thus, it makes sense that, in the first definition, Hume refers to conventions as a synonym of “agreements”, and, in the second definition, Hume argues that conventions derive from “voluntary choices”, “consent”, or a “combination”, as conventions express a kind of convergence between individuals. Conventions, just like promises, entail voluntary agreements that express the willingness of the parts to behave or not to behave in some particular way. However, conventions are not to be confused with promises. There is a difference between both of them. Promises express a willingness to assume an obligation to perform a future action. However, in the second definition, Hume argues that we are not “bound to keep our word” just because “we have given our word to keep it”. In the second chapter, I have pointed out that, for Hume, individuals are not endowed with a natural disposition that motivates them to will an obligation and comply with it. They simply do not naturally recognize the behavior of promising as intelligible or obligatory. Thus, it is not by means of an exchange of promises that a convention of justice can be established. For Hume, the claim that the convention of justice is a kind of promise is fundamentally inconsistent. Instead, to understand how promises come to be intelligible and morally binding, a more general kind of voluntary agreement must be presupposed, a convention of justice. This is what is meant when Hume argues that promises
arise from “human conventions”. Thus, unlike the promises, the convention of justice is not grounded on a supposed prior moral obligation.

Second, in the first definition, Hume argues that it is by means of a convention of justice that each individual expresses to one another the intention to conditionally regulate their behavior according to “certain rules of conduct”. In the second definition, Hume expands on this point by claiming that “each man” is carried “in concurrence with others” to take part in a “general plan or system of actions” defined by a set of rules of justice. And individuals are willing to act according to a set of rules justice for they believe that, by behaving in this way, they can expect socially advantageous outcomes, as the “general plan or system of action” “tends to the public utility”. It follows that the convention of justice should be interpreted in two ways: i) an agreement among individuals intended to establish a set of rules of justice, and ii) an agreement in which a convergence among individuals towards a pattern of behavior of compliance with a set of rules of justice is displayed. More specifically, the convention of justice is a regularity in the cooperative behavior of the participants to afford socially advantageous outcomes. Given that individuals are not willing to bear losses from this social structure of interaction, each one is only disposed to regulate their behavior according to the set of rules of justice, provided that the others do the same. The convention of justice is advantageous to its participants only “if all perform their part”. It would lose “all advantage” if only one individual “performs” the own part, which, as a result, would provide “no motive” for any individual “to enter into that scheme of conduct”. Thus, what is expressed by each individual in the convention of justice is a disposition of conditional cooperation. This is the disposition chosen by them as the rational strategy for this social structure of interaction. This point highlights the conditional
nature of the cooperative behavior that characterizes the standard behavior of
individuals in the convention of justice.

Third, the fact that the convention of justice is established by the
mutual expression of a shared perception of individuals on how they should
behave, it entails that this convention is common knowledge. A situation of
common knowledge is one in which each individual knows that a particular
state of affairs holds and is aware that all others know about it too. To the
extent that, as Hume argues in the first definition, “all the members of the
society express to one another” this intention and they are “sensible of a like
interest”, it is then “known to both” the willingness to cooperate, provided
that the others do the same. In the second definition, Hume makes the same
point by stating that this interest is a “sense each man feels in his own breast,
which he remarks in his fellows”. For Hume, common knowledge is a
relevant feature of the convention of justice, since it generates mutual
expectations among individuals that actions will be “perform’d upon the
supposition, that something is to be perform’d on the other part”.

Fourth, Hume depicts the convention of justice as gradually evolving
over time. According to the first definition, the pattern of behavior defined by
the rules of justice arises “gradually” and consolidates itself “by a slow
progression” from the “repeated experience of the inconveniences of
transgressing it”. This description suggests that, for Hume, the convention of
justice is a kind of agreement that, in order to generate a regular pattern of
compliance behavior, results from a long process of social learning, adaption,
and consolidation of habits over time. Thus, Hume is not only concerned with
clarifying the conditions upon which the convention of justice can be agreed
upon, but also how the pattern of behavior defined by the rules of justice can be consolidated as the standard behavior among individuals.

By now, it should have been already noticed that the definitions of the convention of justice, both in *Treatise* and in the *Enquiry*, are constituted by elements that are similar to the analogy of the Stag Hunt game model suggested in the second chapter. The description of the convention of justice parallels the description of the solution for the Stag Hunt game model in the most fundamental points. First, given the background conditions that characterize the pre-conventional structure of interaction, in order for a cooperative outcome to arise as a solution under such circumstances, a kind of mutual agreement is required. Without such an agreement, it is highly unlikely that a collective effort can ever emerge. Hume is aware of this point. However, since individuals are mainly rational utility maximizers, they are mainly motivated by the passion of self-interest, this agreement should not be thought of as resulting from an exchange of promises, or the making of a pact or contract, or any kind of conduct grounded on a supposed prior moral obligation. Rather, this agreement is a function of the converging interests of individuals, as it is claimed both in the analogy and in Hume’s text.

Second, through this agreement, it is expressed the intention of each individual to regulate their behavior according to a set of rules of justice, provided that the others do the same. In other words, the strategy chosen by individuals as the rational one for this social interaction is that of conditional cooperation. Individuals are willing to cooperate in view of the ever-increasing expected advantages yielded by a “general plan or system of actions” defined by the rules of justice. However, they are not willing to cooperate only for the sake of cooperation. They are willing to cooperate only conditionally. More specifically, they are willing to cooperate only with other
potential cooperators, as the expected advantages of this particular social structure of interaction can only arise from a regular pattern of compliance behavior. This claim, which is quite central to the account, is also shared by both the analogy and Hume’s account.

Third, the very convention of justice allows for this disposition for conditional cooperation to be known by each individual and that it be known that all others know it too. Common knowledge is a feature of the convention of justice, as it requires from its members that they express, in one way or another, the intention to behave in this way. This kind of knowledge strengthens the pattern of behavior defined by the rules of justice by allowing each individual to know what to expect from others. Strictly speaking, this point plays a greater role in Hume’s definition of the convention of justice than in the Stag Hunt game model. The feature of common knowledge of the strategy chosen by the players might accelerate the consolidation of a specific pattern of behavior, as it might naturally arise from the observation of repeated interactions, but it is not a necessary condition for the establishment of the convention of justice, either as the set of rules justice or the pattern of behavior defined by the set of rules.

Fourth, this description addresses the convention of justice from a dynamic standpoint. In fact, according to this account, not only is the pattern of behavior defined by the rules of justice a result of a kind of social interaction extending over time, but also that this kind of social interaction is infinitely extended. This is because, whereas the establishment of the convention of justice might eventually come at a slow and gradual progression, its maintenance is a never-ending process. This infinitely extended social interaction involves a social learning process by which individuals gradually and slowly adjust expectations, converge in behavior,
and reinforce the habit of compliance from the repeated past experience in this dynamic structure. This claim also fits well both the description by analogy and Hume’s description of the convention of justice.

In sum, the convention of justice seems to resemble, in its most fundamental elements, the description that characterizes the solution for the Stag Hunt game model. First, the convention of justice is not grounded on a supposed prior moral obligation. Second, the convention of justice is characterized by a conditional cooperation between the individuals in relation to the compliance with the pattern of behavior defined by a set of rules of justice. Third, the convention of justice is common knowledge among all individuals. And, fourth, the convention of justice should be thought of as developing and consolidating gradually and slowly over time, from a dynamic standpoint. And this convention is a kind of agreement among individuals by which the intention to establish a set of rules of justice is mutually conveyed and by which a convergence towards the pattern of behavior defined by this set of rules is displayed. This set of rules of justice is established by individuals motivated by a restrained sense of self-interest, as they expect to reap more beneficial results by behaving in accordance with this set of rules. This is how the nature of the convention of justice could be properly summarized.

3.2.2 The Rules of the Convention of Justice
In the Treatise Book 3, Section 2, Part 2, most of Hume’s account of the convention and the virtue of justice is focused on the rule of the stability of possessions. This is because, for Hume, the first most pressing problem for individuals in society is the prospect of violent competition for the acquisition of scarce goods. However, the set of rules of justice comprises “three
fundamental laws of justice” for Hume (T 3.2.8.2/SBN 541). Consequently, as a general rule, the logic of the convention of justice is supposed to apply to all other rules as well. And although the content of the set of rules established by the convention of justice has already been previously mentioned, it is worth considering them more carefully here.

Hume’s account of the convention of justice changes slightly from the Treatise to the Enquiry. As Baier points out, in the Treatise, Hume introduces his theory of justice as a set of successive conventions that establishes the rules of justice. The first convention stabilizes the possessions among individuals by establishing a rule of private property rights. The second convention allows for the optimal distribution of possessions among individuals by establishing a rule of transference of private property by consent. The third convention secure mutual agreements involving the future exchange of goods and services among individuals by establishing the rule of the obligation of promises.56 In this way, Hume provides an account of justice that plays an interesting dialectical role. It depicts an image in which the rules of justice are successively introduced to solve individual and collective problems as they arise. The rule of private property rights is introduced to solve the problem of the threat faced by each individual of possible aggression from others and subsequent deprivation of their possessions due to potential conflicts over competition for scarce resources. The rule of transference of private property by consent is introduced to solve the problem of optimal distribution and allocation of resources suitable to satisfy the present needs and demands of individuals in a very primitive and rudimentary market structure after the stabilization of possessions among them. The rule of the

56 See Baier (1991: 243-244).
obligation of promises is introduced to solve the problem of widespread mutual distrust among individuals with regard to exchanges of resources to satisfy the future needs and demands of individuals. In this way, Hume provides a very logical account of the evolution of the rules of justice in terms of the individual and collective problems that they are designed to respond to. However, in the Enquiry, Hume seems to abandon this account. The rules of justice are still said to arise from conventions, but they are not introduced successively to solve individual and collective problems as they arise. Presumably, in the Enquiry, Hume thought that the rules of justice should be introduced at once by one or many conventions that gradually grow hand in hand.\(^{57}\) This is because the nature of the individual and collective problems that individuals face would require the establishment of the entire set of rules of justice beforehand in order for them to be solved.

For Hume, the rule of private property rights, the rule of the translation of private property by consent, and the rule of the obligation of promises are the three “fundamental laws of nature”. This does not entail that these rules are the only rules of justice that exist, but further rules would depend to some extent on a background structure set up by this set of rules. However, is it really plausible to argue that this set of rules constitutes the fundamental laws of justice? Intuitively, it could be argued that Hume overlooks other basic rules of coexistence without which a common life is impossible, such as rules that protect freedom, physical safety, and life. Such rules of justice are the body of rules of criminal law. And why did Hume not consider this set of rules on an equal footing with the other as the fundamental laws of justice? The answer to this question can be answered in some ways. On the one hand,

\(^{57}\) See Harrison (1981: 59).
Rawls interprets Hume’s rules of justice as mainly addressing the distributive and economic aspects of justice. They are thought of as principles to regulate and specify the rules of economic competition and production between individuals. Thus, they are supposed to constitute the most basic rules of economic relations between individuals. On the other hand, Baier supports a more comprehensive interpretation of Hume’s theory of justice. According to Baier, Hume’s virtue of justice is first and foremost about “honesty in property dealings and keeping agreements”. Hume’s example of loan repayment is a typical example of this point, insofar as it involves all three rules of justice simultaneously to allow for the general future exchange of goods and services between individuals. However, Baier remarks that criminal law is also taken into account by Hume. Such rules are also artificial, as they are introduced only later after the establishment of further artifices, such as governments and bodies of magistrates.

I believe that the reason why Hume considers this set of rules the fundamental laws of justice refers to a point addressed in the second chapter. Recall that, in the second chapter, it is pointed out that Hume believes that there are three different kinds of goods: i) the mental abilities, ii) the physical attributes, and iii) the external possessions acquired by labor and fortune. And, for Hume, questions of justice only arise with regard to the last kind of goods. Now, if Hume’s point is true and only external possessions matter in questions of justice, it follows that the set of rules suggested by Hume could be sufficiently described as the fundamental laws of justice since such rules are well-suited to regulate and organize the acquisition and transfer of external possessions. However, is it plausible to assume that only conflicting

claims over the acquisition and transfer of external possessions are able to raise questions of justice? Surely not. By arguing in this way, Hume is guilty of neglecting other sources of conflict that may also enable the rise of rules of justice. Buchanan argues that even if all goods were available in superabundance, conflicts among individuals would still arise.\footnote{See Buchanan (1975: 31).} Hobbes identifies “competition”, “diffidence”, and “glory” as the “three principal causes of quarrel” among individuals.\footnote{See Hobbes (1998: 83).} To the extent that conflicts arising from glory, as Hobbes would argue, break out, claims over external possessions are not necessarily relevant. They may arise from conflicts over mental or physical abilities as well. Under such circumstances, freedom, physical safety, or the life of individuals might be at stake, and some rules would be required as fundamental laws of justice. Thus, as kinds of goods other than external possessions are objects of conflicting claims, additional rules of justice are required to settle conflicts of this nature. And such rules could also be regarded as fundamental laws of justice, just like the ones defined by the convention of justice.

Obviously, Hume is not that naïve to disregard the body of rules that constitute criminal law altogether. In the \textit{History of England} (1778), Hume makes sure to distinguish the administration of civil and criminal justice in society (H 2.19/70). However, it is pretty clear that Hume considered the conflict over external possessions the most important incentive for individuals to agree upon a set of fundamental rules of justice. And Hume’s mistaken assumption about the kind of goods that gives rise to questions of
justice entails this rather restricted and ultimately incorrect conception of the so-called fundamental laws of justice.

3.3 Hume’s Convention of Justice, Evolution, Utilitarianism, and the Social Contract

3.3.1 The Evolutionary Interpretation

This analysis of Hume’s convention of justice was carried out with the purpose of understanding it in a much broader interpretative context. The point was to provide a more careful and detailed description of Hume’s convention of justice in order to understand how it can be accounted for by a comprehensive line of interpretation of Hume’s moral and political philosophy. And this contributes to a more coherent and systematic comprehension of Hume’s theory of justice. Three main lines of interpretation have been suggested to understand Hume’s moral and political philosophy: i) the evolutionary interpretation, ii) the rule-utilitarian interpretation, and iii) the contractarian interpretation. I have suggested that each of the lines of interpretation makes some very plausible points for a correct interpretation of Hume’s theory of justice. I intend to weigh the plausible and implausible points of each interpretation in order to evaluate which one of them is the most plausible to describe Hume’s theory of justice. It is useful to start by evaluating the plausibility of the first line of interpretation, namely, the evolutionary interpretation. This line of interpretation could be formulated as such:

Evolutionary Interpretation (E.I.) - The evolutionary interpretation is the view that Hume’s purpose with his theory of justice would be to account for the origin and the development dynamics of the social and
legal institutions related to justice over time. In other words, Hume would claim that these social and legal institutions should be understood as a practice that gradually grows and evolves as natural responses of mankind in the light of the given empirical conditions.

There is compelling textual evidence to believe that (E.I.) is a suitable interpretation of Hume’s theory of justice. Hume frames the whole discussion of justice in terms of an inquiry into the “origin and nature” of this practice. This interpretation relies heavily on the textual evidence characterized by this kind of a causal or descriptive perspective that Hume develops in his theory of justice. This approach seems to suggest that Hume is committed to an account of the historical origins and the social development of justice over time. Such an account could be better characterized as an empirical description of the evolutionary history of this practice than as a proper philosophical inquiry into it. However, (E.I.) is ambiguous. As it is stated, (E.I.) could be differently interpreted depending on how the first statement is comprehended. In this way, (E.I.) could be understood as formulating two different statements, which I call (E.I.1) and (E.I.2):

(E.I.1) – The evolutionary interpretation is the view that Hume’s purpose with his theory of justice would be to account for the origin and the development dynamics of the social and legal institutions related to justice over time. In other words, Hume’s convention of
justice is supposed to be *a historically accurate description* of the origins and the evolution of this practice.

(E.I.2) – The evolutionary interpretation is the view that Hume’s purpose with his theory of justice would be to account for the origin and the development dynamics of the social and legal institutions related to justice over time. In other words, Hume’s convention of justice is supposed to be *understood in an evolutionary logic*. It is supposed to be thought of as gradually evolving and consolidating itself as a social and legal practice.

Both (E.I.1) and (E.I.2) clearly differ in terms of how the evolutionary interpretation should be comprehended. This difference could be spelled out in the following way. Whereas (E.I.1) commits Hume to the stronger claim that his account of the convention of justice is a historical explanation of how this practice has been established and evolved in societies, (E.I.2) commits Hume to the weaker claim that his account of the convention of justice works rather as a method to understand how this practice is supposed to evolve and consolidate itself in societies. The distinction between (E.I.1) and (E.I.2) is a distinction between a historical account of the evolution of justice and a methodological account of the evolution of justice. But, while one of the formulations is a plausible interpretation of Hume’s theory of justice, the other is surely not. It might be tempting to think that (E.I.1) is the right interpretation of Hume’s theory of justice given the fact that this account is focused on the origins of justice, but there are reasons to doubt it. For if (E.I.1) is taken as a plausible interpretation, Hume’s convention of justice should be comprehended as a historically accurate account of the evolution of justice.
And this entails that Hume’s account of the empirical conditions under which a convention of justice is likely to arise should be historically accurate as well. What is particularly relevant for this discussion is the Humean state of nature. So, according to (E.I.1), Hume is committed to the thesis that the Humean state of nature should be comprehended as a real and historical state in which individuals found themselves before a civil society could be organized and ruled by a political authority. However, textual evidence in Hume does not seem to support (E.I.1). Take, for instance, the following passages. In the Treatise Book 3, Section 2, Part 2, Hume refers to the state of nature as such:

I shall make only one observation before I leave his subject, viz. that tho’ I assert, that in the state of nature, or that imaginary state, which preceded society, there be neither justice nor injustice, yet I assert not, that it was not allowable, in such a state, to violate the property of others. (*T* 3.2.2.28/SBN 501)

And in the Treatise Book 3, Section 2, Part 3, Hume provides an additional description of the conditions of individuals in the state of nature:

To illustrate this, I propose the following instance. I first consider men in their savage and solitary condition; and suppose, that being sensible of the misery of that state, and foreseeing the advantages that wou’d result from society, they seek each other’s company, and make an offer of mutual protection and assistance. I also suppose, that they are endowed with such a sagacity as immediately to perceive, that the chief impediment to this project of society and partnership lies in the avidity and selfishness of their natural temper; to remedy which, they enter into a convention for the stability of possession, and for mutual restraint and forbearance. (*T* 3.2.3.3/SBN 502-503)

And since it is addressed the plausibility of Hume’s theory of justice as a historically accurate account of the evolution of justice, it is worth considering how Hume views the question of the inquiry into the remote origins of societies from his historiographical methodology. In the *History of England*, Hume begins this discussion as such:

The curiosity, entertained by all civilized nations, of enquiring into the exploits and adventures of their ancestors, commonly excites a regret that the history of remote
ages should always be so much involved in obscurity, uncertainty, and contradiction. Ingenious men, possessed of leisure, are apt to push their research beyond the period, in which literary monuments are framed or preserved; without reflecting, that the history of past events is immediately lost or disfigured, when intrusted to memory and oral tradition […] The only certain means, by which nations can indulge their curiosity in researches concerning their remote origin, is to consider the language, manners, and customs of their ancestors, and to compare them with those of the neighbouring nations. The fables, which are commonly employed to supply the place of true history, ought entirely to be disregarded; or if any exception be admitted to this general rule, it can only be in favor of the ancient Grecian fictions, which are so celebrated and so agreeable, that they will ever be the objects of the attention of mankind. (H 1.1/3–4)

From a philosophical perspective, the first two passages show that Hume is not making a case for (E.I.1). In such passages, the state of nature is supposed to be an “imaginary state” and the conditions of individuals in this “imaginary state” are characterized as “assumptions” or “suppositions”, as Hume points out. This indicates a merely hypothetical character of the state of nature. From a historiographical perspective, the last passage casts even more doubt on (E.I.1) as a plausible interpretation of the convention of justice. In this passage, Hume suggests that historical attempts to inquiry into the remote origins of societies should be faced with skepticism due to the “obscurity”, “uncertainty”, and “contradiction” that such periods of history necessarily entail. Whenever “memory” or “oral tradition” is brought as historical evidence to ascertain the historical truth of past events, history is sometimes so “lost” or “disfigured” by this kind of report that they should be disregarded in favor of more reliable sources. By implication, as a matter of historical method, the very idea of inquiring into a supposed state that precedes civil society makes little sense to Hume. Therefore, given textual evidence, (E.I.1)
is surely a wrong interpretation of Hume’s theory of justice, which can be ascertained from both his philosophical and historical perspectives.

Still, Hume’s convention of justice is portrayed in an evolutionary logic. It is said to arise “gradually” and “by degrees” in “a slow progression” from the “repeated experience”. But, given that it does not refer to a historical account of the evolution of justice, it follows that it can only refer to a methodological account of the evolution of justice. In this sense, the most plausible formulation of (E.I.) as a proper interpretation of Hume’s theory of justice is (E.I.2). According to (E.I.2), the state of nature should not be interpreted as a real and historical condition in which individuals found themselves before the emergence of civil society. Instead, the state of nature should be interpreted in the logic of the so-called conjectural history. This historical method, which was very popular during the 18th century Scottish Enlightenment, is an attempt to rationally reconstruct the history of mankind in a speculative way. It is based on a reflection on causes and effects related to the uniform principles of human nature, on the one hand, and the external circumstances in which people are likely to have found themselves in the remote past, on the other. Thus, the point of conjectural history is not much to describe how individuals have actually behaved in the past, but rather to speculate how individuals would likely to have behaved in the past. Conjectural history is then an attempt to describe the evolution of individual behavior in the course of time in a purely methodological way. Since (E.I.2) is the only formulation of (E.I.) that provides a description of the evolution of

---

individual behavior from a methodological perspective, it follows that only (E.I.2) qualifies as the plausible interpretation of Hume’s theory of justice.

However, even if there is textual evidence for (E.I.) as a suitable interpretation of Hume’s theory of justice, this line of interpretation still faces relevant limitations. Evolutionary interpretations usually entail a purely descriptive account of the gradual evolution of the rules of justice. And, in the particular methodological view implied by (E.I.), rules of justice are thought of as gradually evolving over time. But, although (E.I.) provides a speculative account of the evolution of justice in terms of a reflection on individual behavior in the establishment of rules of justice, it does not refer to them as normative and rational responses to the basic social needs and interests of individuals. Ultimately, (E.I.) understands the rules of justice as an imagined result of evolutionary pressures. However, it overlooks an important aspect of Hume’s account, which is precisely the emphasis of the rules of justice as normative and rational responses to the basic social needs and interests of individuals. In the *Enquiry Appendix III*, Hume makes this point very clear by referring to the rules of justice, in particular, the rule of private property rights, as a result of a rational collective response of individuals to their “necessities”, given their “inclinations” and “passions”. And this rational collective response is based on their “understanding” and “experience”. In other words, it describes justice not only in terms of how individuals would have happened to behave, but also in terms of the reasons why they would have behaved in this way:

Men’s inclination, their necessities, lead them to combine; their understanding and experience tell them that this combination is impossible where each governs himself by no rule, and pays no regard to the possessions of others: and from these passions and reflections conjoined, as soon as we observe like passions and reflections in others, the sentiment of justice, throughout all ages, has infallibly and certainly had
place to some degree or other in every individual of the human species. (*EPM App. 3.9/SBN 307*)

As a result, it can be concluded that, whereas (E.I.) makes a very important contribution to the overall understanding of Hume’s theory of justice, as it correctly describes justice as thought of in a logic of gradual evolution, it also falls short to account for a relevant normative aspect that seems to be implied in the general description provided by Hume of the convention of justice.

3.3.2 The Rule-Utilitarian Interpretation

The second line of interpretation of Hume’s theory of justice is the rule-utilitarian interpretation. This line of interpretation could be formulated as such:

Rule-Utilitarian Interpretation (RU.I.) - The rule-utilitarian interpretation is the view that Hume’s theory of justice would be best described if the social and legal institutions related to justice are understood as a set of social rules designed to maximize the overall sum of happiness or welfare of society. In this sense, these social and legal institutions would maximize the sum of happiness or welfare of society, if each individual, in similar circumstances, performs actions in accordance with this set of institutional rules.

With regard to (RU.I.), there is also evidence to believe that it is a suitable interpretation of Hume’s theory of justice. In fact, this line of interpretation has become the most accepted one among scholars to understand Hume’s theory of justice. I suspect that this line of interpretation has become the
standard line of interpretation of Hume’s theory of justice among scholars for the following reasons. First, Hume comes up with some intuitions that seem to foreshadow and pave the way for the 18th and 19th century rule-utilitarian theories. For instance, in one passage of the Enquiry, his later work, Hume refers to his moral theory as a “just calculation” and a “steady preference” for the “greater happiness” (EPM 9.15/SBN 279). Second, the concept of utility or public utility features both in Hume’s theory of justice and in rule-utilitarian theories. Both concepts are very important in Hume’s moral and political philosophy and in rule-utilitarian theories. This might suggest that both concepts are supposed to have the same meaning in both theories. So, I intend to assess the plausibility of (RU.I.) by comparing Hume’s theory of justice and rule-utilitarianism. I intend to start by explaining the ethical theory better known as rule-utilitarianism.

Rule-utilitarianism is a kind of consequentialist moral theory that states that the rightness of actions is to be judged in such a way that the set of social rules that prescribes this kind of action is capable of maximizing the overall sum of happiness or welfare of society. This maximizing principle by which utilitarian theories are typically characterized is called the “principle of utility”. Since the aim of rule-utilitarianism is the maximization of the overall sum of happiness or welfare of society, it is implied that a social welfare function needs to be established. A social welfare function must be constructed by means of a specific type of aggregation of subjective utility functions, which are specified by the particular interests of individuals. In general, the kind of aggregative conception that rule-utilitarianism typically assumes is the combination and coherent arrangement, in one single and

---

63 See Bentham (2011: 13), Mill (2003: 186, 194). Mill also calls it the “greatest happiness principle”.
consistent set, of the particular interests of individuals from the perspective of an impartial point of view. In a rule-utilitarian theory, this kind of impartiality in the aggregation of particular interests for the construction of a single social welfare function requires a measure of interpersonal comparison of utilities, in order to properly evaluate, arrange, and rank the utility functions of all individuals in one single and consistent set. How this measure of interpersonal comparison of utility should be specified is a matter of a long-lasting debate among utilitarians over history. I do not wish to enter into the specific details of the discussion about the correct utilitarian value theory, so I will assume, for the sole purpose of this interpretative work, that the content of the particular interests or preferences of individuals are purely determined by their subjective attitudes and not subjected to any sort of objective moral norm or standard to constrain them, just like in Hume’s description of instrumental rationality. So, it is the impartiality in arranging and ranking individual ranking utility functions, constructed in this way, that properly constitutes the moral value in rule-utilitarian theories. As a result, the social welfare function is defined as the arithmetic mean of the individual utilities. Since rule-utilitarianism requires the maximization of this social welfare function, it amounts to aiming at the greatest satisfaction for the greatest

---


65 See Smart (1973: 12-13). J. J. C. Smart provides a helpful discussion on the kinds of utilitarianism with regard to the nature of the measure of subjective utilities functions, which are to constitute to social welfare function. In this sense, Smart characterizes Bentham as a hedonistic utilitarian, Moore as a ideal utilitarian, and Mill as a quasi-ideal utilitarian.

66 See Harsanyi (1977: 52). This conception is supported by Harsanyi under the suggestive name of principle of consumers’ sovereignty.

number of individuals. It can be concluded, then, that a moral theory is rule-utilitarian when it is concerned with providing a set of social rules that maximize the social welfare function. In other words, it should satisfy the overall sum of particular interests involved. In general, rule-utilitarianism, as a moral theory, is identified by these three characteristics: i) a given value theory, ii) social welfarism, and iii) a set of social rules designed to maximize social welfare. And as far as a rule-utilitarian theory of justice is concerned, what is being asked is whether the social and legal institutions of a society are designed to maximize the overall sum of happiness or welfare in order for them to be qualified as just.\(^68\)

Is (RU.I.) a plausible alternative to describe Hume’s theory of justice? For this hypothesis to be correct, Hume’s convention of justice must be adequately framed in a rule-utilitarian theory of justice. What has traditionally spoken in favor of (RU.I.) was mainly Hume’s own conceptual framework. Given that the concept of utility features in a central position in Hume’s theory of justice, it is tempting to characterize his theory as rule-utilitarian or, at least, related the classical rule-utilitarianism.\(^69\) But, in order for (RU.I.) to be plausible, three points must be proved. First, the concept of utility in Hume must mean the same as it does for utilitarians. Second, Hume must share the same value theory as utilitarians. Third, Hume must be committed to a kind of social welfarism.

With regard to the first point, Hume’s concept of utility fairly coincides with the one adopted by classical utilitarians. For Hume, utility or usefulness is defined as the property of things to have the tendency to satisfy

---

\(^68\) This is the conclusion of classical utilitarians like Mill, for instance, who understand justice as a “particular kind or branch of general utility”. See Mill (2003: 217).

the interests or ends of individuals. In this sense, Hume’s concept of utility is very basic, conceptually associated with other more complex concepts such as, for instance, the concepts of private and public interest. In the Treatise Book 3, Part 3, Section 1, Hume provides a definition of usefulness in the following way:

In this case, the qualities that please me are all consider’d as useful to the person, and as having a tendency to promote his interest and satisfaction. They are only regarded as means to an end, and please me in proportion to their fitness for that end. The end, therefore, must be agreeable to me. (T 3.3.1.25/SBN 588)

In the Enquiry Section V, Hume provides a similar definition of usefulness as the one presented in the Treatise:

Usefulness is only a tendency to a certain end; and it is a contradiction in terms, that anything please as means to an end, where the end itself no wise affects us. (EPM 5.17/SBN 219)

And in the Enquiry Appendix I, Hume suggests that utility and usefulness are synonyms to the extent that they are defined in the same way:

Utility is only a tendency to a certain end; and where the end totally indifferent to us, we should feel the same indifference towards the means. It is requisite a sentiment should here display itself, in order to give a preference to the useful above the pernicious tendencies. […] Here, the therefore reason instructs us in the several tendencies of actions, and humanity makes a distinction in favor of those which are useful and beneficial. (EPM App.1.3/SBN 286)

However, it is important to notice that this definition, by itself, is not committed to any utilitarian claim in particular. This concept of utility is a basic and fairly accepted notion in many contexts. For instance, it features in the economic behavior theory, the rational decision theory, game theory, and even incorporated by moral philosophers in moral theories, among which the

---

70 This definition matches precisely the one put forward by Bentham. See Bentham (2011: 13).
rule-utilitarian ones as well.\textsuperscript{71} In this sense, although Hume’s concept of utility bears a resemblance to the one espoused by classical utilitarians, the usage of this concept by Hume does not necessarily speak for or against (RU.I.) or any other particular interpretation. Now, when it comes to Hume’s convention of justice, it is the concept of public utility that really matters in the discussion. The concept of public utility is built as a logical generalization of the concept of utility. It is defined as the property of things to have the tendency to promote certain interests or ends of society as a whole. In this sense, Hume’s concept of public utility is conceptually associated with his concept of public interest, insofar as certain social interests or ends must be assumed so that a tendency to promote them can exist at all. But, the important point is how to interpret the concept of public utility. In this context, (RU.I.) might still be the most plausible explanatory alternative to accurately describe Hume’s theory of justice, for the principle of utility, which prescribes the maximization of the overall sum of happiness or welfare of society, might still apply to Hume’s concept of public utility. In order to assess whether Hume’s convention of justice makes reference to the maximization of the overall sum of happiness or welfare of society or not, it is necessary to address the third point concerning social welfarism. But, for now, it can be concluded that, at least for the concept of utility in Hume, (RU.I.) is not particularly favored as the most accurate reading over any other possible interpretative alternative.

With regard to the second point, Hume shares with some utilitarians the same value theory based on the purely subjective attitudes, interests, or preferences of individuals. However, even if it is true that Hume sides with

\textsuperscript{71} Many economists, who are generally inclined to support utilitarian theories, take this concept, roughly as formulated by Hume, for granted.
some utilitarians with respect to the value theory, other utilitarians support opposite conceptions. And also non-utilitarians take for granted the value theory based on the subjective attitudes, interests, or preferences of individuals. Therefore, for the like reasons of the first point, this mere coincidence in the value theory does not particularly speak in favor of (RU.I.).

With regard to the third and most important point, Hume’s convention of justice does not seem to correspond to social welfarism in the terms rule-utilitarians support. Social welfarism would require that Hume’s convention of justice is designed to maximize the social welfare function. According to rule-utilitarianism, this social welfare function must be defined as the average of individual utilities arranged by means of an impartial aggregating method, which, in turn, presupposes a possible measure of interpersonal comparison of utilities. The maximization of this social welfare function affords the greatest satisfaction for the greatest number of individuals. Hence, in order for (RU.I.) to be plausible, Hume’s convention of justice must be established exactly for this purpose, namely, to maximize the overall sum of social welfare. In other words, not only Hume’s concept of public utility must correspond to the overall sum of happiness or welfare of society, but also Hume’s convention of justice must reflect the principle of utility. And the social welfare, according to the rule-utilitarians, is the result of a calculus that balances and compensates individual utilities taken together in one single and consistent set, so that the arithmetic mean of utilities can be attained. However, I think that Hume addresses this point from a quite different perspective. In the Enquiry Section III, Hume introduces the relation between justice and public utility for the first time as such:

That Justice is useful to society, and consequently that part of its merit, at least, must arise from that consideration, it would be a superfluous undertaking to prove. That public utility is the sole origin of justice, and that reflections on the beneficial
For Hume, public utility accounts for both the origin and the moral value of justice. So, in principle, it could be correctly stated that Hume’s convention of justice establishes a set of rules designed to maximize social welfare, or, in his own terms, the public utility. However, the convention of justice is not designed to maximize the arithmetic mean of individual utilities, as required by (RU.I.). Hume’s convention of justice is not intended to afford the greatest satisfaction for the greatest number of individuals. It is not intended to reflect the principle of utility. Rather, it is established with the purpose of promoting interests that are shared and common to each individual. In other words, it is established to assure mutual advantage. This is not a necessary requirement for (RU.I.). Recall that, according to (RU.I.), the maximization of the overall sum of happiness or social welfare entails the greatest satisfaction for the greatest number of individuals. In this sense, Hume’s convention of justice reflects a quite different rationale from (RU.I.), as it can be observed from some passages of his works. Take, for instance, the passage in the Treatise Book 3, Part 2, Section 2. Here, Hume presents a definition of the convention of justice in which the “whole system of actions” generated by the set of rules defined by the convention of justice is infinitely advantageous “to the whole” and “to every part”. In the same vein, Hume defines the convention of justice as “a sense of interest” that is “supposed to be common to all”:
induce any one of them to perform an act of justice, who has the first opportunity. This becomes an example to others. And thus justice establishes itself by a kind of convention or agreement; that is, by a sense of interest, suppos’d to be common to all, and where every single act is perform’d in expectation that others are to perform the like. \((T\ 3.2.2.22/SBN\ 497-498)\)

Similarly, in Book 3, Part 2, Section 6 of the \textit{Treatise}, Hume points out that the “system of conduct and behavior” of justice “comprehends” the “interests of each individual”, after a proper “adjustment” of the “self-love” or “interested passions” of each one in a single system of actions:

So that the laws of justice arise from natural principles in a manner still more oblique and artificial. ’Tis self-love which is their real origin; and as the self-love of one person is naturally contrary to that of another, these several interested passions are oblig’d to adjust themselves after such a manner as to concur in some system of conduct and behavior. This system, therefore, comprehending the interests of each individual, is of course advantageous to the public; tho’ it be not intended for that purpose by the inventors. \((T\ 3.2.6.6/SBN\ 529)\)

Such passages do not correspond to (RU.I.). Recall that (RU.I.) requires that individual interests or preferences are taken into account in the construction of a social welfare function, but only indirectly. And this is due to the conception of social welfare that rule-utilitarians claim to be morally right to maximize. The maximization of social welfare, interpreted as a function of the average of individual utilities, makes it necessary to disregard certain particular interests or preferences in favor of the total sum of happiness.\(^72\)

This is precisely the opposite of Hume’s description of the convention of justice as presented in the aforementioned passages. Hume emphasizes that the convention of justice results from a convergence of interests, which are supposed to be shared and common to each individual. This logically entails that the convention of justice yields mutually advantageous outcomes, as “the

\(^{72}\) This is a problem that utilitarians generally recognize. But, unfortunately their answers have not been very convincing. See Smart (1973: 67-73).
system of actions” accommodates the “interests of each individual”. In this sense, this system is advantageous not only “to every part”, but also “to the whole” which amounts precisely to Hume’s concept of public utility. However, (RU.I.) is not an account based on the convergence of interests that are shared and common to each individual. By implication, it is not an account based on mutual advantage either. This is why Hume does not endorse social welfarism, the conception of social welfare supported by rule-utilitarianism. As a conclusion, it can be argued that this line of interpretations is inadequate to accurately describe Hume’s theory of justice, for Hume’s concept of public utility, in the specific context of the convention of justice, is not the same as that of rule-utilitarianism. Similar conclusions are also drawn by Mackie and Haakonssen.73

Lastly, rule-utilitarians might still argue that (RU.I.) could satisfy the condition of mutual advantage required by Hume in the description of the “system of actions” defined by the rules of the convention of justice. Rule-utilitarianism can be formulated as flexible enough to incorporate certain Pareto conditions in order to ensure that no one ends worse off by the establishment of social rules in accordance with the utilitarian calculus. In this way, mutually advantageous outcomes can be attained. For instance, Hardin believes that rule-utilitarianism does not need to be necessarily conceived in an aggregative way. In this sense, it should not necessarily aim at the greatest satisfaction for the greatest number of individuals. In fact, his claim is that the notion of mutual advantage is the logical collective analog of individual self-interest. Rule-utilitarianism, understood as preserving mutual

73 See Haakonssen (1981: 5-6) and Mackie (1980: 151-154). For both of them, whenever Hume refers to the concepts of public utility or public interest, he is not committed to a utilitarian conception.
advantage, could, then, adequately accommodate Hume’s theory of justice, according to Hardin.74

Still, (RU.I.), reformulated in this way, is not a plausible interpretation of Hume’s theory of justice. Even if Pareto conditions are introduced and no individual interests or preferences are disregarded in the process of establishing a set of rules of justice, the reformulated (RU.I.) still does not make reference to any kind of convention or convention in which individuals are conceived as expressing their sense of common interest. (RU.I.), however reformulated it may be, still operates in a rule-utilitarian framework, in which a utility calculus still exists in order to maximize happiness or welfare of society. But, the convention of justice is not designed by its members with the purpose of maximizing social welfare. As evidenced by the passage aforementioned, the advantage “to the public” is “not intended” by “the inventors”. For Hume, public utility is promoted as an unintended outcome of a convention established by individuals with the purpose of attaining mutually advantageous outcomes. What (RU.I.) seems to overlook is that the rule-utilitarian description of Hume’s theory of justice makes individuals extrinsic to the process of establishing rules of justice. If it is assumed that the rules defined by the convention of justice are already specified by a utility calculus, individuals need not be conceived as agreeing to the set of rules of justice so established in light of their specific interests or preferences. Rule-utilitarianism alienates individuals from the establishment of justice. And this is precisely the opposite of Hume’s account of the convention of justice. For Hume, the origins of justice are related to a convention agreed upon by its members in order to satisfy their interests.

Hume’s theory of justice, in particular, his account of the convention of justice, cannot be accurately comprehended under any variation of (RU.I.).

3.3.3 The Contractarian Interpretation

3.3.3.1 Original, Explicit, and Tacit Contractarianism

The third line of interpretation of Hume’s theory of justice is the contractarian interpretation. This line of interpretation could be formulated as such:

Contractarian Interpretation (C.I.) - The contractarian interpretation is the view that Hume’s theory of justice would be best described if it is conceived as resulting from a hypothetical agreement. The suggestion is that the social and legal institutions related to justice, insofar as they favor the interests of each individual concerned and assure that no one of them ends worse off, could be thought of as earning unanimous assent among everyone. This implies that this set of rules could be considered as morally justified in order to regulate the structure of mutual interaction among each and every individual.

As far as (C.I.) is concerned, this line of interpretation of Hume’s theory of justice might, at first, be faced with skepticism, given the more explicit textual evidence in Hume’s work. Hume is known to have provided a critique of the traditional social contract theories, such as formulated, for instance, by Hobbes and Locke. In fact, Hume’s critique helped to pave the way for a broader group of important objections in the 18th and 19th centuries by Hegel, Marx, and the tradition of utilitarian philosophers against the traditional social contract theories. Thus, it is important to address Hume’s arguments against some of the possible formulations of this kind of theory. Here, it is necessary
to make a brief remark. Hume’s arguments are originally intended to target the nature of the government and the justification of political obligations according to the traditional social contract theories. However, it is safe to assume that Hume’s arguments can be generalized, without further complications, to target the nature of justice and the justification of legal obligations as well. This is because, for Hume, the government is a derivative artifice of justice. The government operates as a supportive artifice of the convention of justice in that it is designed to adjudicate and enforce the rules of justice in particular cases and to solve collective action problems when the size of the population gradually increases. In this sense, individuals are led to establish the government motivated by the same interests that prompted them to agree upon the convention of justice in the first place, namely, the interests in social cooperation and peaceful coexistence.\(^\text{T3.2.7.1-8/SBN 534-539}\).

Thus, since, for Hume, the government is grounded on the same interests as those that ground the convention of justice, then the arguments that Hume puts forward against a contractarian conception of the government and the political obligations could also be directed against a contractarian conception of the convention of justice and the legal obligations.

Hume can be interpreted as providing arguments against three kinds of social contract theories in the essay *Of the Original Contract*: i) *original contractarianism*, ii) *explicit contractarianism*, and iii) *tacit contractarianism*.\(^{75}\) The first one addresses the origins of justice and political authority from a contractarian perspective, while the other two address the justification of justice and political authority from a contractarian perspective.

Original contractarianism states that justice and political authority have their

\(^{75}\) This characterization of Hume’s critique is suggested by Gauthier. See Gauthier (1979: 11-12).
origin in a real and historical social contract agreed upon between individuals at a given moment in the past. Explicit contractarianism holds that legal and political obligations can be legitimate and justified if and only if they derive from the explicit consent of individuals in actual agreements to a specific pattern of behavior to be followed. Tacit contractarianism states that legal and political obligations can be legitimate and justified if and only if they derive from the implicit consent of individuals to the legal and political institutions. The idea is that, to the extent that individuals remain within a society, enjoy the benefits and abide by the rules of justice and the laws of the political authority, without protesting the legitimacy of them, this means that they implicitly assent to these rules.

Hume rejects all these accounts on different grounds. Against original contractarianism, Hume argues that there is no empirical evidence that a contract, in such terms, has ever been agreed upon at any given moment in the past. Hume even speculates that societies might have originated from a kind of association of free and equal individuals that voluntarily decided to adopt rules for the mutual restraint of behavior with the expectation of advantageous outcomes. But, the pact upon which original contractarianism is based is not a mere conjecture, but an event referred to as though it had really occurred. And Hume finds this thesis rather implausible (OC 4/473-474). For this reason, Hume concludes that the origins of the convention of justice and the government cannot be traced back to any plausible real and historical social contract agreed upon by individuals.

Against explicit contractarianism, Hume raises several objections. First, Hume argues that, even if it is conceded the extremely controversial existence of an explicit social contract at any time in the past, it does not follow that the justification of the rules and obligations defined by the
convention of justice and the political authority lies in the consent explicitly expressed by this social contract. Hume believes that individuals, when bound by an obligation established by a promise or a contract, must be able to be aware of their own will declared in such pacts (T 3.2.8.9/SBN 547-548). In this particular case, Hume asks us to single out the particular “act of mind” by which it is ascertained a kind of commitment to comply with the rules of justice and to obey the political authority and which each individual is supposed to be aware of if they are really parts of a pact. Hume’s point is that no one really supposes that the act of mind that grounds legal and political obligations is a kind of consent explicitly expressed in a social contract (OC 7/475-476).

Second, Hume also argues that, even if such a social contract had actually been made by the members of society at any time in the past, it would still require that the forthcoming generations come into explicit agreements in order to legitimize the rules and obligations previously defined by the convention of justice and the political authority (OC 8/476). Therefore, an explicit social contract agreed upon at a remote time in the past could not bind generations of individuals who did not take part in the original agreement. It would then be a mistake to consider that they have obligations for the sake of a contract that they did not take part in.

Third, Hume concedes that the consent of individuals expressed by some explicit agreement might constitute a justification for legal and political obligations. However, it does not follow that they cannot be considered justified if they are upheld in some other way (OC 20/480). In fact, according to Hume, the origins of “almost all the governments” can be traced back to either “usurpation or conquest”, and not to the consent expressed by any kind of explicit agreement among individuals (OC 9/476-477). But, governments
can still be regarded as legitimate or justified. This entails that an account of the legitimacy of legal and political obligations requires a kind of justification which should be distinct from a consent expressed by individuals through an explicit agreement. Hume’s suggestion seems to be that the relation between legitimacy and consent is that the legitimacy of legal and political rules and obligations ensures consent and not the other way around. Ultimately, for Hume, the justification of legal and political obligations should refer to the interests of individuals.

Fourth, this social contract theory is based on the assumption that the association of individuals in a society under the rule of a political authority must be established through a contract, or a promise, or a pact, or whatever device that could guarantee the expression of consent of individuals. However, if Hume’s theory of justice, in particular, his account of artificial virtues is correct, it follows that the very obligation to honor a contract or to keep a promise is itself the result of the convention of justice, which, in turn, is a necessary condition for the association of individuals in a society in the first place. Therefore, in order for suitable devices for expressing consent to exist, a convention of justice must have been previously established. And, consequently, a society must also pre-exist as a result of the establishment of a convention of justice. Yet, social contract theorists attempt to make the obligation to keep promises the basis of society and the political authority. For Hume, this idea reverses the proper order of explanation.

Finally, Hume also rejects the tacit contractarianism, which was supported, for example, by Locke and Plato. Against this kind of social contract theory, Hume argues that individuals do not suppose that the legal

---

77 See Locke (2003: 152-153), Plato (1931: 154)
and political obligations to which they are bound depend on their tacit choice or consent. Hume claims that individuals usually observe their legal and political obligations motivated by a sense of utility, interest, and ongoing habit that has been reinforced by the constant permanence in a society shaped by these social arrangements. Therefore, it is not tacit consent, as a conscious act of will, but an acquiescence that grounds the legitimacy of these legal and political rules and obligations, as long as they keep favoring interests of each one of them. For this reason, Hume also rejects the tacit contractarianism as well (OC 23/481). I tend to agree with Gauthier that Hume’s understanding of the traditional social contract theories seems to come down to these three kinds of theories in the Of the Original Contract. Whereas original contractarianism and explicit contractarianism are the theories that Hume has in mind when he refers to Hobbes’ political philosophy, tacit contractarianism is the theory Hume has in mind when he refers to Locke’s political philosophy, as they were Hume’s main philosophical influences in terms of political philosophy.

3.3.3.2 Hypothetical Contractarianism
Despite the critique raised by Hume against the traditional social contract theories in Of the Original Contract, it still remains to be examined whether Hume’s theory of justice could be interpreted as another kind of social contract theory, namely, the hypothetical contractarianism. Hypothetical contractarianism states that legal and political obligations can be legitimate and justified if and only if they can be thought of as earning unanimous assent among rational individuals. This is the kind of social contract theory that resembles (C.I.). Recall that, according to (C.I.), Hume’s theory of justice would be best interpreted if it is conceived as resulting from a hypothetical
agreement. Legal institutions could be thought of as earning unanimous assent among rational individuals, insofar as they favor the interests of each individual concerned and assure that no one of them ends worse off. Now, unlike the other kind of social contract theories, explicit or tacit agreements are ruled out of the justification of legal and political obligations altogether. The point is that, to the extent that legal and political obligations can promote the interests of each individual and afford mutually advantageous outcomes, they manifest a kind of convergence of the interests of all. In this sense, they can be thought of as earning unanimous assent. They can be thought of as resulting from a purely hypothetical agreement.

\( (C.I.) \) was first suggested by Gauthier as a suitable interpretation of Hume’s theory of justice. In Gauthier’s formulation, \( (C.I.) \) involves interpreting Hume’s convention of justice as a kind of contractual convention. Such kind of convention is defined by some device, such as a promise, a contract, or a covenant, by which individuals enter into an agreement on the basis of their individual interests or preferences. Agreements resulting from the convergence of individual interests or preferences are referred to by Gauthier as requiring “interested recognition”. And insofar as agreements are entered into on the basis of interested recognition, they command rational adherence, which Gauthier refers to as “interested obligation”. So, for Gauthier, the contractual convention, which is supposed to describe Hume’s convention of justice, should be defined by a device that allows for an agreement among individuals and that commands compliance as interested obligation by relying on interested recognition. Such kind of convention allows for mutual expected advantages, which, for Gauthier, is essential to
(C.I.) and to justice, as understood by Hume.\(^\text{78}\) In Sugden’s formulation, (C.I.) also characterizes mutual expected advantages as the central point of the account, for it constitutes the basis for both coordinative interactions and voluntary agreements.\(^\text{79}\) In this way, mutual expected advantages constitute the common ground on which both coordinative behavior in Hume and hypothetical voluntary agreements stand.

(C.I.) focus on the perspective of the individual choice. More specifically, it explores the textual evidence in Hume that supports the idea that justice is established in terms of the rational and self-interested choices of individuals. As this study has shown so far, there is actually considerable textual evidence to believe that Hume’s theory of justice can be adequately described by (C.I.). For instance, it should be recalled that, for Hume, each individual is led to agree to establish a convention of justice motivated by a sense of self-interest, more specifically, a sense of restrained self-interest. By this sense of self-interest, which is supposed to be common to all, is meant the intention to conditionally cooperate by complying with a set of rules of justice, as behaving in this way can afford socially advantageous outcomes for each individual. This entails that the individual interests or preferences must be taken into account in the establishment of a kind of agreement, which is, for Hume, the convention of justice. Ultimately, the set of rules defined by the convention yields a mutually advantageous scheme of actions. In Gauthier’s terms, Hume is arguing that each individual is led by “interested recognition” to agree upon an agreement enabled by the device of a

\(^{78}\) See Gauthier (1979: 9, 17-18, 35-36). For Gauthier, (C.I.) is additionally supported by the way Hume’s concept of public utility or public interest should be interpreted. In a passage of the *Enquiry*, the term “public utility”, which is stated as the “sole origin of justice”, is interpreted by Gauthier as a synonym of “mutual expected utility”. In this way, for Gauthier, Hume’s theory of justice could be understood according to (C.I.).

\(^{79}\) See Sugden (2009: 21)
convention, which commands compliance of rational individuals based on “interested obligation”. And it should also be recalled, as it is shown by (E.I.), that the convention of justice should not be interpreted as a kind of original or explicit agreement. It should not be comprehended as depicting a real or historical agreement. Although Hume’s theory of justice is mainly intended to address the origins of justice, it is supposed to address it in a purely methodological way. Hume’s theory of justice must be interpreted from the perspective of conjectural history. The convention of justice must be imagined as if it has been agreed upon and followed by individuals in order to satisfy their interests. The convention of justice must be conceived as a rational response to the specific interests and necessities of individuals given the circumstances in which they would find themselves in a supposed and imaginary state of nature. Such characteristics of Hume’s theory of justice, which have already been systematically articulated by the game theory model of the Stag Hunt, seem to support (C.I.) as a suitable interpretation of Hume.

The point that should be emphasized here in the following one. The most distinctive aspect of (C.I.) is that, unlike (RU.I.) and (E.I.), it makes that individual interests or preferences are taken directly into account as rational responses in the establishment of the convention of justice and the compliance with the pattern of behavior defined by its rules. While in (RU.I.) some individual interests or preferences might be disregarded in the establishment of the convention of justice, since they are considered only indirectly in an aggregative conception of social welfare, the only factor that is rational to maximize, in (C.I.) not any single individual interest or preference is disregarded in the establishment of the convention of justice, since, according to this description, it is intended to afford mutual expected advantages. In (C.I.), the expectation of mutual advantage is a necessary condition for the
establishment of the convention of justice, while, in (RU.I.), the expectation of maximization of the total sum of happiness is a sufficient condition for the establishment of the convention of justice. In this sense, (C.I.) seems to be a description that properly fits with Hume’s theory of justice.

3.3.3.3 The General Scheme of Justice

However, it is also worth bringing textual evidence that casts doubt on the belief that (C.I.) qualifies as a suitable interpretation of Hume’s theory of justice. Take, for instance, the following three passages. In paragraph 22 of the Treatise Book 3, Part 2, Section 2, Hume considers a relevant point and provides some examples to illustrate it:

To make this more evident, consider, that tho’ the rules of of justice are establish’d merely by interest, their connexion with interest is somewhat singular, and is different from what may be observ’d on other occasions. A single action of justice is frequently contrary to public interest; and were it to stand alone, without being follow’d by other acts, may, in itself, be very prejudicial to society. When a man of merit, of a beneficent disposition, restores a great fortune to a miser, or a seditious bigot, he has acted justly and laudably, but the public is the real sufferer. Nor is every single action of justice, consider’d apart, more conducive to private interest than to public; and ‘tis easily conceiv’d how a man may impoverish himself by a signal instance of integrity, and have reason to wish, that with regard to that single act, the laws of justice were for a moment suspended in the universe. (T 3.2.2.22/SBN 497)

Later, still in this same passage of the Treatise, Hume expands on this point:

But however single acts of justice may be contrary, either to public or private interest, ‘tis certain, that the whole plan or scheme is highly conducive, or indeed absolutely requisite, both to the support of society, and the well-being of every individual. ’Tis impossible to separate the good from the ill. Property must be stable, and must be fix’d by general rules. Tho’ in one instance the public be a sufferer, this momentary ill is amply compensated by the steady prosecution of the rule, and by peace and order, which it establishes in society. And even every individual person must find himself a gainer, on ballancing the account; since, without justice, society must immediately dissolve, and every one must fall into

See Gauthier (1979: 9-10).
that savage and solitary condition, which is infinitely worse than the worst situation that can possibly be suppos’d in society. (T 3.2.2.22/SBN 497)

And the same point is also brought in paragraph 3 of the Enquiry Appendix III, which suggests the importance that Hume attaches to this point:

The case is not the same with the social virtues of justice and fidelity. They are highly useful, or indeed absolutely necessary to the well-being of mankind: but the benefit resulting from them is not the consequences of every individual single act; but arises from the whole scheme or system concurred in by the whole, or the greater part of the society. General peace and order are the attendants of justice or a general abstinence from the possessions of others; but a particular regard to the particular right of one individual citizen may frequently, considered in itself, be productive or pernicious consequences. The result of the individual acts is here, in many instances, directly opposite to that of the whole system of actions; and the former may be extremely hurtful, while the latter is, to the highest degree, advantageous. Riches, inherited from a parent, are, in a bad man’s hand, the instrument of mischief. The right of succession may, in one instance, be hurtful. Its benefit arises only from the observance of the general rule; and it is sufficient, if compensation be thereby made for all the ills and inconveniences which flow from particular characters and situations. (EPM App 3.3/SBN 304)

In all three passages, Hume raises a point that his theory of justice seems to imply. Hume notices that while particular acts of justice may be harmful to “private” or “public interest”, the general scheme of justice, in particular, the system of private property rights, is still beneficial to the “public interest” or to “society”. Public interest is defined by the set of interests or ends of society as a whole. Peaceful coexistence and social well-being stand out as the fundamental interests or ends of society, as general “peace” and “order” and the “well-being of mankind” are particularly highlighted by Hume in the passages. As examples that illustrate this problem, Hume observes that when an individual respects private property rights by “restoring” property, upholding “inheritance rights”, and accepting “succession” rights in favor of a selfish individual, the “public” is the real “sufferer”, for resources tend to be concentrated in the hands of a “miser”, a “seditious bigot”, or a “bad man”
in this particular case. However, if this single act of justice is understood in
the logic of a regular pattern of behavior of compliance with the rules of
justice, which should be “concurred in by the whole, or the greater part of the
society”, then the “whole plan” or “scheme” of justice is “highly useful” to
society, insofar as it satisfies the public interest by “supporting” society and
promoting the “well-being” of “every individual”. But, the way by which the
general scheme of justice is useful to society is very particular. Hume points
out that the eventual “ills” to private or public interest derived from a single
act of justice cannot be distinguished from the “goods” to society derived
from the general scheme of justice, which consists precisely of the total sum
of single acts of justice. For Hume, the “momentary ills” of a single act of
justice are “compensated” by the greater good afforded by “the steady
prosecution” of the rules of justice in upholding a peaceful social coexistence,
so that, when this social structure of interaction is considered and the gains
and losses “balanced” by each individual, they must find themselves
“gainers” in this case.

Now, the question is how to understand the point conveyed by the
three passages aforementioned. As Mackie points out, Hume could be
interpreted as making one of the following four statements: i) single acts of
justice are not harmful in themselves, they might be harmful only in the
absence of the general scheme of justice; ii) single acts of justice only seem
to be harmful in the light of the short-term interests, but, in reality, they are
beneficial in the light of their contribution to the general scheme of justice in
satisfying of their long-term and rationally considered interests, which
coincides with the public interest; iii) single acts of justice might be harmful
in the light of the short-term interests, but the harm is worth suffering in the
light of their contribution to the general scheme of justice in satisfying long-
term and rationally considered interests, which coincides with the public interest; and iv) single acts of justice might be harmful both in the absence or in the existence of the general scheme of justice.  

Although the four statements may seem to be incoherent or even paradoxical with each other, they do capture Hume’s point both from the perspective of the concurrence or isolation of singular acts of justice and the general scheme of justice and from the perspective of the different interests that they are supposed to address. So, it is necessary to define which of the four statements better describes Hume’s point.

At first, Hume’s point seems to focus on highlighting an apparent incongruity by imagining the harmful effects of single acts of justice when considered apart from the beneficial regular pattern of behavior of compliance with the rules of justice, as the first passage suggests. However, in the following two passages, this focus seems to shift. Hume proceeds to address the incongruity between the harmful effects of single acts of justice and the beneficial effects of the general scheme of justice when considered together, which indicates that Hume’s real concern is not much with single acts of justice considered apart from the general scheme of justice but with the different interests that both are supposed to address when they are considered together. This can be noticed when Hume points out that it is “impossible to separate the good from the ill” that a single act of justice entails in the context of the general scheme of justice. Or when Hume points out that the “momentary ill” that a single act of justice entails is “amply compensated” by “peace and order” promoted by the general scheme of justice. So, it is safe

\[81\] See Mackie (1980: 91).
to assume that the first and fourth statements are ruled out as suitable interpretations of Hume’s point.

Now, with regard to the different interests addressed by single acts of justice and the general scheme of justice, there are two possible interpretations. Either single acts of justice are seemingly harmful in the light of short-term interests, but are really beneficial in the light of long-term and rationally considered interests or single acts of justice are really harmful, but, in the light of long-term and rationally considered interests, the losses are bearable. As it stands, Hume never seems to suggest that single acts of justice cannot be harmful in themselves based on a distinction between the short and long-term interests at stake. In fact, single acts of justice are characterized as really harmful in particular cases by Hume. It is pointed out that the results of “individual acts” might be “extremely hurtful” or are “frequently contrary” to “private” or “public interest”. The “public” might be a “sufferer” of the result of single acts of justice which might be “directly opposite to that of the whole system of actions”. For Hume, single acts of justice are not necessarily meant to directly address the long-term and rationally considered interests. Instead, they address this kind of interest only indirectly. Every single act of justice is indirectly beneficial by contributing to a more general scheme, which, in turn, is capable of satisfying long-term and rationally considered interests, which also constitute the most fundamental public interests at stake, such as “peace and order” in society. However, single acts of justice still might be directly harmful to the short-term public and private interests in particular cases. But, the point is that the satisfaction of the long-term and rationally considered interests is able to “compensate” the harm to the short-term public and private interests in particular cases. As Hume puts it, the “ills and inconveniences” from “particular characters” and “situations” are “compensated” by the social
benefits, so that, “on balancing the account”, individuals must find themselves “gainers”. So, it can be concluded that the second statement is also ruled out as suitable interpretations of Hume’s point, which leaves only the third statement as a suitable interpretation of Hume’s point.\textsuperscript{82}

As far as (C.I.) is concerned, the only two possible statements that could qualify as suitable interpretations of Hume’s point are the first and second statements. Recall that, according to (C.I.), Hume’s theory of justice would be best interpreted if social and legal institutions could be thought of as earning unanimous assent among rational individuals. To the extent that the social and legal rules and obligations promote the interests of each individual concerned and assure that no one of them ends worse off, they can be thought of as resulting from a convergence of the interests of all. In this sense, they can be thought of as earning unanimous assent. What (C.I.) implies is that, by definition, the social and legal rules and obligations must be able to afford mutually advantageous outcomes for rational individuals. They can only be thought of as earning unanimous assent if rational individuals could agree with them based on the expectation that they can satisfy their interests. And, rational individuals, insofar as they are rational, would agree only with the kind of social and legal rules and obligations that can satisfy their rationally considered interests. This entails that, by

\textsuperscript{82} As a matter of fact, Hume is really ambiguous here. In the Enquiry Appendix III, Hume also suggests, by the example of Cyrus in the fourth paragraph, that the balancing that characterizes the rules of justice is between short and long-term and rationally considered interests, so that, upon a careful deliberation, the latter eventually overcomes the former as the kind of interest that really matter in this discussion (EPM App 3.4/SBN 304-305). In this sense, this passage supports the second statement as a suitable interpretation of Hume’s point. However, in the third and sixth paragraphs, Hume brings passages that suggest otherwise. So, given the contrasting textual evidence, it is reasonable to hold the position that is emphasized the most by Hume in his work, which is precisely the one conveyed in multiple passages of the Enquiry and the Treatise, as I have previously quoted.
definition, they cannot be harmful in the light of rationally considered interests. When I say that social and legal rules and obligations must be able to afford mutually advantageous outcomes for rational individuals, I do not mean that their rationally considered interests cannot be short-term interests or that there is no conflict of interests in particular cases. Rather, I mean that, according to (C.I.), the social and legal rules and obligations cannot be harmful in the light of rationally considered interests. In this sense, (C.I.) requires that single acts of justice are beneficial in the light of the long-term and rationally considered interests. For this reason, only the statements that claim that single acts of justice cannot, in principle, be harmful are suitable interpretations of Hume’s point according to (C.I.), which, as I have pointed out, is not the case.

For scholars such as Harrison and Rawls, the three passages aforementioned constitute textual evidence for (RU.I.). This is not an implausible interpretation at all. These passages seem to suggest that Hume shifts from the perspective of the individual choice to the perspective of the scheme of justice itself, which favors (RU.I.). Hume is strongly emphasizing the global advantages afforded by the system of justice. Hume’s description does seem to resemble a rule-utilitarian argument in favor of the duty to comply with the rules of justice even if, in particular cases, it might be directly harmful to the short-term private or public interests, for the eventual losses would be compensated by the benefits resulting from the satisfaction of long-term interests. However, it is not certain that Hume’s point, as depicted by the three passages aforementioned, only by itself, can authorize the conclusion that Hume’s theory of justice would be best described by (RU.I.). For

instance, Mackie interprets Hume as claiming precisely the opposite of (RU.I.) in the three passages. For Mackie, the point is that Hume is not so concerned with single acts of justice, but with a practice of justice. Hume wants to highlight that, in face of a beneficial regular pattern of behavior defined by a general scheme of justice, eventual losses resulting from single acts of justice are balanced and compensated by the benefits in the end. If the discussion is framed in terms of choosing between regular practices and cultivating a disposition accordingly rather than choosing between acts in particular cases, the problems concerning the isolation of acts from practices or the different interests that both are supposed to address fade away. In this sense, not only is Hume’s account not incoherent or paradoxical as it seems at first glance, but it also does not particularly speak for (RU.I.), which is a moral theory primarily based on the morality of individual actions.\footnote{See Mackie (1980: 92-93).}

More important than settling which interpretation fits better with Hume’s point, these passages reveal a relevant element of Hume’s theory of justice. It shows that Hume’s concept of public utility or interest differs from that of private utility or interest. According to (C.I.), the concept of public utility is built as a function of the private utilities of all individuals. The public interest should be defined by the set of shared and common interests of each individual, insofar as they are rationally considered. It is only when rational individuals can expect that the set of shared and common interests of all could be satisfied by rules of justice that they could be thought of as earning unanimous assent. As a result, according to (C.I.), the public interest is satisfied only if the shared and common interests of all individuals are also satisfied. However, this is not what Hume claims. Or, at least, Hume was
unsure about how both concepts could be articulated. The point raised in the three passages suggests that Hume does not think that the public interest and the shared and common individual interests coincide. In fact, given that rules of justice might cause eventual losses to private interests in particular cases while satisfying long-term public interests, rational individuals might even question themselves whether it is rational to comply with this kind of rules in the first place, as Stroud rightly points out. And this point is reinforced by other passages where Hume clearly feels the need for a distinction between the “publick interest” and “that of every individual”. For Hume, the rules of justice depend on “these different interests” (T 3.2.2.20-21/SBN 496-497).

This kind of textual evidence casts doubts on (C.I.) as a suitable interpretation of Hume’s theory of justice.

In this chapter, I have carried out the analysis of Hume’s convention of justice, which is the central concept of Hume’s theory of justice. I have addressed the origins and foundations of the virtue of justice by surveying the motive that induces individuals to agree upon a convention of justice and follow the pattern of behavior defined by its rules. This motive need not be a moral motive, but a specific sense of self-interest, the enlightened self-interest, which amounts to a restrained sense of self-interest directed to establishing a set of rules of justice and to behaving accordingly.

I have also argued that, contrary to what a commonly accepted conventionalist interpretation holds, the nature of the convention of justice is that of an agreement. I have also pointed out that the rules defined by Hume’s convention of justice are three: i) the rule of private property rights, ii) the rule of transference of private property by consent, and iii) the rule of the obligation of promises. I have tried to show how Hume comes up with this particular set of rules and not another one. In this sense, I have not only tried to make explicit the logic and the development dynamics of this set of rules of justice given the empirical circumstances in which individuals face themselves in, but also to show possible objections and limits to Hume’s account.

Finally, I have carried out this analysis with the purpose of evaluating Hume’s theory of justice in the light of the three lines of interpretation suggested to far. As I have suggested, the three lines of interpretation make important contributions to the general understanding of Hume’s theory of justice, but they also fall short of providing a fully adequate description of it.
However, this account of the convention of justice does not exhaust Hume’s treatment of the subject. In fact, most of Hume’s account in the *Treatise* Book 3, Section 2, Part 2 and in the *Enquiry* Section III, concerns the rule of the stability of possessions among individuals, which is firstly introduced by the convention. This rule grounds private property rights in Hume. But Hume also provides details about the other rules of justice. For this reason, it might be useful to inquiry into Hume’s account of the rules that specify the private property rights, the rule of the transference of property by consent, and the rule of the obligation of promises in order to understand how they relate to the lines of interpretation. In this way, by surveying the details of the “three laws of justice”, I will be able to secure a fully systematic understanding of Hume’s theory of justice as a whole. I will address them in the next chapters.
4.1 The Initial Distribution
4.1.1 Immediate or Present Possession

In the last chapter, I have put forward careful and detailed analysis of the central concept of Hume’s theory of justice, the convention of justice. I have attempted to clarify the nature of the convention of justice and to understand it in light of the three main lines of interpretation that have been suggested in the tradition of Hume’s scholarship. The convention of justice is defined by a specific set of rules: i) the rule of the stability of possessions, ii) the rule of the transference of property by consent, and iii) the rule of the obligation of promises. For Hume, this set of rules is conventionally established. However, recall that, in the Treatise, the convention of justice is first introduced mainly as an agreement establishing the rule of the stability of possessions. It is only after the introduction of this rule that Hume proceeds to introduce the others. So, taken for granted a rule to stabilize possessions, which ultimately grounds the private property rights, it is necessary to account for the rules that specify the particular distribution of possessions among individuals and the rule of the transference of possessions in order for Hume’s theory of property to be completely explained. In this way, Hume’s theory of property, which is a
central part of Hume’s theory of justice, could be properly interpreted in light of the three main lines of interpretation suggested so far.

At the beginning of the *Treatise* Book 3, Part 2, Section 3, Hume claims that a kind of criterion is required by which possessions should be distinguished and assigned to each individual. Even if the social recognition of private property rights depends on the stabilization of possessions among individuals established by the convention of justice, it is only through a specific distribution criterion that particular possessions can be recognized as private property of particular individuals in a society. This is why it is useful for the general rule of the stability of possessions to be further specified by a particular distribution rule. This is Hume’s point in the following passage:

Tho’ the establishment of the rule, concerning the stability of possessions, be not only useful, but even absolutely necessary to human society, it can never serve to any purpose while it remains in such general terms. Some method must be shewn, by which we may distinguish what particular goods are to be assign’d to each particular person, while the rest of mankind are excluded from their possession and enjoyment. *(T 3.2.3.1/SBN 501-502)*

And such a specific distribution criterion defines the share of possessions that each individual is allowed to enjoy after the establishment of the rule of the stability of possessions. This is Hume’s point in another passage:

’Tis evident, then, that their first difficulty, in this situation, after the general convention for the establishment of society, and for the constancy of possessions, is how to separate their possessions, and assign to each his particular portion, which he must for the future inalterably enjoy. *(T 3.2.3.4/SBN 503)*

In other words, a distributive pattern of possessions must also be somehow set by a specific distributive criterion at the starting point of a society, when a convention for the stabilization of possessions is agreed upon and private property rights are established. And, since the starting point of a society is defined by a convention of justice, which marks the departure of a hypothetical
state of nature, Hume needs to come up with a distributive criterion that defines the initial distribution of possessions in a society. Without such a convention, there is no use in coming up with a distributive criterion by which possessions should be distinguished and assigned to each individual. Under this circumstance, no “method must be shewn” for defining the initial distribution, as possessions would not be recognized or protected as private property anyway.

Hume argues that the distributive criterion that defines the initial distribution should be the rule of immediate or present possession (T 3.2.3.4/SBN 503; EPM App. 3.10/SBN 309). The idea is that the convention for the stabilization of possessions should be agreed by taking into account the immediate or present possessions of each individual to establish private property rights, for this rule is promptly induced by the imagination and easily reinforced by habit. Hume believes that this rule is the most obvious that individuals, who need to decide together how to distribute their possessions, can come up with. Since the distributive criterion of immediate or present possession is considered, in Hume’s terms, as the “most natural expedient” to be followed, it follows that this rule could secure an easy “acquiescence” or a natural “agreement” among individuals on how the distribution should be conducted, as Hume points out in the following passage:

This difficulty will not detain them long; but it must immediately occur to them, as the most natural expedient, that every one continue to enjoy what he is at present master of, and that property or constant possession be conjoin’d to the immediate

---

86 There are different ways by which the idea of an initial distribution can be defined. For instance, Buchanan thinks that an initial distribution is defined by a random arrangement of goods or resources among individuals which precedes the equilibrium established by a natural distribution. See Buchanan (1975: 31). This is not what I mean by an initial distribution in Hume’s theory of justice. Instead, by initial distribution I mean simply the arrangement of possessions among individuals that must be established when a convention for the stabilization of possessions is agreed upon and private property rights are established.
possession. Such is the effect of custom, that it not only reconciles us to any thing we have long enjoy’d, but even gives us an affection for it, and makes us prefer it to other objects, which may be more valuable, but are less known to us. What has long lain under our eye, and has often been employ’d to our advantage, that we are always the most unwilling to part with; but can easily live without possessions, which we never have enjoy’d, and are not accustom’d to. ‘Tis evident, therefore, that men wou’d easily acquiesce in this expedient, that every one continue to enjoy what he is at present posses’s’d of; and this is the reason, why they wou’d so naturally agree in preferring it. (T 3.2.3.4/SBN 503-504)

While the idea of an initial distribution based on a criterion of the immediate or present possession is quite intuitive, the account, by itself, is hardly convincing and ends up sparking more questions than it actually answers. Is Hume justifying this distributive criterion for the initial distribution or is he explaining how the initial distribution eventually comes to be established by this distributive criterion? Is it really true that the rule of immediate or present possession is the most obvious one for defining the initial distribution? Does it really secure an easy acquiescence or a natural agreement among all? This account leaves many questions unanswered. On the one hand, it seems clear that Hume is partially concerned with the description of the initial distribution. After all, Hume’s theory of justice is a theory of the origins of justice. On the other hand, it is also plausible to make the stronger claim that Hume is also partially concerned with the justification of the distributive criterion that defines the initial distribution. In fact, given the textual evidence, it could even be argued that Hume went as far as to outline a theory of distributive justice. Now, in order to understand how Hume attempts to justify the distributive criterion of the immediate or present possession, it is necessary to bring into the discussion Hume’s views on possible alternative distributive criteria for the initial distribution. What I intend to show is that Hume did not believe that alternative distributive criteria are plausible candidates for defining the initial distribution for several reasons. Thus, the distributive criterion of immediate
or present possession could be regarded as a rationally acceptable rule for the initial distribution, at least as Hume conceived it. After this, I intend to assess this rule in light of the lines of interpretation, in particular, given the premise that a distributive criterion should be able to earn acquiescence and acceptance of individuals. So, let us take a look at possible alternative distributive criteria.

4.1.2 Original Acquisition

Instead of the immediate or present possession, it could be imagined that a more rational distributive criterion for defining the initial distribution would be that of the original acquisition. The original acquisition is commonly regarded as a libertarian distributive criterion. In the philosophical tradition, it is supported by John Locke and Robert Nozick. The idea is that, in order to define the initial distributive pattern at the starting point of a society, instead of asking which possessions each individual currently holds, one should ask which possessions each one of them originally came to hold until this time. But, this acquisition should not be morally arbitrary. Possessions must be acquired fairly. By “fairness” it is meant that individuals, in their activity of appropriation in the natural state, must, firstly, respect the appropriation of others in the natural state, without forcefully depriving them of what they have fairly acquired themselves, and must, secondly, restrict their appropriation so that no one ends up worse off than if the possession remained in its natural state. The intuition here is that the appropriation in the natural state should not be predatory to the point that possessions end up running out for others.\(^7\)

Hume addresses Locke’s argument that supports this distributive criterion in the second footnote of the *Treatise* Book 3, Part 2, Section 3. The

---

\(^7\) The complete formulation of the distributive criterion of the original acquisition, including the idea of the Lockean proviso, is originally found in the chapter V of
argument goes roughly as follows. Individuals are entitled to possessions by original acquisition insofar as they mix their own labor to the object they intend to acquire in the natural state. This entitlement is based on the fact that each individual is endowed with the fundamental moral status of self-ownership, as Locke claims. Self-ownership means a kind of basic entitlement or, as Locke puts it, a “property right”, that each individual possesses over their own selves, bodies, and powers, such as, for instance, their labor and the fruits thereof. Therefore, by mixing labor to unclaimed possessions in the natural state, individuals can secure them as private property fairly. A similar argument is also held by Nozick.

This account could be contested by Hume for two reasons. First, in the footnote, Hume argues that it is only in a “figurative sense” that labor can be said to be joined to possessions by individuals. Labor can only transform possessions. Labor does not transmit any kind of moral or metaphysical property from individuals to possessions. Nonetheless, Hume points out that labor is still a very relevant element for the purpose of defining private property rights. It establishes a relation between individuals and possessions which can eventually lead them to the creation of other private property rules through an imaginative process. In fact, Hume admits that the original acquisition or, as he calls it, “first possession” or “occupation”, is an additional criterion for the establishment of private property rules, but it presupposes previous distributive criteria (T 3.2.3.6/SBN 505-506. EPM App. 3.10/SBN).

---

88 In Locke, the argument is formulated exactly in such terms. See Locke (2003: 111-112). In Nozick, the argument is intended to make the same point, but it also addresses the fact that one should not seize the results of others’ labor, for it implies the appropriation of others’ working time, which violates self-ownership. See Nozick (1974: 171-172).
Second, it seems that this argument depends on very demanding assumptions. What Hume demands is an explanation of how an initial distribution can be secured in a way that is easily acquiesced or naturally agreed among individuals. Locke and Nozick are able to provide such an explanation, but only at the cost of assuming, from the start, substantial and yet questionable moral and metaphysical views, which involve, for instance, the idea of self-ownership. Hume was too skeptical to accept such assumptions. Instead, Hume would rather start from a more neutral point of view. Hume would reject the claim that individuals are endowed with the basic entitlement of self-ownership, which allows for the initial distribution to be defined and, ultimately, grounds private property rights. Individuals are not entitled, by self-ownership, to specific shares of possessions in the hypothetical state of nature. Neither do the possessions acquired in this way become private property by a quality of the agents alone, independently of the social recognition of all others. This is implied by Hume when he claims that there could be no violation of private property in the hypothetical state of nature, for there was simply “no such thing” as “property” and “justice” or “injustice” (*T* 3.2.3.28/SBN 501). For Hume, the acquired possessions need not be respected on grounds of self-ownership. It can be noticed that the distributive criterion of the original acquisition is not so simple as it seems. In fact, it assumes quite questionable moral and metaphysical concepts. It can be concluded that this distributive criterion cannot be regarded as more natural or rational than Hume’s. In virtue of a justification based on a rather mysterious notion of self-ownership, the original acquisition is not a
convincing candidate to secure easy acquiescence or a natural agreement among individuals.

4.1.3 Equality
A second distributive criterion that could be imagined for defining the initial distribution would be that of the egalitarian one. In a scenario where the initial distribution is defined by an egalitarian distributive criterion, the sum of the social possessions must be arranged so that the resulting distributive pattern guarantees equal shares of possessions for each individual. The idea is very intuitive. An initial distribution can only be considered acceptable if it ensures that each individual enjoys equal treatment with regard to the distribution of possessions. Although this interpretation of egalitarian distribution may be regarded as problematic for several reasons, for the purpose of this discussion, I suggest that the model for an egalitarian distribution of possessions should be interpreted simply as that of strict equality. Other interpretations could be suggested, but it would be useless to spell all of them out here, as none of them would be suitable for defining the initial distribution, as Hume believed. In interpreting egalitarian distribution in this way, I attempt to ensure that, in the context of an initial distribution, the criterion is introduced at the starting point of a society, according to which everyone should enjoy equal treatment with

---

There are many ways to interpret an equal distribution. It can be interpreted in a wide variety of ways, such as strict egalitarianism, welfare egalitarianism, luck or resource egalitarianism, opportunity egalitarianism, and so on. In this spectrum, there are major differences in what should be equally distributed and what should be accomplished by an equal distribution.
respect to this distributive aspect, but also that it can leave room for later unequal distributions that may arise from voluntary choices of each individual.

If, on the one hand, the egalitarian distributive criterion seems to be a more rational candidate for defining the initial distribution, on the other, Hume would find it very unlikely that individuals would accept it. It is well known that Hume has a critique of perfectly egalitarian societies, which I will still address later on in this chapter (EPM 3.26/SBN 194). However, I would like to highlight another point here. For Hume, individuals are, in general, most directly concerned with themselves, their relatives, and their acquaintances. In other words, they are most directly concerned with their “narrow circle” of relationships. As I have previously pointed out, this idea stems from Hume’s theory of passions and theory of association of ideas. For Hume, the impression of ourselves is one of the strongest in our psychological constitution. Therefore, each one of us tends to focus more closely on the corresponding idea of ourselves than on any other, whenever it is presented in our consciousness. Principles of association of ideas such as, resemblance, contiguity in time and space, and causation, operate in such a way that redirects this tendency for each one of us to focus on objects related to us, which explains this particular concern with our “narrow circle” of relationships. For Hume, this is the psychological source of our natural partiality and the wide range of our self-referential passions, such as pride, humility, self-love, self-interest, ambition, avarice, and so on. And, in the hypothetical state of nature, where circumstances of moderate scarcity and easy transition of the possessions of some to others prevail, our natural partiality takes the form of selfishness and avarice, given the urgent need to acquire and secure scarce external goods. Self-interest turns into a strong avidity for acquiring increasingly larger shares of scarce external goods. When
individuals strive to acquire external goods for themselves, their relatives, and their acquaintances in this way, they are likely to engage in potential violent competition over the acquisition of such goods. In this sense, Hume holds a similar view to that of Hobbes with regard to the social interaction in the hypothetical state of nature, as I have already pointed out.

Given this description, it seems that the egalitarian distributive criterion, for Hume, would fall short of the conflicting claims of individuals. It is hard to imagine that, in the context of a convention for the stabilization of possessions and the establishment of private property rights, each individual would acquiesce to the egalitarian distributive criterion that requires that the sum of the social possessions be equally distributed, while each one seeks to maximize the acquisition of possessions only for themselves. Thus, rather than giving away their possessions for the purpose of establishing a more egalitarian society, they would probably prefer to keep the possessions they have acquired for a future convention that introduces private property rights. Even if the egalitarian distributive criterion may ensure fairness by granting each individual with equal shares of the total sum of social possessions, some of them might still question themselves why they should contribute for the rearrangement of the distributive pattern of possessions towards a more egalitarian society if they may end up worse off than before. 90 Certainly, some of them would have no incentives to accept this distributive criterion. This entails that the egalitarian distributive criterion is not a very

90 This argument is developed by Stemmer (2000: 244-245). However, given Hume’s assumptions in his theory of passions and theory of association of ideas, it is very plausible to argue that this outcome would follow as a rational generalization of the premises established by the theories. So, it is reasonable to argue that Hume would agree with this argument.
persuasive candidate to secure easy acquiescence or a natural agreement among individuals for defining the initial distribution, as Hume requires.91

4.1.4 Desert

A third distributive criterion that could be imagined for defining the initial distribution could be that of desert. Traditionally, desert has been considered a very relevant criterion in distributive justice theories. So, it could be argued that desert could be thought of as a more rational rule than the immediate or present possession for defining the initial distribution. The problem with this distributive criterion is that it requires the specification of a basis for what it is actually deserved. In the *Enquiry* Section III, Hume has considered the distributive criterion of desert for the distribution of possessions in general. The basis for deserving, in turn, would be defined by the virtuous or vicious character of each individual. In other words, merit would be defined by virtue or vice. As far as the initial distribution is concerned, the idea is that the arrangement of possessions among individuals should be patterned in such a way that the more virtuous one is, the larger share of social possessions one would be granted. In this way, for Hume, the most virtuous individuals would be assigned the largest sum of possessions and the largest amount of power too. Ultimately, this social arrangement would “best promote the public

---

91 Interestingly, the distributive criterion of immediate or present possession could end up establishing an egalitarian distribution of possessions, but this would certainly not be designed beforehand. Rather, it would be a random outcome, which is completely contingent upon how the immediate or present possessions would actually end up being arranged.
interest” by ensuring “peace” and “security among mankind” (*EPM* 3.23/SBN 192-193).

This description is similar to the one supported by Aristotle. For him, a fair distribution should be the one that rewards individuals in proportion to their merit. And, since individuals are not equal, it is not fair that they are rewarded with equal shares in the distribution. But, as Hume points out, the idea that some individuals deserve shares of possessions based on their virtuous characters faces a serious difficulty. The problem is with the uncertainty of merit evaluation or, as Hume puts it, the “uncertainty of merit”. For Hume, there are two reasons for this uncertainty. Firstly, because individuals are, in general, “self-conceited”, each one of them would tend to claim the merit only for themselves (*EPM* 3.23/SBN 193). I suppose that Hume believed, just like Hobbes did, that each individual would hardly admit that others are as talented or righteous as themselves. Consequently, given the partial self-evaluation of their own virtuous character, they would frequently claim the highest merit for distributive purposes. And, when a large number of individuals claims merit in this way, its evaluation becomes increasingly uncertain. Secondly, and an even more fundamental point, the concept of merit, which is essential for the distributive criterion of desert, is characterized by a kind of “natural obscurity”, in Hume’s own terms (*EPM* 3.23/SBN 193). This characterization seems to imply that, for Hume, the very concept of merit is somewhat indistinct and confusing, maybe even chimerical. I suspect that Hume’s concern with the concept of merit is based on his belief in determinism. In the *Treatise* Book 3, Part 3, Section 4 and in

---

92 See Aristotle (*EN* 1131a20-30).
94 Baier holds the same view. See Baier (2010: 105).
the *Enquiry* Appendix IV, Hume discusses the alleged distinction between natural abilities and moral virtues. Traditionally, it has been claimed that the former are devoid of any moral value, while the latter are the suitable objects of moral evaluation. This conception is supported by the common belief that moral virtues are voluntary, while natural abilities have “no dependency on liberty and free-will” (*T* 3.3.4.3/SBN 608. *EPM* App. 4.2/SBN 313). Against this conception, Hume claims that moral virtues are not as dependent on choice or free will as any other natural ability. In a fully deterministic world, “external bodies” are as uniformly determined in their interactions by other bodies as “voluntary actions” are uniformly determined by “motives, tempers, and circumstances” (*T* 2.3.1.3-4/SBN 399-401. *EHU* 8.5-7/SBN 82-84). As both natural abilities and moral virtues are not especially distinct with regard to universal causal determinism, Hume concludes that there is no reason for the moral evaluation of virtues and not of the natural abilities, at least not because of the notions of “voluntary” and “involuntary” behavior. It follows that, for Hume, there is no more reason for claiming merit by virtuous behavior than by any natural ability.

This conclusion does not favor the distributive criterion of desert, since the very idea of merit is challenged. Since Hume is concerned with settling “some verbal disputes” in the context of this discussion, the very application of the concept of merit would be completely unwarranted, as the concept is built upon a very obscure notion. In fact, Hume suggests that both natural abilities and mental dispositions should be morally evaluated insofar as both qualities can convey pleasant or unpleasant sentiments of approbation or disapprobation in a spectator from a general point of view. Both natural abilities and mental dispositions may have a tendency to be useful or agreeable to the individuals themselves or to society (*T* 3.3.4.3/SBN 608-609. *EPM* App.
In this account, Hume dispenses completely with the concept of merit and the notions of what is “voluntary” or “involuntary”.

It could be argued that Hume’s argument is not very convincing. I tend to agree with this claim. There seems to be a real distinction between both qualities. Perhaps Hume should not consider the virtuous or vicious character of each individual as the basis for deserving in the first place. The amount of individual effort or social contribution could be suggested as a more plausible basis for the distributive criterion of desert. But, even in face of the many consistent objections that could be raised against this approach of merit and deserving, the point that Hume is making, although only superficially, is of extreme importance. In fact, it reflects the contemporary criticism of distributive justice theories that take merit or desert as the criterion for a fair distribution. Rawls, for instance, argues that the distributive criterion of desert fails to properly take into account arbitrary elements, such as social contingencies and natural chance, which invariably play a role in determining what is necessary for individuals to be held responsible for their actions. And this is because of the deterministic intuition that Hume has also realized. Since behavior is at least partially determined by arbitrary elements over which individuals lack control, merit cannot be, in general, precisely ascertained. In this sense, Rawls would agree that the distributive criterion of desert would be characterized by a kind of “uncertainty”, which is how Hume depicts this notion as well. Since the distributive criterion of desert is based on rather confusing, uncertain, and indeterminate grounds, it can be concluded that it

---

95 See Rawls (1971: 74, 311-312).
cannot secure an easy acquiescence or a natural agreement among individuals for the purposes of defining an initial distribution.

4.1.5 Natural Distribution

So, it can be concluded that, for Hume, the initial distribution should not be determined by any distributive criterion other than that of the immediate or present possession. Only the possessions that individuals manage to acquire and secure at the time must be taken into account in a convention for the stabilization of possessions and the establishment of private property rights. And, although Hume does not explain it in detail, this distributive criterion is not entirely unjustified. In fact, it could also be understood as more rational than others in important ways. The point is that it does not matter how much other distributive criteria may seem fair or rational for arranging the sum of social possessions if individuals may simply not agree to dispose of their possessions for this purpose. Such considerations of fairness or rationality would not be convincing for individuals, as Hume depicts them. At the same time, however, it is in their interest to “separate” or “distinguish” the possessions in some way, so that they can be “assigned” to “particular persons” as private property (EPM App. 3.10/SBN 309). So, Hume claims that it would be better to leave each individual with their immediate or present possessions, which they are “always the most unwilling to part with”. Notice, however, that Hume’s claim implies that the distribution must fit the way individuals are willing or unwilling to behave. In this sense, Hume’s claim is supported by the premise that a distributive criterion for the initial distribution should be one to which individuals would “easily acquiesce” or upon which they would “naturally agree”. And, if the discussion is framed by the premise that a distributive criterion for the initial distribution should be able to earn
acquiescence and acceptance of individuals, then other distributive criteria would not satisfy this requirement, given the way Hume imagines that individuals would likely to rationally interact in the hypothetical state of nature.

This brings us to another important point. Here, it can be observed that Hume is not exclusively concerned with making a descriptive point. Hume is also concerned with making a normative point. When Hume says that “some method must be shewn” to “separate” and “distinguish” possessions and assign them as private property to “particular persons”, this statement should not be interpreted simply as stating that the “method” of immediate or present possessions happen to emerge as a distribution criterion for the initial distribution in the course of evolutionary history. Even if, for Hume, this is the case, it should be pointed out that Hume also mentions that individuals would “agree in preferring” this distributive criterion over the others. And, this is why this statement should also be interpreted as justifying the “method” of immediate or present possessions as the distributive criterion that defines the initial distribution. In the *Treatise* Book 3, Part 2, Section 3 and in the *Enquiry* Section III, Hume considers alternative distributive criteria and dismisses them precisely because they would be “pernicious” to the “interests of society” (*EPM* 3.22-26/SBN 192-194). It is precisely on such grounds that Hume rejects alternative distributive criteria, such as that of the original acquisition, strict equality, and desert, as I have shown. For Hume, none of them are appropriate criteria by which “property is determined” and “men’s possessions should be separated”. Therefore, although Hume does not systematically engage in this discussion, Hume’s argumentative strategy to justify the distributive criterion of the immediate or present possessions involves ruling out other distributive criteria based on the premise that such a
criterion, in order to be acceptable, should be one that individuals would acquiesce, agree, or prefer, which is not an implausible premise at all. Hume could be interpreted as providing an argument that could be summarized in two premises and the following conclusion: i) “some method must be shewn” to “separate” and “distinguish” possessions and assign them as private property to “particular persons”; ii) this “method” must be one that “men wou’d easily acquiesce” or that “they wou’d so naturally agree in preferring”; and iii) the only “method” that can satisfy this requirement is the distributive criterion of the immediate or present possessions. In this sense, Hume could be comprehended, in the course of his work, as supporting the rationality of this particular distributive criterion by ruling out all others, insofar as only this rule can satisfy the requirement set by Hume himself.

Now, the question is how the distributive criterion of the immediate or present possessions could be specified. Immediate or present possession could be imagined as associated with the idea of natural distribution. In particular, the immediate or present possession could be thought of as defined by a natural distribution. Buchanan and Stemmer are among the philosophers who support this idea. The natural distribution could be described as follows. Suppose that, in a hypothetical state of nature, a scenario without established moral rules or private property rights, individuals are compelled to fight with each other over shares of a scarce good, since its complete consumption by one or some of them imposes subsistence pressures on all the others.

96 Clearly, my interpretation is not unanimous. Others disagree with it. For instance, Harrison argues that Hume could not have consistently argued for the rationality of the rules of property that we actually have. See Harrison (1981: 86-87).
97 Buchanan argues that the idea of natural distribution is closely related to the theory of property rights as formulated by philosophers of the social contract tradition, in which, for Buchanan, Hobbes and Hume are included. See Buchanan (1975: 31-33), Stemmer (2000: 229-231).
Consequently, each individual is willing to invest efforts to acquire and protect shares of the scarce good from all others until the emergence of an equilibrium. This equilibrium is characterized by a situation in which, given the total efforts in acquisition and protection of shares of the scarce good by each individual, the benefits of the effort investment is equal to the cost that this effort investment requires for each individual involved. In other words, it is no longer worth investing more efforts to fight for further shares of the scarce good, for the costs to acquire them by overcoming others’ defenses are just too high. Notice, however, that, given the specific differences in physical and intellectual qualities, it is expected that the equilibrium resulting from actual or potential conflicts over shares of the scarce good would result in unequal distributions of this good among individuals according to their different capacities. Therefore, if this logic is generalized from a scenario in which a particular good is scarce to a scenario in which most goods are scarce, the kind of equilibrium that is likely to emerge is a natural distribution. In this way, it could be claimed that once the scarce possessions are arranged among individuals through a natural distribution, then an initial distribution based on the distributive criterion of immediate or present possession is established.

4.1.6 Utility and Imagination

The idea that the distributive criterion of immediate or present possession should be defined by a natural distribution is quite attractive, but it is not exactly what Hume argues for. Rather, in the first footnote of the Treatise Book 3, Part 2, Section 3 and in the second footnote of the Enquiry Appendix III, Hume claims that all rules that specify private property rights are defined either by “motives of public utility” or by “imagination”. This implies that other rules that specify private property rights such as occupation,
prescription, accession, and succession, are determined, to some degree, by considerations of public interest and, to some other degree, by imagination (T3.2.3.5/SBN 505). And, with regard to the distributive criterion of immediate or present possession, Hume considers that it is mainly determined by imaginative processes. For Hume, human beings have a natural propensity to make imaginative connections between particular individuals and specific possessions based on principles of association of ideas. Therefore, from a psychological perspective, the distributive criterion of immediate or present possessions becomes the salient solution for the assignment of private property rights to individuals, as it facilitates an easy acquiescence or a natural agreement for a convention that stabilizes possessions. The way in which imaginative associations are useful to solve distributive problems by salience can be summarized, in a simplified way, by the example provided by Hume in the fifth footnote of the Treatise Book 3, Part 2, Section 3:

Suppose a German, a Frenchman, and a Spaniard to come into a room, where there are plac’d upon the table three bottles of wine, Rhenish, Burgundy and Port; and suppose they shoul’d fall a quarreling about the division of them; a person, who was chosen for umpire, wou’d naturally, to shew his impartiality, give every one the product of his own country. (T3.2.3.9/SBN 510)

In fact, it should not come as mysterious that Hume thought, from a psychological perspective, that imaginative associations would allow for a gentle mental transition from individual preferences that each one holds to what they “have long enjoy’d” or to what “has often been employ’d” by them to a collectively recognized rule to solve distributive problems. This point is

---

98 The concept of salience was briefly addressed in Chapter 3. See Lewis (1969: 35), Baier (2010: 36).
99 In fact, a rule to settle distributive problems based on particular associations between claimants and objects is a very intuitive idea. There is a natural propensity to come up with solutions that, more or less, mirror pre-existing relations between objects or individuals. There is also psychological and biological evidence that conflicts over
very important. Recall that Hume is not exclusively concerned with describing the conditions under which immediate or present possessions emerge as a distributive criterion for the initial distribution, but also with justifying it, as I interpret him. And, this justification involves the premise that the distributive criterion should be able to earn easy acquiescence or a natural agreement among individuals. Now, given this premise, it seems that Hume considers that the requirement set by the premise can only be satisfied if the imagination is involved in this process. For Hume, the justification of the distributive criterion must involve the natural propensity of human beings to be drawn to particular imaginative associations and not just pure rational considerations. Presumably, Hume thought that rationality was not as a strong factor as imagination to settle distributive problems from the very psychological perspective of individuals themselves. In fact, this solution fits well with Hume’s naturalistic approach to philosophy. It reflects another instance in which, for Hume, the limits of reason are overcome by instinct and custom, as it is also concluded in the discussions about causal inference and the problem of induction (EHU 5-8 /SBN 43-47; EHU 5.22/SBN 55). Thus, it can be concluded that, for Hume, it is the natural instinct of human beings to recognize, by imaginative processes, the salience of a wide range of relations between particular individuals and specific possessions that makes the possessions are settled by rules or conventions in favor of possessors. And, they are very rational solution for settling potentially violent disputes. Sugden attempts to explore all the possible conscious or unconscious rules or conventions to define the ownership of possessions which are based on prominence, salience, precedents, and analogies. See Sugden (2005: 91-93, 97, 106-107).
immediate or present possessions a rational distributive criterion for the initial distribution.

4.1.7 The Initial Distribution, Evolution, Utilitarianism, and the Social Contract

Given that Hume claims that the immediate or present possessions constitute a suitable distributive criterion to define the initial distribution, it is necessary to comprehend this rule, which specifies the convention for the stabilization of possessions, in light of the three main lines of interpretation suggested so far. Clearly, an initial distributive pattern defined by the immediate of present possessions could be understood in the logic of (E.I.). In fact, Hume is concerned with providing a general description of the emergence of an initial distributive pattern from a hypothetical state of nature on the basis of imaginative associations specifically related to the present or immediate arrangement of possessions. This description could be easily read as a methodological account of the evolution of private property rights, as (E.I.) would state. However, although the emergence of this initial distribution the immediate or present possessions can be accurately described by (E.I.) as a result of the collective psychology of human beings, it can capture only part of the general description, as it overlooks Hume’s normative reasons for the selection of this distributive criterion and not any other. So, the account of the distributive criterion could be, in part, understood in light of (E.I.), even if this line of interpretation cannot capture Hume’s general description in its entirety.

As far as (RU.I.) is concerned, the situation is quite different. First, the requirement set by Hume’s premise does not seem to fit well with (RU.I.). Recall that, according to the premise, a plausible distributive criterion should
be able to earn easy acquiescence and natural agreement among individuals with regard to the particular distributive pattern that it is designed to establish. Hume does not mention that this rule, to which individuals acquiesce, is designed to maximize the sum of happiness or welfare of society. Instead, the requirement set by the premise is only that this rule is one that individuals would “easily acquiesce” or “naturally agree in preferring”. Second, even if it is the case that this distributive criterion is able to maximize the sum of social welfare by specifying the rules of private property in a particular way, Hume would still argue that this rule is determined by imaginative associations. Recall that Hume considers that all rules that specify private property rights are defined either by motives of public utility or by imagination. And yet, Hume does not brings public utility as a relevant consideration to determine the distributive criterion. Consequently, it is implausible to argue that (RU.I) qualifies as a plausible interpretation to understand this rule.

It seems that the requirement set by Hume’s premise is more directly related to (C.I.). To the extent that a distributive criterion for an initial distribution should be one that individuals could easily acquiesce or naturally agree, (C.I.) emerges as the most obvious interpretative candidate for understanding the logic of this rule. This is because (C.I.) requires that rules of justice could be thought of as earning unanimous assent. However, for Hume, the distributive criterion of the immediate or present possessions, according to which possessions should be firstly distinguished, sorted out, and assigned to each individual, is determined by imagination. The natural and collective propensity of human beings to make psychological and imaginative associations between particular individuals and specific possessions favors the selection of a salient solution for the initial distributive
pattern in society. Hence, on grounds of salience, immediate or present possessions becomes the distributive criterion for the initial distribution to which individuals would easily acquiesce or upon which they would naturally agree in preferring. For Hume, salience is, then, the basis of his conception of the initial distributive pattern in society.

Now, the challenge is to understand how (C.I.) could account for this rule of Hume’s theory of property. Gauthier, who supports (C.I.), interprets Hume’s appeal to salience as an entirely legitimate coordinating device for the selection of a particular system of private property rights, even if it might be questioned whether this device is really the best one for defining the initial distributive pattern in society. This is because, in Gauthier’s view, Hume takes individuals as concerned with agreeing on some system of private property rights than with choosing among other possible systems, as the “expected differential advantages” between a system of private property rights and no system at all are very large, while the “expected differential advantages” between two systems are comparatively small. However, Gauthier’s real point is that Hume’s appeal to salience as a coordinating device for the selection of a particular system of private property is still based on the logic of (C.I.), for it takes into account the interests and preferences of each individual, who would easily acquiesce or naturally agree with such salient solution in comparison to other costly bargaining procedures.100

However, there are limitations to this interpretation. While Gauthier is correct in pointing out that, for Hume, individuals would prefer to establish some system of private property rights than not to establish no system at all, it does not follow that, for Hume, other systems of private property rights are

---

entirely indifferent to individuals. If this were the case, Hume would not care to discuss other distributive criteria, such as the original acquisition, equality, and desert. In fact, Hume rejects all other criteria, for they are either detrimental to private and public utility or too counterintuitive to the imagination, given the circumstances in which individuals find themselves. But, the real point addresses Hume’s appeal to salience as a coordinating device to settle the initial distributive pattern. It should be noticed that some salient solutions, in particular, Hume’s, are not rational in the sense required by (C.I.). As Gauthier himself points out, salience plays a purely coordinative role. It operates as a device that, by its uniqueness, stands out from the rest in a particular way to guide individuals towards certain actions or state of affairs. But, in Hume’s account, the distributive criterion of immediate or present possession is selected without a suitable rational endorsement by each individual. Instead, it is selected only by imaginative associations. In fact, given the psychological influence of salience, it might turn out that individuals could choose a distributive criterion that is harmful to their own interests. Unlike in (C.I.), where the unanimous assent is a result of a convergence of interests of all upon a rational consideration, the unanimous assent earned by salient solutions can only be consolidated upon psychological grounds. According to the requirement set by (C.I.), salient solutions cannot be regarded as rational, for they do not necessarily need to take the interests and preferences of each individual into account in the procedure of selecting a particular distributive pattern in society. In this sense, the distributive rule suggested by Hume, which is introduced as a salient solution for a particular distributive problem, is not based on interested recognition, which means that it cannot give rise to the kinds of interested obligations required by (C.I.). Notice that I am not claiming that Hume’s account of the distributive criterion
of the immediate or present possession could not, in any sense, be considered rational. I am just claiming that Hume’s account of the distributive criterion of the immediate or present possession is not rational in the sense required by (C.I.). And this seems to be an insurmountable difficulty for (C.I.) in the description of Hume’s theory of justice.

4.2 The Transference of Property by Consent
Hume’s theory of property can only be fully described if the rule of transference of property by consent is addressed as well. Hume does not take long to explain this rule, as it is evident by the discussion in the Treatise Book 3, Part 2, Section 4. On the one hand, the rules that specify private property rights such as, the immediate or present possession, occupation, prescription, accession, and succession, are introduced to regulate private property rights with regard to the initial distribution and acquisition of possessions in general. On the other hand, the rule of transference of property by consent regulates private property rights with regard to the optimal distribution of possessions by allowing the acquisition of private property by mutual exchange or trade. The rationale for a rule of transference of property by consent is quite straightforward. The initial distribution does not necessarily ensure that possessions are arranged in such a way that individuals will be able to satisfy their needs and desires. A rule is necessary to allow individuals to acquire possessions to satisfy their needs and desires, for, in the absence of it, they might engage in violent appropriations of possessions. And the rule of transference of private property by consent is appropriate for this purpose. However, Hume interprets this scenario as far more complex in comparison to the one depicted in the hypothetical state of nature. Now, Hume seems to assume that individuals are defined by a utility function and a production
function. Not only are they defined by a set of particular needs and desires, but also by a set of productive capacities to manufacture specific products or perform specific services.\(^\text{101}\) The rule of transference of private property by consent operates as the origin of a market structure by which individuals, who “produce different commodities” and “are “fitted for different employments”, are able to engage in “mutual exchange” and “commerce” (T 3.2.4.1/SBN 514).

As far as the interpretation of the rule of transference of property by consent is concerned, it does not seems that it stands in contradiction with any of the lines of interpretation suggested so far. This rule seems to satisfy the requirement set by (E.I.) since it is introduced by Hume as one of the several rules that gradually specify the rule of stability of possessions until the emergence of a complete system of private property rights. This rule seems to satisfy the requirements set both by (RU.I.) and (C.I.) as well. One the one hand, it seems to be able to maximize the sum of social welfare by operating as a general rule that allows for an optimal distribution of possessions to take place, which is capable of satisfying the greatest amount of utility functions. On the other hand, it also seems to be able to earn a rational unanimous assent of individuals, since no one would likely to end up worse off by the establishment of a rule that depends on the expression of consent. In this sense, Hume’s description of the rule of transference of property by consent does not necessarily speak for any interpretative claim in particular. In fact, it seems to

---

\(^\text{101}\) Such concepts are developed by Buchanan. See Buchanan (1975: 72).
be such a general rule for social coexistence that it could be comprehended by any line of interpretation.

4.3 Utility and Inequality

4.3.1 Standard Interpretation

It could be argued that, as far as distributive justice is concerned, this account is all that Hume has to offer on the subject. Hume’s theory of justice would be just a matter of defining the initial distribution and safeguarding voluntary exchanges among the consenting parties in society. As a result, Hume’s theory of justice could be thought of as generating large and harmful social inequalities. These worries are commonly related, and not without good reasons, to Hume’s views in the *Enquiry*. More specifically, I tend to believe that this interpretation is based on the following passage in the *Enquiry* Section III:

> But historians, and even the common sense, may inform us, that, however specious these ideas of perfect equality may seem, they are really, at bottom, impracticable; and were they not so, would be extremely pernicious to human society. Render possessions ever so equal, men’s different degrees of art, care, and industry will immediately break that equality. Or if you check these virtues, you reduce society to the most extreme indigence; and instead of preventing want and beggary in a few, render it unavoidable to the whole community. The most rigorous inquisition too is a requisite to watch every inequality on its first appearance; and the most severe jurisdiction, to punish and redress it. But besides, that so much authority must soon degenerate into tyranny, and be exerted with great partialities; who can possibly be possessed of it, in such a situation as it here supposed? Perfect equality of possessions, destroying all subordination, weakens extremely the authority of magistracy, and must reduce all power nearly to a level, as well as property. (*EPM* 3.26/SBN 194)

In view of this passage, it is comprehensible that many scholars have interpreted Hume as indifferent to issues of social inequalities or

---

redistribution. Instead, Hume’s theory of justice would be focused on defining the rational conditions for the initial distributive pattern in society, and specifying acceptable rules and procedures for gradual changes in its distributive order, such as, for instance, rules of mutual exchange and trade. In sum, the idea is that the establishment of the initial distribution should result from the collective psychology of individuals in a hypothetical state of nature and distributive changes should result from the voluntary exchanges of individuals in a social state.

4.3.2 Social Inequalities and Negative Effects
I argue that this interpretation is incorrect. It is false that Hume is indifferent to issues of social inequality or redistribution. Hume is not opposed to redistributive rules to specify private property rights in case of large and harmful social inequalities, provided that appropriate circumstances are obtained. In fact, it is surprising that this interpretation could have prevailed for so long among Hume scholars. Clearly, relevant passages of Hume’s work were overlooked by this interpretation. For instance, in the essay *Of the Populousness of Ancient Nations* (1741-42), when Hume discusses successful policies of ancient states that resulted in population growth and prosperity, we come across with the following passage.104

Before the encrease of the ROMAN power, or rather till its full establishment, almost all the nations, which are the scene of ancient history, were divided into small territories or petty commonwealths, where of course a great equality of fortune prevailed, and the center of the government was always very near its frontiers. This was the situation of affairs not only in GREECE and ITALY, but also in SPAIN, GAUL, GERMANY, AFRIC, and a great part of the Lesser ASIA: And it must be owned, that no institution could be more favourable to the propagation of mankind.

104 It is important to notice that Hume considers population growth as an indication of the prosperity of a state. While it is clear that not every state with a large population is prosperous, I would tend to agree with Hume that there is a particular correlation between states that provide suitable conditions for population growth and prosperity.
For, though a man of an overgrown fortune, not being able to consume more than another, must share it with those who serve and attend him; yet their possession being precarious, they have not the same encouragement to marry, as if each had a small fortune, secure and independent. [...] Where each man had his little house and field to himself, and each county had its capital, free and independent; what a happy situation of mankind! How favourable to industry and agriculture; to marriage and propagation! The prolific virtue of men, were it to act in its full extend, without that restraint which poverty and necessity imposes on it, would double the number every generation: And nothing surely can give it more liberty, than such small commonwealths, and such an equality of fortune among citizens. (PA 46/401)

The same point is brought in the essay *Of Commerce* (1741-42):

Not to mention the great equality of fortunes among the inhabitants of the ancient republics, where every field, belonging to a different proprietor, was able to maintain a family, and rendered the number of citizens very considerable, even without trade and manufactures. (Co 8/259)

Notice that, despite the passage in the *Enquiry*, there is textual evidence to indicate that Hume held a very favorable view of more egalitarian social arrangements. Hume argues that “a great equality of fortune” would provide better conditions for “each man”, so that “industry and agriculture” would be fostered and “marriage and propagation” would be encouraged. In fact, Hume goes as far as to say that, in comparison to the situation of the modern states of his time, greater equality in society, as it was typical in ancient republics, would be a preferable policy for the population growth and prosperity of nations. This is evident from the following passage in *Of the Populousness of Ancient Nations*:

It must be owned, that the situation of affairs in modern times, with regard to civil liberty, as well as equality of fortune, is not near so favourable, either to the propagation or happiness of mankind. EUROPE is shared out mostly into great monarchies; and such parts of it as are divided into small territories, are commonly governed by absolute princes, who ruin their people by a mimicry of the great monarchs, in the splendor of their court and number of their forces. SWISSERLAND alone and HOLAND resemble the ancient republics; and though the former is far from possessing any advantage either of soil, climate, or commerce, yet the number of people, with which it abounds, notwithstanding their enlisting
themselves into every service in EUROPE, prove sufficiently the advantages of their political institutions. (PA 50/402-403)

This passage suggests that Hume makes a normative claim in favor of more egalitarian societies in general. For Hume, egalitarian and free societies provide better conditions for the “propagation or happiness of mankind” than unequal and despotic societies. So, the interpretative claim that Hume is indifferent to issues of social inequality seems to be implausible in light of the textual evidence. In addition, the passages also provide the reasons why Hume believes that a more egalitarian social structure is desirable in a society. In a further passage in Of Commerce, the reasons are presented by Hume most clearly:

A too great disproportion among the citizens weakens any state. Every person, if possible, ought to enjoy the fruits of his labour, in a full possession of all the necessaries, and many of the conveniencies of life. No one can doubt, but such an equality is most suitable to human nature, and diminishes much less from the happiness of the rich than it adds to that of the poor. It also augments the power of the state, and makes any extraordinary taxes or impositions be paid with more cheerfulness. Where the riches are engrossed by a few, these must contribute very largely to the supplying of the public necessities. But when the riches are dispersed among multitudes, the burthen feels light on every shoulder, and the taxes make not a very sensible difference on any one’s way of living. […] Add to this, that, where the riches are in few hands, these must enjoy all the power, and will readily conspire to lay the whole burthen on the poor, and oppress them still farther, to the discouragement of all industry. (Co 17-18/265)

This last passage is central in this discussion because it presents a very plausible argument. Notice that, in the previous passages, Hume has attempted to make a case for greater equality among individuals by stressing the idea that, provided that each family unit is given equal shares of fortunes, they will be able to live safely and contribute steadily for the population growth of the society. But it is difficult to see how this argument relates to social equality in the first place. Social equality is not necessarily required either for securing
the well-being of individuals or for the subsequent population growth. Rather, the argument seems only to conclude that, provided that each family unit is given enough for their living, they will thrive. But, social equality, in itself, has nothing to do with that. In this last passage, however, the point is slightly different. Now, Hume is concerned with pointing out the pernicious effects of social inequality in a whole social perspective. The claim is that that social inequality undermines or, as Hume puts it, “weakens” the state. Hume’s argument to support this claim sets up the following scenario. Suppose that, in order to provide its citizens with public goods, the state must tax them, and the burden of taxes must somehow be socially borne. In this case, there are two possible alternatives. On the one hand, the burden of taxes could fall much more heavily on the rich, since they have more resources available to contribute with. On the other hand, the rich could, by seizing political power, for instance, manage to avoid the burden of taxes and make it fall much more heavily on the poor, who already lack the resources available to do so. In the first alternative, it is the rich who are most directly harmed in their interests, while, in the second alternative, it is the poor who are most directly harmed in their interests.

But, as Hume points out, the expected utilities of the individuals of the two relevant social groups decrease at very different rates depending on which alternative is selected, since the rich, because they have comparatively more resources than the poor, are able to bear the social burdens much more easily. Even if the difference in the rate of decrease in expected utilities of the individuals of the two social groups might constitute sufficient reason for the selection of the first alternative, Hume’s point is intended to be even stronger. Hume claims that, in the second alternative, a scenario that tends to reinforce social inequalities, individuals of both relevant social groups will likely to end
up worse off in the long run. As Hume points out, a whole set of negative effects follow from this unequal state of affairs, such as the hindrance of the enjoyment of one’s “fruits of his labor”, the harm of “one’s way of living”, and the discouragement of “all industry”. And, such social problems, generated by a scenario of sensitive social inequalities, weaken the social fabric itself. Therefore, for Hume, even if the selection of a more egalitarian social structure may cause some losses for the group of rich people in the short run, it is still be preferable than the selection of an unequal social structure in the long run, for the negative effects resulting from such a state of affairs are far more harmful to everyone in a broader social perspective.

Scholars have interpreted Hume as indifferent to issues of social inequality or redistribution. Based on the aforementioned passages, I have argued that Hume makes normative claims in favor of more egalitarian societies in general. What should we make of the two conflicting interpretations? It is important to put things in perspective. Hume’s problem was not with social equality itself, but with the enforcement of egalitarian measures in society. For Hume, the enforcement of a “rule of equality” is not only “highly” useful, but it is also, to some extent, “practicable”. In a passage of the Enquiry Section III, historical evidence from Sparta and Rome is brought by Hume to support this claim, as it follows:

It may appear withal, that the rule of equality, as it would be highly useful, is not altogether impracticable; but has taken place, at least in an imperfect degree, in some republics; particularly that of Sparta; where it was attended, it is said, with the most beneficial consequences. Not to mention that the Agrarian laws, so frequently claimed in Rome, and carried into execution in many Greek cities, proceeded, all of them, from a general idea of the utility of this principle. (EPM 3.25/SBN 194)

The real concern is more about how the enforcement of a rule of equality would be carried out in society. In this sense, two worries are particularly
pressing. First, it could be enforced recklessly. This is suggested by Hume in the discussion about the Levellers, a political group that pushed for a more equal distribution of property among peasants and lords in the 17th century in England. For Hume, the group’s strong religious background could incite superstitious beliefs in the population, which, at the time, were extremely dangerous in the political context (EPM 3.24/SBN 193). Hume believed that the enforcement of a rule of equality should be carried out gradually. It should take into account the “nature” and “situation of man” in society. Second, it could be directed toward the wrong goals. It is important to notice that Hume does not endorse a perfectly equal society. For Hume, social equality is not desirable in itself. A rule of equality is only an efficient means for promoting the public interest. Just as social inequalities are, at times. This is why Hume considers the enforcement of a strict rule of equality as pointless and socially dangerous. Thus, the real concern is more with such a rule of equality that may prove to be too socially costly, not only in the enforcement of the rule itself, but also in the goals that this rule is intended to accomplish.

Therefore, from the interpretative point of view, it should be concluded that Hume is not indifferent to issues of social inequality. In fact, Hume would go as far as to recommend changes in the distributive order of deeply unequal societies to promote the public utility, for the inconveniences of this state of affairs are potentially harmful to society. However, it should be
recalled that Hume never loses sight of the adequacy of this distributive change to achieve this goal.

4.4 The Redistributive Rule
4.4.1 Hume’s Theory of Property
Hume’s theory of property, which forms part of Hume’s theory of justice, addresses the rules that specify private property rights. So far, the rules that specify the private property rights introduced by Hume could be summarized as such. First, the initial distribution is defined based on the distributive criterion of immediate or present possession in the context of a hypothetical state of nature. This distributive rule, which involves the psychological propensity of human beings to recognize salient solutions and make imaginative associations between particular individuals and specific possessions, defines the distribution of possessions among individuals in the convention for the stabilization of possessions and assignment of private property rights. Rules of occupation, prescription, accession, and succession are introduced by Hume as well, as possessions need to be separated and distinguished by considerations of public interest. Second, the rule of transference of property by consent is introduced by Hume to allow voluntary exchanges to take place among the consenting parties in society. It allows for distributive changes from the initial distributive pattern in society in such a way that the acquisition and allocation of possessions are determined according to the needs and desires of individuals by mutual exchange or trade. Third, Hume has a real concern with possible sensitive social inequalities that could emerge from the current distributive order in a social state. So, as long as egalitarian measures are not too socially costly to be enforced, Hume would recommend them to be enforced with the purpose of correcting this unequal
state of affairs. However, even if Hume had a real concern with possible sensitive social inequalities, for some reason, he did not discuss any further the enforcement of a possible redistributive rule. Perhaps, Hume did not even consider this as a problematic issue. Consequently, a more careful and detailed reconstruction of a possible redistributive rule in Hume’s theory of justice would have to be resort to other concepts of Hume’s moral theory. But, at the same time, Hume already provides us with some points that should be taken into account in this reconstruction as well. And, even if a redistributive rule can only be enforced in a political context, the explanation of this rule would complete the description of Hume’s theory of property, which forms part of Hume’s theory of justice, as, for Hume, questions of justice are fundamentally questions of private property. Therefore, I intend to explain how Hume could account for a possible redistributive rule and how this rule should be properly interpreted.

4.4.2 The Natural Virtue of Equity

It is useful to start this discussion by addressing two interpretations that have been suggested to explain how Hume could account for a possible redistributive rule. In this way, it is possible to distinguish between plausible and implausible alternatives. The two interpretations are suggested by Annette Baier and Mark Yellin. The interpretative claim made by Baier states that distributive changes would be better explained by the natural virtue of equity in Hume’s theory of justice. Baier argues that, once a convention has been agreed upon by individuals to establish a set of rules of justice and to determine the compliance with the pattern of behavior defined by it, the

---

105 See Baier (2010), Yellin (2000).
expansion of this basic structure to incorporate additional rules of justice, such as those that regulate fair return to labor, equal opportunities, and social justice, depends on a common and innate disposition of human beings to behave equitably. The point is that the establishment of a basic structure of justice contributes to the flourishing of the natural sentiment of equity in human beings, which, among other things, could guide redistributive actions in society. Notice that this close relationship between the artificial convention of justice, on the one hand, and the natural virtue of equity, on the other, reflects the interpretative claim put forward in the first chapter and shows not only how the basic structure of justice can be expanded but also how the distributive pattern of society could be rearranged towards a more egalitarian one.\(^\text{106}\)

Although Baier is right in pointing out that Hume considers equity to be a natural virtue, which is literally asserted in the *Treatise* Book 3, Part 3, Section 1, it is a very inadequate concept for explaining a possible redistributive rule. I concede that it is possible for individuals to make substantial contributions to society motivated solely by the virtue of equity, even if the assumptions that it is a natural virtue or that it is a suitable disposition to orient possible distributive changes in society are highly questionable. However, I think that the interpretation that explains distributive changes in society based on mere displays of a disposition of individuals to be equitable with each other, however deep-rooted this disposition may be, is very inadequate.

Natural virtues are kinds of common and innate dispositions that motivate individuals to behave in a way that is likely to convey a sentiment of

---

\(^\text{106}\) See Baier (2010: 78-79, 92, 97-98).
approbation in spectators from the impartial stance of the general point of view. For Baier, changes in the distributive pattern in society would be morally approved in that they could be motivated by a natural propensity of each individual to be equitable with others and, in principle, this disposition could be recognizable by all. This description is not very plausible. For Hume, questions of distributive justice involve partial conflict problems over possessions. In general, Hume considers that individuals are primarily self-interested. They would be willing to act to the benefit of others only if they expect some advantage in return. They would not like to be deprived of what they possess without convincing reasons that it would somehow turn out to be advantageous for them in the future. So, it is difficult to imagine that the natural disposition of equity could overcome the equally natural disposition of self-interest that induces individuals to keep and extend their possessions as much as possible. Redistribution addresses the more fundamental issues of individuals. It refers not only to the natural sentiments, passions, and dispositions of human beings, but also, and most importantly, to their relevant interests. So, even if the natural virtue of equity can play an important role in orienting distributive changes in society, this natural virtue could hardly qualify as the appropriate foundation of a possible redistributive rule, as Baier intends. In fact, it is not necessary to rely on the natural virtue of equity as the main concept to settle problems of distributive justice, as Hume provides us with other conceptual resources to explain how a possible distributive rule could be carried out.

4.4.3 Indirect Utility and Rule-Utilitarianism

The interpretative claim made by Baier draws attention to the natural virtue of equity as a factor of Hume’s explanation of a possible redistributive rule.
However, it is Yellin’s interpretation that makes the most important contribution to a rational reconstruction of a possible redistributive rule in Hume’s theory of justice. The interpretative claim made by Yellin is that a possible redistributive rule would be better explained by the notion of indirect utility. By the notion of indirect utility, Yellin means simply the property of things, such as virtues, practices, rules, and institutions, to have the tendency to be more widely advantageous in the long run, even if they may be disadvantageous in the short run. The most relevant applications of the notion of indirect utility by Hume are the cases of the convention of justice and the government, as both require the establishment of a set of rules that prescribes mutual restraint of each one’s utility-maximizing behavior, for the collective action in accordance with this set of rules proves to be conducive to greater future gains for them. Yellin argues that the notion of indirect utility in Hume should be interpreted as normative guidance for individuals in the maintenance or gradual reform of social and political institutions. In this sense, in light of the notion of indirect utility, the assessment of social and political institutions must take into account the extent to which, on the one hand, Hume’s conservatism with regard to sudden and radical social changes or, on the other hand, Hume’s evolutionary perspective with regard to gradual and steady institutional improvements, should prevail. And, for Yellin, Hume would endorse the claim that social and political institutions would be designed to, among other things, enforce a possible redistributive rule to correct sensitive social inequalities for reasons of indirect utility, for this kind of distributive change could maximize the overall sum of happiness or welfare.
of society. Therefore, the interpretative claim made by Yellin is that a redistributive rule would be better explained by the notion of indirect utility.107

Yellin’s interpretative claim of Hume’s theory of justice raises a very important point. Notice that it portrays Hume as making the claim that a possible redistributive rule, based on the notion of indirect utility, must be conceived in a consequentialist logic. In this sense, this interpretation seems to commit Hume to (RU.I.) with regard to the possibility of a redistributive rule to specify private property rights. Therefore, given this interpretative point, it is important to analyze the example brought by Hume in the passage in *Of the Populousness of Ancient Nations* and in *Of Commerce*. Recall, once again, that, for Hume, sensitive social inequalities are socially harmful, as it is claimed in both texts. Given that social burdens could be borne very differently depending on how they are distributed between the relevant social groups in particular social arrangements, individuals of the relevant social groups, in particular, rich and poor people, may feel compelled to try to avoid contributing with their fair share. But, for Hume, the particular social arrangement that deepens social inequalities is more likely to be socially harmful. It is under such circumstances that the notion of indirect utility is useful. Hume points out that, on certain occasions, sensitive social inequalities could be potentially harmful to the interests of both rich and poor people in the long run, for they prevent the enjoyment of one’s “fruits of his labor”, harm “one’s way of living”, or, even worse, discourage “all industry” in society. In this sense, for Hume, both rich and poor people would have to consider the indirect utility of possible social arrangements, as they could give

rise to sensitive social inequalities, which are likely to be socially harmful in the long run.

Now, the question is how a possible redistributive rule is supposed to be interpreted. Here, once again, Hume does not seem to draw a precise distinction between the concepts of public utility and private utility in the argument. At times, the emphasis seems to be on the distributional effects on individual interests, while, at other times, the public interests seem to be stressed with regard to possible distributive changes. For instance, in Of the Populousness of Ancient Nations, when Hume states that where “each man” had a “little house and field to himself”, “mankind” would enjoy a state of happiness, it seems that Hume is concerned with the individual interests of each person under circumstances of a greater “equality of fortune”. Whereas in Of Commerce, Hume seems to turn his attention to the negative effects of sensitive social inequalities to the public interests, as it is indicated by the distributional effects on weakening or strengthening the “power of the state”. This raises a very important interpretative point. On the one hand, Hume’s concern with the individual interests seems to resemble support for (C.I.), on the other hand, Hume’s concern with the public interests seems to resemble support for (RU.I.). Now, as I interpret it, Hume is making a more rhetorical than a factual statement in the passage in Of the Populousness of Ancient Nations. In Of Commerce, Hume adopts a more realistic point of view. Hume argues that “every person, if possible, ought to enjoy the fruits of his labour” and “equality is the most suitable” condition for this purpose. Here, unlike in the rhetorical statement, Hume introduces a modality to indicate that he is not entirely committed to the thesis that redistributive changes towards a more egalitarian state of affairs should contemplate each and every individual. Instead, a possible redistributive rule should aim at this goal only as far as
possible. Here, once again, Hume seems to focus on the perspective of the global advantages afforded by a redistributive rule, for the claim is that redistributions are mainly intended to strengthen the state by creating, as much as possible, a more egalitarian state of affairs. This perspective favors (RU.I.). Therefore, according to textual evidence, in Hume’s theory of justice, a possible redistributive rule should be interpreted in terms of (RU.I.) and should be carried out in society as such, since the public utility, which is supposed to be favored by a more egalitarian distributive order, is not built as a function that equally satisfies the private utilities of all individuals by Hume.

In this sense, it is important to observe that Yellin’s interpretation makes the most important contribution to a rational reconstruction of a possible redistributive rule in Hume’s theory of justice. Not only is this interpretation able to better explain why relevant social groups could select, based on the notion of indirect utility, social arrangements oriented towards a more egalitarian distributive order in society, but it is also able to most accurately describe the idea conveyed by the textual evidence about a possible redistributive rule to specify private property rights in Hume’s theory of justice, namely, in terms of (RU.I.).

4.4.4 Redistribution and Contractarianism

The textual evidence in the Essays indicates that a possible redistributive rule in Hume’s theory of justice should be interpreted in terms of (RU.I.). However, it is possible to articulate a description of a possible redistributive rule, based on the similar arguments brought by Hume, which fits (C.I.). Note that, in order for redistributive changes to be appropriately explained from the perspective of (C.I.), the concept of public utility should be defined in terms of the expected utilities of each individual. For Hume, the particular
distribution of social burdens in society can contribute to the decrease in the expected utilities of individuals of either the relevant social group of rich or poor people depending on how the displacement of the burdens of social contributions falls on either one of them. However, Hume goes on to claim that, among the possible distributions of social burdens, the one that generates sensitive social inequalities are potentially more harmful to individuals of both relevant social groups in the long run. The conclusion is that sensitive social inequalities should be avoided in general. In this sense, Hume would be inclined to consider a possible redistributive rule to minimize sensitive social inequalities, provided its enforcement is not too socially costly. Had Hume assumed that a redistributive rule should be based on the expected utilities of each individual, it could be adequately accounted for from the perspective of (C.I.). I will attempt to prove that such a description of this subject is possible by briefly surveying the suggestions of James Buchanan, John Rawls, and R. B. Braithwaite.

Buchanan holds that redistribution should be carried out as a collective income insurance plan.\textsuperscript{108} In other words, a rule that prescribes a degree of income redistribution among members of relevant social groups should be enforced by a collective intervention. For Buchanan, the justification for income redistribution depends on the succession of periods of good and bad fortunes of individuals in terms of income revenue. It is assumed that income redistribution operates in such a way that it increases the individual utility functions when it adds to the income during periods of bad fortune and subtracts from the income during periods of good fortune. If individuals are fortunate enough to actually go through periods of good fortune, they can

\textsuperscript{108} See Buchanan (1965: 193-197).
anticipate periods of bad fortune by purchasing a private insurance plan. However, as Buchanan puts it, at the stage of the constitutional choice, in which rational individuals must select the institutions that constitute the government, they lack information about how the succession of periods of good and bad fortune will unfold in the course of their lives. Given that income is the primary resource on which all life planning depends, each individual has incentives to maximize their utility function by choosing to establish a redistributive rule to be collectively enforced by the government. Thus, for Buchanan, the choice to establish a collective income insurance plan with the purpose of reducing individual risks in face of the uncertainty regarding future income revenues is a rational one for each individual, especially if everyone is included in this plan.

Rawls holds that redistribution should be carried out in accordance with the difference principle. This is the principle that properly defines Rawls’s conception of distributive justice. This principle states that social and economic inequalities are arbitrary unless they are arranged in such a way as to contribute to the greatest expected social benefits for the least advantaged. And, for Rawls, the justification of the difference principle is, in some ways, very similar to Buchanan’s, for it involves reflections about lack of information, uncertainty, and risk aversion. Rawls thinks that the justification of the difference principle should make reference to the veil of ignorance. The difference principle should be regarded as justified insofar as it could be selected by rational individuals who imagine themselves as choosing, from a specific choice procedure, what their conception of distributive justice should be. In other words, rational individuals would be required to choose, once and

109 See Rawls (1971: 83)
for all, the principles of distributive justice that inform the main social and political institutions of society. In turn, redistributions would be carried out by social and political institutions in order to maximize the expected benefits of the least advantaged. Here, the specific choice procedure that Rawls has in mind is precisely the veil of ignorance. The veil of ignorance is a scenario in which rational individuals lack information about their natural talents or abilities and their position in society. In this way, by depriving individuals of this kind of information in the choice procedure, arbitrary advantages and disadvantages among individuals are ruled out. And, in such circumstances of lack of information, Rawls supposes that rational individuals would be compelled to make a conservative decision in the selection of the principles of distributive justice that they could expect to benefit them in the worst-case scenario. Since the difference principle would ensure the greatest expected social benefit for the least advantaged, it is assumed that it would be selected by rational individuals under a veil of ignorance.

However, it is Braithwaite who suggests a description of a possible redistributive rule which is most closely related to Hume’s account. For Braithwaite, redistributions must primarily address the distributive rearrangement of the cooperative surplus in society. This situation is typically characterized by a partial conflict over the redistribution of the sum of the socially produced resources. It is neither a situation of pure conflict, in which the interests are diametrically opposed, nor a situation of pure coordination, in which the interests are perfectly convergent. Instead, it is a situation in which conflicting claims exist, but, nonetheless, there is still the possibility of accommodating them in a more advantageous state of affairs than the previous one. Thus, the redistribution of the cooperative surplus must conciliate claims
that individuals raise based on their expected utilities. The fact that redistributions must reflect the expected utilities of individuals of the relevant social groups implies that they should be carried out by means of an agreement among the concerned parties to choose a single joint strategy for action. Therefore, a redistributive rule should not be imposed on individuals by means of a utility calculation. Rather, decisions regarding distributive rearrangements of the cooperative surplus in society should refer to the interests of all parties concerned in this process.

By definition, redistributions must somehow change the distributive arrangement in society towards a more egalitarian one. In other words, they are considered as successful as they can be effective in narrowing the gap between individuals or relevant social groups. In an agreement for the choice of a suitable redistributive rule, this amounts to defining the relative advantage of one individual or relevant social group over the other in the bargaining process of the redistribution of the cooperative surplus. Since such an agreement must reflect the individual expected utilities, then the relative advantage of one individual or relevant social group over the other reflected in the bargaining process must be defined precisely by the preferences of individuals. In other words, the degree of redistribution of the cooperative surplus in society must reflect a convergence of interests of both parties involved.

However, it should not be simply assumed, without further remarks, that individuals would be willing to share with others part of whatever they have acquired so far without any reserve, even if this share is minimal. The logical supposition is that, on the one hand, some would be willing to claim

---

110 See Braithwaite (1955: 22).
as much as they can in this process, while, on the other hand, others would be willing to make no concessions whatsoever. Common ground is needed so that both conflicting claims can converge. And, it is Hume’s scenario that presents a clearer example of a possible common ground. As Hume points out, sensitive social inequalities could be potentially harmful to the interests of both rich and poor people in the long run. They are likely to be socially harmful, for they tend to prevent the enjoyment of one’s “fruits of his labor”, harm “one’s way of living”, and discourage “all industry”. Under such circumstances, both rich and poor people would have incentives to enter into an agreement for the redistribution of the cooperative surplus. In other words, they are supposed to agree on a single joint strategy for action not on the basis of what seems to be advantageous for each one of them now, but rather on what is actually more advantageous for each one of them in the long run. The agreement on a possible redistributive rule should be conceived by the concerned parties from the standpoint of long-term benefits. Thus, if a deeply unequal state of affairs is, on certain occasions, harmful for the interests of both rich and poor people, it would not be rational for individuals of the group of rich people to choose to make no concessions under such circumstances, for this disadvantageous situation would simply persist for the worse of everyone. For this reason, it is difficult to see why the group of rich people could rationally refuse to agree with a possible redistributive rule whenever sensitive social inequalities prove to be socially harmful. In this sense, for Braithwaite, agreements on a possible redistributive rule, which are constituted by the claims and based on the expected utilities of individuals
from antagonistic groups, can become relatively advantageous to the most vulnerable group, and this for the overall advantage of all.

Notice that Buchanan, Rawls, and Braithwaite have all developed accounts of redistribution from the perspective of a contractarian theory. The accounts provided by Buchanan and Rawls are more focused on the rationality of individuals in ideal circumstances characterized by a lack of information, uncertainty, and risk aversion. The account provided by Braithwaite bears a strong resemblance to Hume’s in that they are more focused on the rationality of individuals in factual circumstances characterized by a state of affairs that represents a common ground for conflicting claims. However, the idea is very clear. A possible redistributive rule could only be regarded as justified if it carried out by means of a hypothetical agreement among the parties concerned, insofar as they are imagined as ideally rational agents who take into account their long-term interests in the process of deliberation. Such agreements, which are meant to reflect a convergence of interests of the parties, should be the ground for rearrangements of the distributive pattern in society. This is the way in which redistribution is properly explained from the perspective of a contractarian theory. And, even if Hume did not articulate this kind of description, it can be noticed that the ideas laid down by his account could easily be developed into a contractarian account of redistribution. It is Hume’s slight ambiguity in relation to the concepts of public and private
utility that ultimately settles the controversy between (RU.I.) and (C.I.) to describe his account of redistribution.
In this chapter, I have examined Hume’s theory of property, which is a central part of Hume’s theory of justice. In particular, I have addressed the rules that constitute Hume’s theory of property as they feature in the *Treatise* and in the *Enquiry*. They are meant to specify the general rule of stabilization of possessions and private property rights, which were the first rules introduced by the convention of justice. In this sense, rules of immediate or present possession, occupation, prescription, accession, and succession must be introduced to specify how possessions can be separated and distinguished in order to be further acquired as private property rights. Such rules are primarily based on a combination of the psychological propensity of human beings to make imaginative associations between particular individuals and specific possessions and the considerations of public utility. I have also addressed the rule of transference of property by consent, which is also introduced by Hume as part of his theory of property. In establishing a rudimentary market structure, this rule defines the acquisition of private property rights by voluntary exchanges or trade. Finally, in the *Essays*, I have also analyzed Hume’s account of a possible redistributive rule, which is meant to mitigate sensitive social inequalities that could emerge from the given distributive order. Since that Hume considers that sensitive social inequalities might, on certain occasions, be socially harmful, it is necessary to introduce a rule to specify the redistribution of private property. I have concluded that, given textual evidence, a possible redistributive rule should be interpreted in terms of (RU.I.) for Hume. It should be enforced and carried out in society in accordance with a logic of maximization of public utility. And, since the concept of public utility is not necessarily built by Hume as a function of the private utilities of each individual, it follows that (C.I.) is ruled out as a suitable description of Hume's account of redistribution.
CHAPTER 5

The Rule of the Obligation of Promises

5.1 Promises and the Natural Law Theory

In the third chapter, I have attempted to provide a careful and detailed analysis of the central concept of Hume’s theory of justice, the convention of justice. For Hume, individuals agree to establish a set of rules of justice by means of a set of successive social conventions or, in other words, a convention of justice: i) the rule of private property rights, ii) the rule of transference of private property by consent, and iii) the rule of the obligation of promises. This set of rules is taken by Hume to be of the utmost importance, to the point that he would refer to it as the “three fundamental laws of nature”. This description is meant to emphasize that this set of rules defines a practice that is common and universal among human beings, as it plays a pivotal role in solving basic problems and satisfying basic necessities of all. In his exposition, Hume begins by focusing on the rule of stabilization of possessions and private property rights, which are the first introduced by the convention of justice. In the fourth chapter, I have examined Hume’s theory of property, which a central part of Hume’s theory of justice. Different kinds of rules that further specify the rule of stabilization of possessions and private property rights are addressed, among which, the rule of transference of private property by consent. Now, it is important to conclude the study of Hume’s
convention of justice by discussing the third rule of justice, the rule of the obligation of promises. Thus, I intend to address Hume’s theory of promises, which is the other central part of Hume’s theory of justice. In this discussion, I will also be concerned with engaging in the interpretative debate in order to assess whether Hume’s theory of promises could be interpreted in light of the three main lines of interpretation suggested so far.

Hume’s theory of promises is very important in the history of moral philosophy. Hume is regarded as the philosopher who first developed the conventionalist theory of promises. This theory states that the promises are established by means of a social convention. In other words, it claims that the practice of promising is defined by a set of social rules. And, the only reason why promises are invented and regarded as binding is that the practice of promising yields socially advantageous outcomes to the participants of this social convention. Promises are a kind of device designed to ensure mutual trust among individuals. In this way, the practice of promising tends to foster a cooperative social structure of interaction. Therefore, according to this theory, promises should be considered as social inventions designed to promote socially advantageous outcomes to those who engage in the practice of promising. In contrast to this theory, promises have been traditionally conceived according to the natural law theory, in particular, in Hume’s time. This theory states that promises are binding because they are prescribed as a moral duty to individuals by God’s divine command. The idea is that God created humans as rational beings and, as part of their rational nature, is the ability to recognize and act according to moral rules. And the moral duty to keep promises is one of these moral rules. Consequently, it is precisely against this kind of theory that Hume argues in the course of his exposition. Strictly speaking, Hume’s account of promises could be divided into two parts: a
negative and a positive one. The negative one concerns Hume’s critique of the account of the obligation to keep promises according to natural law theory, whereas the positive one concerns Hume’s account of the obligation to keep promises based on social conventions. Therefore, it is useful to start this analysis by briefly introducing the account of promises according to natural law theorists of the 17th century, such as, for instance, Hugo Grotius, Samuel Pufendorf, and John Locke, in order to understand the historical context in which Hume found himself.

Natural law theory states that promises are binding because individuals, insofar as they are rational, are capable of recognizing as a moral obligation the duty to keep them, for it is commanded by God as a moral rule. This rational recognition stems from human nature itself, as designed by God. This kind of metaphysics of human nature implies that individuals are, by their very nature, endowed with certain essential features. These essential features include, for instance, the property of being rational and the property of having a moral personality, which makes each human being, in principle, capable of rationally recognizing certain moral duties. Thus, it can be inferred that it is a natural law, derived from this very conception of human nature, that individuals ought to keep promises as a moral duty. Notice that, according to this kind of theory, promises have a very particular normative force. Promises bind individuals based on their rational nature. Ultimately, individuals are required to keep promises because of the way they are. As a result, promises bind individuals regardless of any preexisting social institution. In the Second Treatise of Government (1690), Locke characterizes the moral duty to keep promises in this way, as the following passage suggests:
The promises and bargains for truck, &c. between the two men in the desert island, mentioned by Garcilasso de la Vega, in his history of Peru; or between a Swiss and an Indian, in the woods of America; are binding to them, though they are perfectly in a state of nature, in reference to one another: for truth and keeping faith belongs to men as men, and not as members of society. (2003: 106).

Similarly, Pufendorf also argues for the natural law theories, as it is stated in a passage of *De iure naturae et gentium* (1672):

> Now if any pacts are entered into among men, the sociable nature of man requires that they be religiously observed. For if it required less than this the greatest part of the utility arising for humankind from the reciprocal exchange of duties would be lost. [...] Also, unless it were necessary to keep promises, one could by no means count firmly on other men’s assistance. And indeed, the most just causes of quarrels and of war appropriately spring from trust betrayed. [...] Therefore, it is a most sacred precept of the natural law, and one that regulates the grace, the measure, and the order of human life in general, that “Everyone should keep his given word, or fulfill his promises and pacts.” (1994: 166-167)

Notice that Pufendorf points out that “promises and pacts” are intended to promote socially advantageous outcomes, in that they allow “reciprocal exchanges of duties” by ensuring mutual trust among individuals. Promises and pacts ensure that each individual can “firmly” count on other “men’s assistance”. However, Pufendorf also stresses that the duty to keep promises is “a most sacred precept” of “the natural law”. The duty to “religiously observe” promises and pacts is based on human nature or, as Pufendorf puts it, on the “sociable nature of man”.

Thus, for both Locke and Pufendorf, the moral obligation of promises is based on certain essential features of human nature itself. Consequently, the duty to keep promises, which is prescribed by the natural law, is supposed to bind individuals regardless of social conventions or participation in society. Therefore, it is appropriate to conclude that, for natural law theorists, promises have a particular normative force based on a conception of human nature.
According to this kind of theory, the duty to keep promises is prescribed by God as a natural law and is required as a moral obligation on grounds of the very conception of human nature held by philosophers of this tradition. Now, with regard to the source of promises and, by implication, to the nature of this kind of moral obligation, natural law theorists hold that promises are constituted by an act of will of those who commit themselves to perform it. The idea is that promises, insofar as they restrict the freedom of individuals and compel them to perform some future action that they would otherwise not perform, must require their consent to be considered as a moral obligation. This idea is illustrated by Pufendorf in the following passage of De iure naturae et gentium (1672):

> Since promises and pacts as a rule restrict our freedom and impose on us a burden of necessarily doing something we were formerly free to do or not do, there is no more appropriate reason that someone cannot rightly complain about having to carry such a burden than the fact that he himself has previously consented to it and willingly taken upon himself what he was formerly able to refuse. (1994: 170-171).

The same point is made by Hugo Grotius in De jure belli ac pacis (1625):

> A third Degree is, when to this Determination we add a sufficient Declaration of our Will to confer on another a real Right of demanding the Performance of our Promise. And this is a compleat Promise, as having the same Effect as the Alienation of a Man’s Property. For it is either an Introduction to the Alienating of a Thing, or the Alienation of some Part of our Liberty. To the former belong our Promises to give, to the latter our Promises to do something. And of this the Scriptures give us a notable Proof, where they tell us, that GOD himself, who cannot be obliged by any Law imposed by another, would act contrary to his own Nature, not to perform what he promised [...] whence it is plain, that to perform Promises is a Duty arising from the Nature of immutable Justice, which as it is in GOD, so it is in some Measure common to all such as have the Use of Reason. (2005: 704-705)

Natural law theorists have placed great importance on the will of the promising parties as a suitable source of promises. Unsurprisingly, both Pufendorf and Grotius have devoted a lot of effort not only to accurately...
describe the kind of “consent” or “declaration of will” that properly constitutes a promise, but also the kind of external behavior most suitable to express this consent or declaration of will. For if the will is in any way impaired, it makes promises void or invalid, as well as their morally obligatory character. So, given that promises require an act of will as a constitutive source for them to be valid at all, it follows that obligations related to the duty to keep promises also depend on this expression of consent or will.

The “will theory” became predominant in the discussion of promises in the 17th century. It is only in the 20th century that this kind of theory was strongly challenged due to the emphasis on subjective or psychological elements as the constitutive source of promises. Instead, contemporary legal theories tend to focus more on objective criteria of liability and non-liability than on any kind of evidence that consent or will has, in fact, been successfully expressed. It is in the context of the critique of the “will theory” that Hume points out the difficulties of the account of promises and their morally obligatory character according to the natural law theories.

5.2 Hume’s Critique of the Traditional Theories

5.2.1 The Intelligibility and the Moral Obligation of Promises

In the Treatise Book 3, Part 2, Section 5, Hume starts the negative part of his account of promises by arguing against the central thesis of the natural law theory. Hume claims that the rule of the obligation of promises is not a natural rule. Thus, individuals are not, by nature, morally obliged to keep promises.

---

112 According to Atiyah, this view is supported by H. L. A. Hart (1981: 19).
And, this claim is supported by two theses that Hume dedicates himself to prove. Hume presents the two theses in this way in the following passage:

That the rule of morality, which enjoins the performance of promises, is not natural, will sufficiently appear from these two propositions, which I proceed to prove, viz. that a promise would not be intelligible, before human conventions had established it; and that even if it were intelligible, it would not be attended with any moral obligation. (T 3.2.5.1/SBN 516)

This way of framing the discussion is very useful, as it shows that there are two different problems at stake here. One the one hand, the first thesis addresses the intelligibility of promises. It is supposed to answer which conditions should be satisfied so that a specific kind of behavior can be comprehended as a promise. On the other hand, the second thesis addresses the moral obligation of the practice to keep promises. It questions whether this specific kind of behavior, even if it could be, in principle, comprehended as a promise, could have any moral significance for us, that is, whether it could generate a moral obligation or not.

The first thesis states that promises are unintelligible in the absence of preexisting social conventions. It amounts to the question as to which conditions must be satisfied for a specific behavior or a linguistic expression to be comprehended at all. So, what Hume is arguing here is that signs or verbal expressions, such as “I promise to x”, “I will do x”, or “I hereby place myself under an obligation to do x in the future”, made or uttered by any particular individual, cannot be understood before the establishment of a social convention that prescribes a specific behavior specifically related to the performance or utterance of such linguistic expressions. For Hume, determining the intelligibility of a specific behavior or linguistic expression is a matter of identifying the impression which it can be traced back to. This is the method employed by Hume for determining the intelligibility of
concepts in general. Basic metaphysical concepts such as substance, mind, and causality are analyzed in this way in the *Treatise* and in the *Enquiry* (T 1.1.1.11/SBN 6-7; *EHU* 2.9/SBN 22). Similarly, the analysis of the intelligibility of promises requires the clarification of linguistic expressions, such as signs or verbal expressions, as, for instance, “I promise to x”, “I will do x”, or “I hereby place myself under an obligation to do x in the future”, which, in most of the times, refer to a promise. Therefore, it is necessary to survey the human nature to single out the impression, which is, in this case, an act of will, corresponding to this kind of linguistic expressions, in order for promises to be intelligible at all.\(^{113}\) This is the point made by Hume in the following passage of the *Treatise* Book 3, Part 2, Section 5:

> If promises be natural and intelligible, there must be some act of mind attending these words, *I promise*; and on this act of the mind must the obligation depend. Let us, therefore, run over all the faculties of the soul, and see which of them is exerted in our promises. (*T* 3.2.5.2/SBN 516)\(^{114}\)

In the *Treatise*, Hume considers three possible “acts of mind” that, when referred to by such linguistic expressions, could make a promise intelligible: i) the resolution to do something, ii) the desire to do something, and iii) willing to do something. The resolve to do something is not to be confused with a promise to do something. The mental state characterized by one’s decision or resolution to do something does not necessarily bind him or her to do it in the future. It may be possible for one to decide or resolve to do

---

\(^{113}\) It is important to notice that, for Hume, the will is a kind of impression as well. It is basically an internal impression of which individuals are aware when they act or think (*T* 2.3.1.2/SBN 399).

\(^{114}\) Strictly speaking, Hume does not mention non-verbal signs or gestures in his analysis. Under the right circumstances, non-verbal signs or gestures can amount to the making of a promise. Presumably, Hume is restricting his analysis to what he considers the basic form of a promise, which is basically referred to by a verbal expression of the form “I promise”.
something but change his or her mind later. However, a promise requires one’s intention to lay oneself under an obligation to do something in the future. Therefore, there is no conceptual entailment between a decision or resolution and a promise, for a decision or resolution does not involve an obligation of any sort.

Neither is the desire to do something the act of mind of a promise. Clearly, one can bind oneself without really desiring to perform one’s part in a promise. It is common for one to promise something and fulfill it even if one desires to break the promise. The act of mind expressed by a promise cannot be the will to do the promised action either.\textsuperscript{115} To support this claim, Hume presents an argument similar to the one made by Hobbes in the \textit{Leviathan}. For Hobbes, linguistic expressions characterizing promises involve words referring to a present or a future time. On the one hand, in the former case, when a promise is formulated by words in the present tense, it refers to an act of the will in the present to transfer a right in the future. On the other hand, in the latter case, when a promise is formulated by words in the future tense, it refers to an act of the will in the future. However, when additional signs upon which one can safely rely are absent, promises formulated by words in the future tense can barely be said to transfer any right.

\textsuperscript{115} For Hume, the mental states of willing to do something and desiring to do something are not the same thing. Both are kinds of impressions, but Hume distinguishes the mental state of desiring from the mental state of willing in terms of the strength and intensity of the impressions that properly define each one of them. On the one hand, desire is a kind of “direct passion”. Such passions are a stronger kind of reflective impression that arises directly from sensory impressions, such as pain or pleasure. The passion of desire is to be contrasted with the passion of aversion. On the other hand, will is a weaker kind of reflective impression associated with conscious intention or volition in deliberate thoughts or behavior. The impression of will is to be contrasted with unintentional or purposeless thoughts or behavior (\textit{T} 2.1.2.4/SBN 276-277; \textit{T} 2.3.1.1-2/SBN 339).
Hume agrees with this point. Promises, in general, are supposed to orient future performances, but an act of will can only influence present actions. And, it is not certain whether an act of the will in the present will be qualitatively identical to the one in the future. Therefore, the act of mind of willing to do something is not conceptually identical to a promise.

After ruling out the three possible acts of mind, Hume comes to the conclusion that promises are defined by the act of mind of willing the obligation to do something. For Hume, the only act of mind by which promises could, in principle, be rendered intelligible is the act of willing the obligation to do something. In fact, this conclusion is in line with the most common intuition and sound philosophical thinking. Hume’s definition of promise is not different from that of any other philosopher. The definition of promise is the voluntary undertaking of an obligation to do something for someone else in the future. Hume is aware of this. However, the mere identification of the act of mind that defines promises still does not answer how they can be intelligible. What Hume needs to answer is how the notion of obligation can arise in the first place. Ultimately, this is what makes promises intelligible. Hume points out that this notion of obligation must not be merely assumed or taken for granted as natural of individuals. Rather, it begs for a plausible explanation, for it is precisely on this idea that the whole practice of promising is grounded. Hume presents his conclusion as follows:

It follows, therefore, that since the act of mind, which enters into a promise, and produces its obligation, is neither the resolving, desiring, nor willing any particular performance, it must necessarily be the willing of that obligation, which arises from the promise. Nor is this only a conclusion of philosophy; but is is entirely conformable to our common ways of thinking and of expressing ourselves, when we say that we are bound by our own consent, and that the obligation arises from a mere will and pleasure. The only question, then, is, whether there be not a manifest

---

Hume claims that willing the obligation to do something is the act of mind that properly defines a promise. Now, it is necessary to understand why, for Hume, it would be a “manifest absurdity” in “supposing this act of the mind” as natural of individuals. Ultimately, this point boils down to explaining how the notion of obligation, which is implied in the act of mind that characterizes a promise, can emerge. Thus, to support this claim, Hume presents an argument that relies on some of the assumptions made in the context of his moral theory, in particular, on the naturalistic and sentimentalist character of his moral theory. This argument is also related to the second thesis that states that promises, even if they were intelligible, could not generate a moral obligation.

Hume’s argument relies on two theses of his moral theory: i) morality, on which the practice of promising is primarily based on, consists in entertaining pleasant and unpleasant sentiments of moral approbation and disapprobation induced by character traits or actions from the impartial stance of the general point of view, and ii) one can only be said to be morally obliged to behave or act in some way if one fails to display a character trait or perform an action caused by a motive that induces a pleasant sentiment of approbation from the impartial stance of the general point of view. Since morality is based on sentiments, character traits or actions are said to be morally good when they are caused by a motive that induces a pleasant feeling of moral approbation. The failure in displaying these character traits or performing these actions conveys an unpleasant sentiment of moral disapprobation, as it is supposed to be a sign of the absence of a suitable motive to cause the moral
agent to behave morally. The possibility of this practical failure places the moral agent under a moral obligation to display the character traits or perform the actions required. The conclusion is that, for Hume, a moral obligation entails the requirement of a suitable motive that induces a particular moral behavior. Consequently, a moral obligation entails the arousal of a suitable moral sentiment.

However, the fundamental problem is that it is not up to us to change our sentiments in the same way that we want to voluntary undertake any obligations. In general, sentiments are not in our control. We are not free to manipulate them as we wish. Rather, we are born with them and they are only triggered by appropriate external stimuli, regardless of how we might feel about it. Efforts to change them at will would be either fruitless or have marginal effects at best. In contrast, taking on an obligation, in the case where there are no hindrances, is an act of will par excellence. In general, we are free to will whichever obligation we want, whenever we want. Notice that Hume is trying to show an incongruity between sentiments and moral obligations entailed by promises with this argument. The fact that moral obligations are anchored in sentiments poses a challenge to the commonsense idea that individuals could take on any obligations at will and even recognize corresponding utterances of this kind of performance as binding, because the arousal of sentiments, on which moral obligations related to the performance of promises depend, is not itself a voluntary act. This is why individuals are not naturally endowed with an act of mind of willing to undertake any obligation to do something, as it is required as a condition for the successful performance of a promise. Recall once again what natural law theorists argue. They claim that individuals, by a simple act of will and regardless of preexisting social institutions, would be able to take on as many
obligations as they wanted through promises. And, in principle, they would be able to recognize, in every single instance, that they are morally obliged to keep them by nature whenever the relevant linguistic expressions are uttered. This is precisely what Hume denies by this argument. It makes its case by distinguishing two particular mental states of human psychology, sentiments and will. Hume intends to contrast them by emphasizing the property of sentiments to be involuntary and the property of the will to be voluntary. For Hume, this is why individuals would not, by nature, be able to understand the act of willing an obligation to do something for someone as implying a moral obligation, even if this is exactly what promises are supposed to imply. So, the point is that, after a more careful survey of the human nature, Hume argues that the act of mind of willing an obligation, to which the linguistic expressions of a promise should refer, is not naturally implanted in our nature in the way that sentiments, for instance, are. The conclusion of this argument is that promises are not naturally intelligible and, even if they were, they could not generate a moral obligation, as the following passage shows:

All morality depends upon our sentiments; and when any action, or quality of mind, pleases us after a certain manner, we say it is virtuous; and when the neglect, or non-performance of it, displease us after a like manner, we say that we lie under an obligation to perform it. A change of the obligation supposes a change of the sentiment; and a creation of a new obligation supposes a new sentiment to arise. But ’tis certain we can naturally no more change our own sentiments, than the motions of the heavens; nor by a single act of our will, that is, by a promise, render any action agreeable or disagreeable, moral or immoral; which without that act, wou’d have produc’d contrary impressions, or have been endow’d with different qualities. It wou’d be absurd, therefore, to will any new obligation, that is, any new sentiment of pain or pleasure; nor it possible, that men cou’d naturally fall into so gross an absurdity. A promise, therefore, is naturally altogether unintelligible, nor is there any act of the mind belonging to it. (T 3.2.5.2/SBN 517)

If willing an obligation to do something, which is the suitable act of mind that characterizes a promise, is neither naturally intelligible nor can generate a moral obligation for individuals, then Hume has to consider an alternative
hypothesis to explain how this particular act of mind can emerge. The hypothesis is precisely to advance a conventionalist theory of promises. This theory regards promises as a kind of human invention. Promises are established by a social convention in light of the necessities and interests of individuals in society. They are created with the sole purpose of yielding socially advantageous outcomes for all. However, before addressing the details of Hume’s theory of promises, it is worth considering two objections against Hume’s argument about the intelligibility of promises.

### 5.2.2 Two Objections

Two major objections could be raised against Hume’s argument about the intelligibility of promises. First, it could be argued that the assumptions of Hume’s moral theory are not convincing at all. One simply does not need to accept that promises are anchored in sentiments, let alone his very notion of moral obligation. The thesis that moral obligations are grounded on sentiments of approbation or disapprobation might only be acceptable from the perspective of a sentimentalist ethical theory. Other equally plausible and consistent theories of moral obligation have been provided by philosophers throughout history. So, why should one opt for Hume’s theory of moral obligation to the detriment of all others? Clearly, Hume’s thesis that moral obligations are grounded on sentiments of approbation or disapprobation favors the conclusion that promises are neither naturally intelligible nor can generate a moral obligation. Therefore, if one rejects Hume’s sentimentalist premises in the argument, the conclusion might well be the opposite.
One traditional way of justifying the morally binding character of
promises is through the notion of expectations. And, this claim can be made
by resorting or not to preexisting social institutions. It could be argued that
when certain linguistic expressions such as “I promise to x”, “I will do x”, or
“I hereby place myself under an obligation to do x in the future” are uttered
by any particular individual, an expectation is thereby created. This
expectation, in turn, deserves protection, for one is harmed if those
expectations are somehow disappointed. However, the idea of protecting
expectations can be very obscure. Sometimes expectations can be highly
subjective, as the very notion of expectation entails a subjective point of view.
The utterance of the same linguistic expression may cause expectations in
some individuals but not in others. That is why expectations are sometimes
considered as a form of psychological reliance. The same utterance may
assure some individuals, but not others. There is simply no objective criteria
by which individuals can be said to be assured based only on the notion of
expectation. By the same token, even the notion of harm inflicted by the non-
performance of what one said that would perform is highly subjective if it is
understood under this account. These doubts make the whole account very
blurry and confusing.

In addition, it can be asked why the utterance of some linguistic
expressions should create expectations in the first place. Is it even possible
that a set of linguistic expressions, solely by its mere utterance and regardless
of preexisting social institutions, can cause expectations to such an extent that
it morally binds an individual to perform a given action? Does the meaning
of some linguistic expressions semantically imply the generation of

118 See Atiyah (1918: 42-44).
expectations? The utterance of the linguistic expressions aforementioned, without further qualifications, can easily be confounded with a threat, a prediction, a warning, or a declaration of intention. Some may generate expectations typical of promises, others may generate expectations quite different from those expected from promises, and others may generate no expectations at all. Thus, no linguistic expression, by its mere utterance, entail the generation of expectations regardless of the background social institutions. Instead, it should be concluded that if the utterance of a linguistic expression in the appropriate circumstances creates a promissory obligation, then this is more evidence of a previously established social rule than of alleged expectations.¹²⁰ So, Hume’s theses do not need to be supported only by endorsing his moral theory. In fact, they can be supported by multiple theoretical perspectives.

Second, it could be argued that Hume’s framing of the problem is fundamentally wrong on semantic grounds. Like most of the philosophers of his time, Hume is a proponent of the “will theory”. He makes some critical remarks about the theory, such as developed by the tradition of the natural law theory, but he never really abandons it. In fact, in the context of the discussion about the natural intelligibility of promises, Hume makes the claim that a promise is defined by an act of mind, that is, that of willing an obligation to do something. This conclusion is clearly based on the endorsement of a specific semantic theory, the ideational or mentalist theory of meaning. This theory is a class of a more comprehensive semantic theory, the denotative theory of meaning. According to this theory, the meaning of an expression is its referent, which is the object that the expression denotes. In Hume’s case,

¹²⁰ See Anscombe (1978: 318).
the object that the linguistic expression “I promise” denotes is an act of mind. Thus, for Hume, the meaning of a promise is the idea it stands for. This is why it can be said that Hume endorses the ideational or mentalist theory of meaning.

The problem with this account is that the denotative theory of meaning assumes that the sole role of language is either to denote objects or to describe states of affairs. However, not every expression is meant to denote or describe things. Promises, for instance, when they are successfully performed, are not supposed to describe anything in the world. They are a kind of performative expression. Their meaning is given by the fact that, through its utterance, something is performed and not described. Therefore, in a sense, part of Hume’s analysis is fundamentally flawed. It is based on a misunderstanding about the exact nature of the meaning of a promise. Perhaps an even deeper misunderstanding about the exact nature of language in general. However, in another sense, Hume’s general thesis remains untouched. Recall that the Hume’s purpose with this argument is to criticize the hypothesis that claims that promises could be naturally intelligible. And, even the philosophers who endorse different semantic theories do not question the conventional character of the promises. This means that they agree with Hume’s thesis, albeit for different reasons. In this case, it should be concluded that the negative part of Hume’s account of promises is not entirely incorrect. In fact, some assumptions, especially the semantic ones, are wrong. However, this only indicates that the argument made by Hume that is supposed to support the thesis that promises are not naturally intelligible is not very convincing. It does not follow that the whole account is false. In fact, the positive part of

---

121 See Austin (1962: 11), Searle (1964: 51).
Hume’s account of promises, the one in which he lays down the foundations of his conventionalist theory of promises, is far more promising. In the following, the basis of justification of Hume’s theory of promises will be presented through an analysis inspired by game theory.

5.3 The Logic of Promises: A Game-Theoretic Analysis
5.3.1 Coordination Promises and Exchange Promises
The positive part of Hume’s account of promises is meant to show that they are genuine human inventions established by social conventions. Promises operate as devices that ensure mutual trust among individuals, which is a necessary condition for social cooperation in general. This is the central tenet of the conventionalist theory of promises. However, it is important to notice that even if cooperative solutions are available to individuals, it does not simply follow that they will, in fact, cooperate. For Hume, individuals are moderately selfish. They are not naturally prone to cooperate. Thus, the question to be answered here is, in view of the self-interested nature of individuals, how social cooperation can be possible in the first place. Given these seemingly paradoxical circumstances that characterize the possibility of social cooperation between moderately selfish individuals, one of the strategies to analyze this problem is to consider it by means of an analogy with certain game theory models. Thus, by illustrating strategic situations in which individuals typically find themselves in a social structure of interaction, it is possible to explain what are the conditions under which the adoption of a cooperative strategy is a rational choice for each individual.

But, as a preliminary remark, it is important to draw a distinction between two possible kinds of promises, so that it can be determined which is the appropriate object for which Hume must present a basis of justification.
Promises of the first kind can be characterized as “coordination promises”, whereas promises of the second kind can be characterized as “exchange promises”. Coordination promises, as the name already suggests, are especially relevant for solving coordination problems. A coordination problem is a strategic situation defined by the following conjunction of factors: i) a social structure of interaction in which two or more players are uncertain of each others’ decision; and ii) a social structure of interaction in which there are two or more outcomes that each player equally prefers more than anything else. Each one of these preferable outcomes is defined as coordination equilibrium, which is an outcome of a game determined by a set of strategies such that no player would be better off if any player would to choose alone another strategy and unilaterally deviated from this outcome. Given that a coordination equilibrium is an outcome preferred by each player, as soon as it is achieved, they have no incentives to deviate from their current strategy. In this sense, social arrangements that generate a coordination equilibrium have the property of being *strictly self-enforcing*.

In the second chapter, I have already pointed out that, in pure coordination games, the players’ interests fully coincide. There is really no conflict of interest under such circumstances. The equilibrium point in pure coordination games coincides with the *Pareto optimal* outcome. This means that the outcome that amounts to the solution for pure coordination games is one that cannot be changed without at least one player ending up worse off than before. So, what is really the problem with this kind of game? Why would a promise be required under such circumstances? The problem is that, just as in any strategic situation in a social structure of interaction, players

---

must choose strategies given the choice of strategies of all others. Even if there may be a complete coincidence of interests and no communication obstruction, a device may still be needed to assist players in coordinating actions. This is why coordination promises exist. Players, by saying that they oblige themselves to perform actions that lead to a particular outcome in coordination equilibrium, succeed in solving coordination problems by means of a promise. A coordination promise is a device meant to assists individuals to achieve general coordination of actions.

However, it is not always the case that the players’ interests coincide in strategic situations in social structures of interaction. In fact, it is common that there is a partial conflict of interest in situations where a promise is required. Typical cases usually involve non-simultaneous mutual exchanges of goods and services or mutual exchanges in which the expected quality of the good or in which the extension of the service is not easily discernible. In such cases, if it is assumed that players are mainly self-interested and prone to a utility-maximizing behavior, then the player in favor of whom goods or services were given or performed in advance has incentives to be disloyal and not cooperate in return. In turn, the player who gave goods or performed services in advance is at risk of ending up worse off in such cases. Consequently, given the lack of trust, the latter player would not be willing to give the former player the benefit of the doubt in this non-simultaneous mutual exchange. Under such circumstances, the result is that no mutual exchange could take place whatsoever, which would preclude a specific kind of commerce and trade of goods and services in society to the harm of the well-being of all. In the Treatise Book 3, Part 2, Section 5, more specifically,

at the beginning of the positive part of Hume’s theory of promises, the concern is to explain how the lack of trust among individuals could be bridged given the moderately selfish nature of individuals, as the following passage shows:

Men being naturally selfish, or endow’d only with a confin’d generosity, they are not easily induc’d to perform any action for the interest of strangers, except with a view to some reciprocal advantages, which they had no hope of obtaining but by such a performance. Now as it frequently happens, that these mutual performances cannot be finish’d at the same instant, ‘tis necessary, that one party be contented to remain in uncertainty, and depend upon the gratitude of the other for a return of kindness. But so much corruption is there among men, that, generally speaking, this becomes but a slender security; and as the benefactor is here suppos’d to bestow his favors with a view to self-interest, this both takes off from the obligation, and sets an example of selfishness, which is the true mother of ingratitude. Were we, therefore, to follow the natural course of our passions and inclinations, we shou’d perform but few actions or the advantage of others, from disinterested views; because we are naturally very limited in our kindness and affection: And we shou’d perform as few of that kind, out of a regard to interest; because we cannot depend upon their gratitude. Here then is the mutual commerce of good offices in a manner lost among mankind, and every one reduc’d to his own skill and industry for his well-being and subsistence. (T 3.2.5.8/SBN 519-520)

Notice that, unlike in pure coordination games, in which the player’s interests fully coincide, the description of this social structure of interaction resembles that of the Prisoner’s Dilemma. This game is a kind of non-zero-sum game characterized by a partial conflict of interest between players. At first, it reinforces distrust between the players, as Hume points out in the passage. The equilibrium point of this kind of game is the outcome defined by a set of competitive strategies of the players. In the Prisoner’s Dilemma, the players are encouraged to choose competitive strategies, for it ensures that none of them ends up worse off in a social structure of interaction in which players have strong incentives not to cooperate. Each player knows that, by choosing unilaterally to deviate from a cooperative strategy and switch for a cooperative strategy while the other’s strategies remain fixed, then they may
end up bearing the highest possible loss. Thus, a competitive strategy maximizes each player’s payoff, given the maximizing strategy of the other. Nonetheless, even if the outcome in equilibrium is the one defined by a set of competitive strategies, this outcome is still worse than the one in which the players cooperate. In the Prisoner’s Dilemma, the Pareto optimal outcome, the one that cannot be changed unless the situation of at least one of the players is worsened, is defined by a set of cooperative strategies by both players. Thus, the strategic problem in this game is that the outcome in equilibrium is not the outcome that is Pareto optimal, even if the possibility of achieving this outcome is still open to the players. And, this is precisely the situation in which players involved in non-simultaneous mutual exchanges find themselves.

It is under such circumstances that exchange promises are relevant. They are categorically different from coordination promises in that they effectively contribute to ensuring mutual trust among individuals. This is not something that coordination promises are effectively designed to do. In coordination problems, the lack of trust is not really the main issue. Instead, communication and organization problems in large-scale arrangements that are most potentially harmful. Exchange promises are designed to create a mechanism of trust among players in situations of mutual exchange. They allow a social structure of interaction based on mutual exchanges of goods and services to emerge. And, this is why it is with exchange promises that Hume is chiefly concerned. And, Hume introduces the device of the exchange promises as a solution for the cases of a non-simultaneous mutual exchange, which are situations that typically induce mistrust among individuals, just like that described by the Prisoner’s Dilemma game model.
Now, given the distinction between coordination promises and exchange promises, it is important to delimit the subject to the exchange promises for this discussion. Thus, it is necessary to study the logic of the exchange promises in order to understand the basis of justification of this practice as a social invention resulting from the necessities and interests of individuals in a society.

5.3.2 The Farmer’s Dilemma
It could be noticed that Hume is addressing the whole positive part of his theory of promises from the perspective of the individual choice. The decision to perform or not the part in exchange promises is framed as a kind of strategic problem in which individuals have to make a rational choice, given the choice of everyone else. And this invites a game-theoretic account of exchange promises. The logic of the exchange promises could be illustrated in terms of the game theory model of the Farmer’s Dilemma. The social structure of interaction that presents the kind of strategic problem that exchange promises are designed to solve is described by Hume as such in the following passage of the Treatise:

Your corn is ripe today; mine will be so to-morrow. ’Tis profitable for us both, that I shou’d labour with you to-day, and that you shou’d aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains upon your account; and should I labour with you upon my own account, in expectation of a return, I know I shou’d be disappointed, and that I shou’d in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security. (T 3.2.5.8/SBN 520-521)

The game theory models of the Farmer’s Dilemma and the Prisoner’s Dilemma share important similarities. For instance, in both game theory models, the players have the same possible strategies and the same possible
payoffs from the same possible outcomes. And, common knowledge of the players about the possible strategies, payoffs, and outcomes suggests that both are complete information games, as Hume also seems to assume in the description. However, the former model captures Hume’s idea of a non-simultaneous mutual exchange better than the latter model. This is because the Farmer’s Dilemma introduces a relevant property. The Prisoner’s Dilemma is a simultaneous game, that is, a game in which each player chooses a strategy independently and simultaneously, without knowing beforehand the choices of the others. This uncertainty regarding the other players’ previous choices characterizes a game of imperfect information. In such games, the players choose a strategy without information regarding how others have previously chosen to behave. The Farmer’s Dilemma is a sequential game, that is, a game in which, after an initial chance move, each player chooses a strategy after the others’ moves. This sequential structure of the players’ strategy choices suggests a game of perfect information. In such games, the players are informed about all the previous choices in the game.\textsuperscript{125}

The sequential structure that characterizes this strategic problem is understood by Hume, as it is represented in the description as a problem for the players as to how to behave in the present in face of future social interactions.

However, as Hume already anticipates in the passage, this distinction in a particular property that defines each game theory model does not change the fact that, both in the Farmer’s Dilemma and in the Prisoner’s Dilemma, the players face a similar problem. Even if they are able to infer that it is a rational choice to agree to establish the rule of the obligation to perform the

\textsuperscript{125} See Vanderschraaf (1998: 219, 224).
respective parts of mutual exchanges, they may still question whether it is rational or not to behave according to the pattern of behavior defined by this rule. This problem is equivalent to the one posed by Hobbes’ Foole. For Hobbes, the Foole agrees that it is rational to enter into an agreement to perform covenants, but questions the rationale of complying with the pattern of behavior defined by this agreement if there is a more advantageous strategy available to him or her. And, indeed, there is a more advantageous strategy available for the Foole. Given the sequential structure of mutual exchanges in the scenario described by the Farmer’s Dilemma, there is a more advantageous strategy than the mere performance of one’s part in the mutual exchange given the previous performance of others’ part. This strategy is not to perform one’s own part of the mutual exchange, as previously agreed, given the previous performance of the other party. The challenge is to convince the Foole that, given the previous and secure performance of a part of the mutual exchange by the other party, it is still a rational choice to perform the own part of the mutual exchange. For if the players expect that the Foole is not disposed to perform the own part, but is only willing to seize the opportunity to exploit others, the players would have no incentives to perform their parts either. Consequently, it could be concluded that it is rational for individuals to establish the rule of the obligation to perform mutual exchanges, but it is not rational for individuals to act according to the specific pattern of behavior defined by this rule. As Hume correctly assumes, the nature of the strategic problem that players face in the Farmer’s Dilemma game is that, even if the possibility of successful mutual exchanges remains open for them, they still have stronger incentives to choose a Pareto

Therefore, a solid basis of justification for the existence of exchange promises is not possible based only on the structural nature of the Farmer’s Dilemma. However, it is still a useful game theory model upon which one can further analyze.

5.3.3 The Infinitely Repeated Farmer’s Dilemma
The solution to the problem of achieving a Pareto optimal outcome from a Pareto suboptimal outcome involves refining the representation of the Farmer’s Dilemma by introducing further properties. Here, the strategy is similar to that developed for the Stag Hunt game in the second chapter. Recall that the purpose of examining the Stag Hunt game was to explore the structural similarities between this game theory model and Hume’s pre-conventional structure of interaction. The analysis was intended to provide a solution for the Stag Hunt game, which, due to its structural similarities, could be imagined as analogous to Hume’s convention of justice as a solution for the Humean state of nature. Now, the purpose is to apply the same method to understand the basis of justification of the exchange promises in Hume’s theory of promises, as it is based on structural similarities of this social structure of interaction in which individuals find themselves with the Farmer’s Dilemma game. As in the Stag Hunt game, the Farmer’s Dilemma entails a Pareto suboptimal outcome only if it is conceived in a static model, that is, a one-shot game. If the Farmer Dilemma is conceived in a dynamic model, in particular, as an infinitely repeated game, it is expected that the players will choose different strategies and different outcomes are likely to emerge. This is because players will have stronger incentives to consider the

---

strategy that maximizes their payoffs in the long and not in the short run. Consequently, the outcome in equilibrium in an infinitely repeated Farmer’s Dilemma will be defined by a set of strategies that differs from the one that defines an outcome in equilibrium in the Farmer’s Dilemma one-shot game. In other words, to the extent that players are able to understand that they can maximize their payoffs in the long run in such repetitive social interactions by choosing the strategy to perform their parts in mutual exchanges, provided that the others do the same, successful mutual exchanges become a possible outcome for the players in this game. And, Hume seems to agree with this picture, as it is assumed in his description that, in such mutual exchanges, the players are able to “learn” how to behave based on “expectations” that similar exchanges will continue to take place in the future. This allows that the players have the “foresight” of future “consequences” in case of defection, as Hume points out in the following passage:

Hence, I learn to do a service to another, without bearing him any real kindness; because I foresee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me or with others. And accordingly, after I have serv’d him, and he is in possession of the advantage arising from my action, he is induc’d to perform his part, as foreseeing the consequences of his refusal. (T 3.2.5.9/SBN 521)

Notice that, as in the Stag Hunt game, such a description entails two features. First, the players should choose a very particular strategy to achieve a Pareto

---

128 As in the Stag Hunt game, I have pointed out that, strictly speaking, either one of two properties could be assumed in order for the same outcomes to be achieved. It must be common knowledge among the players that either: i) the Farmer’s Dilemma game has the property of infinite repetition of the same base game in a more comprehensive and infinitely repeated game, or ii) the Farmer’s Dilemma game has the property of imperfect information in cases of finitely repeated games. In this way, the players either: i) know that there will not be a last round of mutual exchanges with which they should be concerned, or ii) do not know whether the game will come to a last round of mutual exchanges or not. In either case, mutually advantageous exchanges are expected to be a possible outcome for the players in an infinitely repeated Farmer’s Dilemma.
optimal outcome as a solution for this game. Players should adopt a strategy of conditional cooperative behavior as a rational choice for this game. In this case, cooperative behavior, which is supposed to refer to a wide range of behaviors in strategic situations, is equivalent to performing the respective part of mutual exchanges, whereas non-cooperative behavior is equivalent to not performing it. Ultimately, this strategy of conditional cooperative behavior amounts to a tit-for-tat strategy with regard to the performance of one’s part in mutual exchanges. It prescribes that each player should perform the own part of the mutual exchange, provided that the other players also behave in this way. If, however, the other players choose not to perform their parts of the mutual exchange, then each player should not perform their parts either. As a result, cooperators will look for other cooperators with whom they can engage in mutual exchanges, while non-cooperators will be punished with ostracization of future mutual exchanges based on their bad reputation. In this way, non-cooperators will have strong incentives to cooperate in order to enjoy the social advantages afforded by mutual exchanges in general.

Second, as a condition for the players to properly identify and distinguish cooperators from non-cooperators, it is necessary to assume that this social structure of interaction is restricted to a small population of players. If the infinitely repeated Farmer’s Dilemma is imagined as played by a large population of players, some non-cooperators may get away with their frequent non-cooperative behavior, as they may not be identified by their peers as non-cooperators. In other words, their reputation simply cannot be ascertained. The distinction between cooperators and non-cooperators is undermined by a degree of anonymity, which strengthens the incentives for non-cooperative behavior in general. Thus, for the players to be able to adopt a strategy of conditional cooperative behavior as a rational choice in the first
place, it must be assumed that the infinitely repeated Farmer’s Dilemma should be imagined as played by a small population of players. If the infinitely repeated Farmer’s Dilemma is imagined as played by a large population of players, additional features, such as higher-order coordination devices, must be introduced in order for a Pareto optimal outcome to be achieved.

Now, the basis of justification for the practice of promising becomes firmly established when it is comprehended that the strategy to keep promises is a rational choice for all. On the one hand, for cooperators, the outcome defined by a set of cooperative strategies, that is, the choice to perform the parts of the exchange promise, is a mutually advantageous outcome. On the other hand, for non-cooperators, the prospects of being punished for the choice not to perform the parts of the exchange promise reduces the payoff of the competitive strategy substantially, which is likely to cause them to end worse off in this game. In other words, it is the possibility of punishment that strengthens the incentives to adopt a cooperative strategy. Under such circumstances, while cooperators will keep their prospects of maximizing their expected utility from exchange promises, non-cooperators are unlikely to be able to benefit from this state of affairs, because they will be sanctioned for their non-cooperative behavior. Thus, for non-cooperators, the choice of not basing their actions on a strategy of conditionally cooperative behavior may prove to be so risky and costly to the extent that it could be characterized as irrational to choose a different strategy. In this way, cooperative behavior becomes the rational choice for each player in this game. This is how the outcome defined by a set of cooperative strategies becomes the solution for this game.

After the description of the logic of mutual exchanges in terms of a game-theoretic analysis of the social structure of interaction, it is important
to understand how Hume articulates these intuitions with his conventionalist account of promises. In the *Treatise*, Hume introduces exchange promises as defined by a set of conventional linguistic expressions that individuals, by uttering them, indicate that they voluntarily undertake an obligation to perform something in favor of others in the future, as the following passage shows:

But tho’ this self-interested commerce of men begins to take place, and to predominate in society, it does not entirely abolish the more generous and noble intercourse of friendship and good offices. I may still do services to such persons as I love, and am more particularly acquainted with, without any prospect of advantage; and they may make me a return in the same manner, without any view but that of recompensing my past services. In order, therefore, to distinguish those two different sorts of commerce, the interested and the disinterested, there is a certain form of words invented for the former, by which we bind ourselves to the performance of any action. This form of word constitutes what we call a promise, which is the sanction of the interested commerce of mankind. When a man says he promises anything, he in effect expresses a resolution of performing it; and along with that, by making use of this form of words, subjects himself to the penalty of never being trusted again in case of failure. A resolution is the natural act of the mind, which promises express: But were there no more than a resolution in the case, promises wou’d not create any new motive or obligation. They are the conventions of men, which create a new motive, when experience has taught us, that human affairs wou’d be conducted much more for mutual advantage, were there certain symbols or signs instituted, by which we might give each other security of our conduct in any particular incident. After these signs are instituted, whoever uses them is immediately bound by his interest to execute his engagements, and must never expect to be trusted any more, if he refuse to perform what he promis’d.  

For Hume, in a social structure of interaction modeled in terms of the structural properties introduced to the basic form of the Farmer’s Dilemma game, mutual exchanges could emerge based only on the structural nature of the game. Notice, however, that there are at least two different kinds of mutual exchanges or, as Hume puts it, “two different sorts of commerce, the interested and the disinterested”. As the Farmer’s Dilemma is supposed to represent a situation characterized by a partial conflict of interests involving moderately selfish players, it follows that mutual exchanges should be
performed as exchange promises. And, exchange promises are specified and distinguished from other kinds of mutual exchanges by certain “symbols”, “signs”, and “forms of words”, which are “invented” and associated with them. Therefore, social conventions are required to introduce conventional “symbols”, “signs”, and “forms of words” to single out exchange promises whenever a case of “interested commerce” takes place. In other words, such linguistic expressions “constitute” a promise, which confirms or, as Hume puts it, “sanctions” a case of “interested commerce” in society. In this way, individuals are able to associate this set of linguistic expressions with an obligation to perform a future action in mutual exchanges. The rule of the obligation of promises consists of fulfilling the kind of obligations that are established and signaled by such social conventions.

Recall that, in the positive part of Hume’s theory of promises, the act of mind that defines a promise has been established as the will to place oneself under an obligation to do something for someone in the future. And, as Hume points out, it would be a “manifest absurdity” in “supposing this act of the mind” as natural of individuals. Hume argues that, by uttering the linguistic expressions that are conventionally associated with a promise, individuals express two mental states: i) the resolution to do something, and ii) the acceptance that they “subject themselves” to a kind of “penalty” in case of non-performance. This acceptance is what properly constitutes the obligation of promises, for it compels individuals to perform their parts of a promise by subjecting them to the sanction of “never being trusted again”. In this way, the obligation associated with the fulfillment of promises is created by the utterance of certain linguistic expressions under the appropriate social circumstances. Promises, which are referred to by such linguistic expressions, become the device by which it is expressed the voluntary of an obligation to
perform a future action. Notice, however, that, in the absence of the linguistic expressions associated with exchange promises, there would be “no more than a resolution in the case”, which is a “natural act of the mind”. The notion of obligation entailed by the act of mind is “new”. For Hume, the obligation and, by implication, the act of mind of willing to undertake it is conventionally created by certain “forms of words”, as soon as individuals realize that, by performing their parts of exchange promises, they can expect of “mutual advantages” from this practice. For Hume, this is how promises are conventionally created and the lack of trust among individuals is bridged.

The conclusion is that Hume’s conventionalist theory of promises depends on the plausibility of some assumptions. First, it depends on the representation of a social structure of interaction as a variation of the basic form of the Farmer’s Dilemma game, which is intended to depict the kind of strategic situation in which mutual exchanges are likely to emerge. Second, it depends on the introduction of social conventions, which are intended to define a set of linguistic expressions as specifying a particular kind of mutual exchange, the exchange promises. In this way, Hume explains the origins both of the notion of obligation related to the performance of promises and the act of mind that defines the promises as established by social conventions. Hume’s conventionalist theory of promises is capable of providing a quite reasonable basis of justification for the practice of promising. Unlike in the natural law theories, which state that promises are morally binding for they are commanded by God as a natural law, Hume assumes that promises are social inventions designed to solve a problem of lack of trust among individuals in society. And, insofar as the social structure of interaction is represented as a realistic picture of this kind of strategic situation, it could be
claimed that individuals have a decisive reason not only to establish but also to maintain the practice of promising.

5.4 Promises, Evolution, Utilitarianism, and the Social Contract

After the exposition of the positive part of Hume’s theory of promises, it is necessary to understand how Hume’s account should be interpreted in light of the three main lines of interpretation. As far as (E.I.) is concerned, it is clear that Hume’s account bears a strong resemblance with an evolutionary perspective. In fact, Hume refers to the gradual development of events related to the emergence of the practice of promising throughout the entire description. For instance, it is strongly suggested that it is the origins of promises that are addressed here when Hume claims that a particular kind of mutual exchanges “begin to take place” and eventually “predominate in society”, which, in turn, calls for a social convention to “distinguish” this kind of mutual exchange from others. Still, it should not be inferred that Hume is making claims about a historically accurate description of the emergence of the practice of promising. As I have pointed out in the third chapter, Hume’s method of explaining the origins of social institutions consists of analyzing them in an evolutionary logic. (E.I.) should be better comprehended as a methodological approach as to how the gradual development of the practice of promising could be thought of having occurred given certain assumptions about human nature. So, it is plausible to argue that Hume’s theory of promises can be accurately described by (E.I.). However, I claim that this interpretation oversimplifies Hume’s point in some ways. The point is not that (E.I.) is not able to capture relevant features of Hume’s theory of promises. Instead, the point is that, if one restricts oneself to (E.I.) alone to describe this account of promises, one neglects other very relevant features of this
description as well. In particular, a purely descriptive account of Hume’s theory of promises overlooks the normative meaning that some notions brought up by Hume in this discussion entail, such as, for instance, the notion that promises are performed in view of “mutual advantage” or that individuals are “bound by interest” to “execute” them. (E.I.) could provide a fairly accurate description of how notions brought up by Hume show up in his theory from an evolutionary perspective of its gradual development, but it could not really describe how they explain the theory in terms of its rationale. In this sense, (E.I.) is a rather thin and superficial interpretation of Hume’s theory of promises.

With regard to the plausibility of (RU.I.), it is useful to start by considering how promises have been explained by the utilitarian theory. In this way, I intend to refer to the traditional distinction in the utilitarian theory between act-utilitarianism and rule-utilitarianism to consider how promises are addressed by this theory. Act-utilitarian accounts of promises are, in general, very problematic. Act-utilitarianism and the moral obligation to keep promises have been traditionally conceived as standing in contradiction. The very fact that the duty to keep promises is morally binding is sometimes brought up as evidence of the refutation of act-utilitarianism. And, it is easy to comprehend the nature of this conflict. The reason why act-utilitarianism fails to provide a proper justification of the moral obligation to keep promises is that, according to its central tenet, an act is considered morally right or morally justified insofar as it is capable of maximizing the overall sum of happiness or welfare of society, or, the social welfare function. Act-utilitarianism claims that, for each given action, the agent must be able to calculate its prospect of generating the highest possible sum of welfare. So, if one maximizes the overall sum of welfare of society by breaking a promise,
then it necessarily follows that one, sometimes, is morally required not to fulfill the duty to keep promises. As far as the moral obligation to keep promises is concerned, this rule is typically regarded as obligatory in a deontological perspective. The intuition is that promises must be kept no matter what. This makes consequentialist justifications of the duty to keep promises somewhat inadequate.

Rule-utilitarianism is introduced to solve this kind of problem. Rule-utilitarianism is a variation of the utilitarian theory that holds that social rules are considered morally right or morally justified insofar as it is the case that each individual, by performing actions in accordance with this set of rules in similar circumstances, can maximize the sum of happiness or welfare of society. With this reformulation, it is claimed that rule-utilitarianism can overcome two challenges that undermine the plausibility of the act-utilitarian account of promises. First, it can explain how the rule of the obligation to keep promises can play a pivotal role in ensuring trust among individuals. In contrast, act-utilitarians would have to accept eventual breaks of promises to maximize the social welfare function whenever it is the case. It is hard to understand how promises can strengthen the bonds of trust among individuals according to this account. Second, it dispenses with the unbearable demand that act-utilitarianism places on individuals in the performance of a utility calculus whenever they have to choose how to behave. Rule-utilitarians emphasize the importance of having a set of rules to specify the kind of morally right or morally justified behavior. This is because rule-utilitarians hold that those who suppose that are able to maximize the overall sum of welfare of society by exploring, as much as possible, the exceptions to the set of rules are more likely to wrongly estimate the social welfare function. Rule-utilitarians assume that individuals are too ignorant of the overall effects of
the utility calculus, all things considered. It is because of the ignorance of the aggregate effects related to the social welfare function that a set of rules is needed.\textsuperscript{129}

Two major objections against rule-utilitarian accounts of promises should be considered here. The first objection claims that rule-utilitarianism and, in fact, utilitarian theories in general, are theories exclusively concerned with the maximization of the overall sum of welfare of society. They are wholly indifferent to the particular distribution of this sum of benefits that make up the social welfare.\textsuperscript{130} For the utilitarian theory, any distribution or redistribution of the sum of benefits that make up the social welfare is acceptable, as long as it is somehow maximized. This is, in fact, a major objection against utilitarian theories, for, once again, if instances of the fulfillment of the obligation to keep promises do less good than to break them, the utility calculus will prescribe its non-performance, even if fulfilling a promise, in this particular instance, may entail a more equitable distribution of the overall sum of social benefits in society.

However, rule-utilitarianism could easily come up with an answer to this objection. In rule-utilitarianism, a set of rules is introduced because the utility calculus, when performed by every individual with the intention of maximization, may not lead to the actual maximization of the social welfare function when all actions are taken together. The uncertainty about whether an action can or cannot, all things considered, maximize the overall sum of welfare of society favors the idea of establishing a set of rules that are known for ensuring the average maximization of the social welfare function. Even if it is clear that particular exceptions of the rule, in the appropriate

\textsuperscript{129} See Atiyah (1981: 83).
circumstances, may maximize the social utility in comparison to the rule, the steady compliance with it can still provide suitable compensation for the eventual losses by ensuring the average rate of maximization of the social welfare function, which, in the absence of the rule, is uncertain to happen. In this way, for rule-utilitarians, a rule of the obligation of promises could, in principle, satisfy both conditions. It could ensure the average rate of maximization of the social welfare, while being, at the same time, the most equitable distribution of the sum of social benefits.

The second objection claims that rule-utilitarianism and utilitarian theories in general, as they are conceived, make the practice of promising ultimately superfluous.\(^{131}\) The idea is that, for rule-utilitarians, individuals are already morally required to orient their behavior according to rules that maximize the overall sum of welfare of society, since they are thus demanded by the principle of utility. So, if it is already required that one should, in general, repay a benefit based only on the principle of utility, quite apart from any promise, then the promise itself does not play any role whatsoever. If it is assumed that it is a moral obligation to respect each others’ expectations in mutual exchanges, then promises are morally superfluous. They do not make any relevant contributions to the current state of affairs from the moral point of view. Rule-utilitarians could still reply by arguing that promises are still important in a derivative sense. They can be useful tools for the clarification and reinforcement of preexisting moral obligations.\(^{132}\) However, this explanation could only work if a whole framework of preexisting moral obligations is already presupposed beforehand, which simply attests to the normative prevalence of the principle of utility in this account. Promises are,

\(^{132}\) See Atiyah (1981: 70).
in this sense, only contingently related to the utilitarian moral framework. They could either exist derivatively or not exist at all.

This is something that addresses (RU.I.) in a fundamental way, for Hume would never admit that promises are superfluous or irrelevant in any sense. They do not play any derivative role in morality. Instead, they play a constitutive role in the convention of justice. Promises are not instrumentally connected to preexisting moral obligations. Instead, they are created as moral obligations by social conventions in light of the necessities and interests of individuals in a society. They play a pivotal role in strengthening the bonds of trust among individuals in society, which enables the emergence of a social structure of interaction that affords mutually advantageous outcomes. It is in this sense that Hume refers to the rule of the obligation of promises as one of the “three fundamental laws of nature”. Thus, (RU.I.) contrasts with Hume’s theory of promises, for the rule-utilitarian theory works just fine regardless of the practice of promising.

It is important to take a look at another consequence of this second objection, which, ultimately, renders (RU.I.) inconsistent with Hume’s theory of promises. Individuals are not only morally required by the principle of utility to behave in ways that correspond to a set of rules designed to maximize the overall sum of welfare in society. The social welfare function also defines the set of interests and preferences that individuals should have in the performance of their moral obligations. And, for rule-utilitarians, the interest or preference that each individual should have in their performance of promises is nothing else than a desire to maximize the overall sum of welfare of society. Interpreting Hume’s theory of promises in this way is wrong. For Hume, the rule to the obligation of promises is created by moderately selfish individuals, who consider only their own interests and
preferences, by means of a social convention established with the expectation of mutually advantageous outcomes. In the following, I will present textual evidence to support this claim.

According to (C.I.), Hume’s theory of promises would be best described if it is conceived as resulting from a hypothetical agreement. In other words, the rule of the obligation of promises should be thought of as earning unanimous assent among everyone in society. And, this rule can only be thought of as earning unanimous assent to the extent that it is expected to yield mutual advantages to each individual. This idea is essential here. It expresses the conceptual entailment between the idea of mutually advantageous outcomes and the idea of unanimous assent. As Gauthier puts it, (C.I.) would state that the rule of the obligation of promises should result from the convergence of individual interests or preferences or, in other words, it requires “interested recognition”. Insofar as this rule is agreed upon based on the interested recognition, individuals are rationally required to adhere to it or, in other words, individuals are bound to it by an “interested obligation”, as they expect that their interests or preferences will be satisfied by complying with this rule.

As I understand it, (C.I.) entirely matches the description provided by Hume in his account of promises. This interpretative claim can be supported by the following passage in the Treatise:

All of them, by concert, enter into a scheme of actions, calculated for common benefit, and agree to be true to their word; nor is there any thing requisite to form this concert or convention, but that every one have a sense of interest in the faithful fulfilling of engagements, and express that sense to other members of the society. (T 3.2.5.11/SBN 522-523)

Hume points out that individuals, “all of them”, “agree” to be “true to their word” in mutual exchanges. And, the condition to form the “concert or
convention” to establish a rule of the obligation of promises is that “every one” has a “sense of interest” in the “faithful” fulfilling their parts of mutual exchange. And, this “sense of interest” in establishing a rule of the obligation of promises is common knowledge, as it is “expressed” to “other members of society”. Clearly, this description entails relevant features of (C.I.). It explains the rule of the obligation of promises as a result of the interested recognition of all individuals. In particular, it depicts this rule as being agreed upon by a “concert or convention”, which reflects the particular convergence of interests of all individuals.

However, the description provided by Hume in his account of promises also brings another relevant feature of (C.I.), as it can be noticed by the following passage in the Treatise:

They are the conventions of men, which create a new motive, when experience has taught us, that human affairs wou’d be conducted much more for mutual advantage, were certain symbols or signs instituted, by which we might give each other security of our conduct in any particular incident. (T 3.2.5.10/SBN 522)

For Hume, the performance of promises is associated with a particular act of mind. The convention creates a new act of mind, that is, the act of mind of willing to undertake an obligation, which, for Hume, properly constitutes a promise. Hume argues that this act of mind, which signals the existence of a particular obligation and is expressed by conventional symbols, signs, or linguistic expressions, is introduced by a convention in view of the expected “mutual advantages”. As it is required by (C.I.), Hume’s theory of promises describes a conceptual entailment between unanimous assent, which is manifested a convention or agreement, and mutually advantageous outcomes. Based on the aforementioned passages, it is Hume’s theory of promises that is more closely related to (C.I.).
I have already hinted that (RU.I.) is not an accurate interpretation of Hume’s theory of promises. Now, in contrast to (C.I.), it is possible to understand why. In the context of the origins and the justification of the rule of the obligation of promises, Hume never mentioned that promises are designed to maximize the social welfare function. Hume’s theory of promises emphasizes mutually advantageous outcomes as a result of the satisfaction of the interests and preferences of all. This is the condition that constitutes the origin and the basis of justification for the rule of the obligation of promises. Unlike in (RU.I.), Hume has a fundamental concern with the distribution of utility, as it is required by (C.I.). No one should end up worse off by the establishment of the rule of the obligation of promises. This is not the case in (RU.I.). The fundamental concern is still with the maximization of the social welfare function, regardless of how the distribution of the sum of social benefits will turn out to be. Under such circumstances, individuals could end up worse off, if, for instance, an alternative rule were to supersede the rule of the obligation of promises for an alleged higher average maximization rate. While in (RU.I.), the coincidence between the maximization of the average rate of welfare and the Pareto criterion is merely contingent, in (C.I.) mutually advantageous outcomes and the impossibility of particular losses are the foundation upon which the whole justification is based.

If (RU.I.) was the accurate interpretation, then Hume’s whole framing of the problem in a kind of strategic situation would be completely dispensable. Hume knows that the challenge is to explain how moderately selfish individuals can trust each other in a social structure of interaction that induces mutual distrust. Individuals take into account their interests in mutual exchanges with others in strategic situations, but they are also able to recognize that the possibility for mutually advantageous outcomes is open to
them if only they can cooperate with each other. In this way, the rule of the obligation of promises emerges as a result of the choices of individuals in a particular social structure of interaction. Promises are a device created with the sole purpose of bridging the lack of trust among individuals. This is why the rule of the obligation of promises is capable of earning unanimous assent from each individual, for they expect mutually advantageous outcomes from this practice. This is the central tenet of (C.I.).

In this chapter, I have analyzed Hume’s theory of promises, which is the other central part of Hume’s theory of justice. First, I have attempted to introduce Hume’s theory of promises in contrast to the account of promises developed by natural law theorists of the 17th century in order to emphasize the historical importance of Hume’s theory in the context of the philosophical thought in his time. Second, I have discussed the negative part of Hume’s theory of promises, which consists of a critique of the natural law theorists based on his naturalistic approach. Third, I have addressed the positive part of Hume’s theory of promises, which was intended to be analyzed in a game-theoretic fashion. In this way, I was able to single out the necessary and sufficient conditions upon which the practice of promising is possible in a social structure of interaction characterized by generalized distrust. This is how the origins and the basis of justification for the practice of promising in Hume can be accounted for. Fourth, I have engaged in the interpretative debate in order to analyze Hume’s theory of promises in light of the three main lines of interpretation. I have advanced the interpretative claim that (C.I.) is the accurate one to describe Hume’s theory of promises. In this way, I have concluded the interpretative analysis of Hume’s theory of justice. I have addressed all the “three fundamental laws of nature” that constitute Hume’s convention of justice from the perspective of the three main lines of
interpretation suggested by Hume scholars throughout history. However, I have not provided a general conclusion of this study yet. In fact, I have come to very different and conflicting conclusions with regard to how each rule should be interpreted in the course of this study. So, in the next chapter, I intend to answer how the interpretative dispute with regard to Hume’s theory of justice can be most reasonably settled.
CHAPTER 6
Naturalism and Hume’s Theory of Justice

6.1 Hume’s Naturalism

In the last chapter, I have concluded the analysis of Hume’s theory of justice. The purpose of this study was to examine the central concept of Hume’s theory of justice, the convention of justice. And, in the course of the last three chapters, I have addressed the set of rules of justice established by individuals through a convention of justice. The “fundamental laws” that comprise Hume’s theory of justice are: i) the rule of the stability of possessions, ii) the rule of transference of private property by consent, and iii) the rule of the obligation of promises. As it is, Hume’s theory of justice is partially constituted by his theory of property and his theory of promises. A fundamental part of this analysis is the interpretative discussion of Hume’s theory of justice in light of the three main lines of interpretation suggested by Hume scholars throughout history. As I have pointed out, the three lines of interpretation are: i) the evolutionary interpretation, ii) the rule-utilitarian interpretation, and iii) the contractarian interpretation.

I have hinted that very different and conflicting points can be inferred about the correct interpretation of Hume’s theory of justice in this study. For instance, at first, (E.I.) seems to be the most natural alternative for a suitable interpretation of Hume’s theory of justice. However, even if (E.I.) is able to portray Hume’s theory of justice in most of its details, it eventually fails to
capture other relevant features of this theory. There is also textual evidence to support both (RU.I.) and (C.I.) as suitable interpretations to describe the set of rules established by the convention of justice, but, with regard to the rules that specify private property rights, the diagnosis is slightly different. Textual evidence indicates that the rules that specify private property rights should be comprehended as based on a combination of imaginative associations and considerations of public utility. Finally, the rule of the obligation of promises is best described by (C.I.), as the textual evidence suggests. This diagnosis raises very relevant questions that need to be answered. Why do the three main lines of interpretation fail to fully explain Hume’s theory of justice? Why have the three main lines of interpretation been so influential among Hume scholars for so long? How can the interpretative dispute about Hume’s theory of justice be most reasonably settled? It is important to understand that the answers to these questions take on greater importance in view of the interpretative claim presented in the first chapter. Recall that I have argued for the interpretative claim that the natural virtues are conceptually dependent on the artificial virtue of justice in Hume’s moral theory. A general sense of virtue can only be inculcated in human beings if the artificial virtue of justice is assumed beforehand. That is why this theory of justice should be considered the focal point of this moral theory. In my interpretation, this theory of justice deserves a prior analysis for any possible interpretative discussion of this moral theory. Thus, I will be concerned with answering these three questions in this chapter.

Clearly, it seems that this interpretative debate cannot be settled from the perspective of the interpretation of Hume’s theory of justice alone. That is because the analysis of Hume’s theory of justice in light of the three lines of interpretations points to very different and conflicting conclusions. This
analysis, by itself, does not allow for a single and definitive conclusion. That is why it is necessary to resort to another strategy to settle this interpretative dispute. My suggestion is that the interpretation of Hume’s theory of justice should be comprehended by reference to a broader interpretative perspective. More specifically, they should be comprehended by reference to Hume’s approach to philosophy in general. The idea is simple. The strategy is to take a step back from the interpretative discussion of Hume’s theory of justice itself and consider it from a more comprehensive interpretative level related to Hume’s philosophical project and his understanding of what philosophy is supposed to be. In a sense, this strategy entails a brief mention of Hume’s metaphilosophy.

Norman Kemp Smith introduced a very influential tradition of interpretation of Hume’s philosophy in the past century. Philosophers like Terence Penelhum and Barry Stroud also contributed to consolidate it in the 20th century. This tradition of interpretation associates Hume’s philosophy with a kind of naturalism. A very broad definition of naturalism characterizes it as a philosophical thesis that rejects that anything supernatural can exist. Naturalism holds that everything that exists can be explained by the physical laws of nature. As a result, naturalism is commonly associated with a scientific view of the world, as it favors a method that seeks precisely to unravel the physical laws of nature. Clearly, Hume endorses the philosophical thesis of naturalism. More specifically, Hume is interested in the study of human nature. That is why Hume sometimes refers to his philosophy as the “science of man” (T xix-xx/3-4). The subjects of the science of man are the powers and faculties of the human mind and cognition, which operate according to psychological laws. And, for Hume, philosophical inquiry is devoted to ascertaining these laws.
Hume’s naturalism is illustrated by his method of surveying human nature. Hume rejects metaphysical conceptions in general. Instead, Hume favors the experimental method used in the natural philosophy of his time, which is based on experimentation and observation. The properties of the bodies and matter are ascertained upon the survey of a sufficiently wide range of experiments. Hume attempts to apply this method in his philosophy in an analogous way. This method is used to survey a sufficient number of instances of human behavior in the widest possible range of circumstances and situations. Hume identifies the most basic types of mental states, such as impressions and ideas. He formulates the psychological laws that define the forms of causal interaction between the mental states, such as the principles of association of ideas. And he sets the principles that systematize and shape human nature, such as habit, sympathy, and other psychological tendencies. So, Hume identifies the simplest elements and explains their possible ways of articulation that build up the most complex phenomena of human behavior. In this way, Hume is able to build a theory that provides a full description of human experience, from the tiny episodes of knowledge to the higher forms of human social organization. And, to the extent that Hume delimits the study of human nature to such facts of observation and experience, it can be concluded that the foundations of Hume’s philosophy are firmly established on the thesis naturalism. As Hume puts it, in the science of man, it is “still certain we cannot go beyond experience” (Txxi/SBN 5).

It is Kemp Smith who best describes the kind of naturalism that characterizes Hume’s philosophy. In the following passage, Kemp Smith portrays the relation between philosophy and nature in Hume:

His entire philosophy, both theoretical and practical, is built around the view of Nature as having an authority which man has neither the right nor the power to challenge. By way of the natural beliefs, Nature has determined the scope and
Kemp Smith argues that the “chief assumption” of Hume’s philosophy is the “authority” of “nature” over “man”. This assumption entails that the truth of some basic facts of nature must be assumed in Hume’s philosophy. As Kemp Smith puts it, the “natural beliefs”, which are inherent tendencies of human beings set by nature, delimit the horizon of human thought and experience. And they also determine the very content of human thought and experience. Causal inference, induction, psychological tendencies to recognize saliences and resemblances, instrumental rationality, the moral significance of utility are all examples of such natural beliefs and inclinations of human beings. They determine the “scope”, “character”, and “possibility” of the “moral standards” and the “theoretical thinking” in Hume. And, for Hume, it is not up to anyone to call the authority of these natural beliefs and inclinations into question here. Nature “has” to be “trusted” and “justifies” this trust, for its authority cannot be challenged by any other ultimate test.

The natural beliefs and inclinations reflect the influence of the natural feelings, passions, instincts, and propensities of human beings for Hume. As Kemp Smith understands it, Hume’s philosophy is characterized by the primacy of feelings over reason. 133 This thesis means that it is a basic fact of nature that human beings, just like other animals, are much more prone to think and behave by feelings, habits, instincts, and propensities than by reason alone. That is the thesis that characterizes of Hume’s naturalism. In this

133 See Kemp Smith (1966: 11).
philosophical system, the more fundamental principles of human nature are such natural beliefs and inclinations, which are defined by instincts and feelings. They illustrate the way by which nature determines the content of human thought and experience.

This kind of naturalism should be understood as a constitutive part of Hume’s philosophy. It should be taken into account in every aspect of his philosophy. In particular, the lines of interpretation of Hume’s theory of justice should be assessed according to the kind of naturalism endorsed by Hume.

6.2 The Descriptive and the Normative in Hume
It is intuitive to think that (E.I.) is the most plausible line of interpretation to describe Hume’s theory of justice. This is suggested by the way in which Hume frames the discussion about justice. Hume is interested in explaining the origins of justice. This account fits very well with the theses of naturalism and evolutionism. On the one hand, it should be borne in mind that Hume puts forward a causal explanation of how increasingly complex social arrangements result from the set of individual behavior of human beings induced by their particular mental states. This description can be said to embody a naturalistic perspective of justice. On the other hand, any explanation of a phenomenon that sets out to trace its origins and describe its gradual development over time refers to evolutionary logic as well. This description also can be said to embody an evolutionary perspective of justice.

There is a distinction between two ways of interpreting Hume’s theory of justice. This distinction concerns the descriptive and normative interpretations of Hume’s theory of justice. (E.I.), as described by Hume scholars, can only be characterized as a descriptive interpretation. And, while
(E.I.) could be able to provide a suitable description of Hume’s theory of justice, it also forsakes any normative character to which this theory can aspire. So, if one endorses (E.I.), as described by Hume scholars, as the best interpretation to account for Hume’s theory of justice, it follows that one commits oneself to the interpretative thesis that this theory is purely descriptive.

As I tried to show in the last three chapters, I claim that (E.I.) is an inadequate interpretation to describe Hume’s theory of justice. This is because (E.I.), as described by Hume scholars, oversimplifies Hume’s theory of justice by neglecting its normative character. This interpretative claim is highly contentious. In fact, the claim that (E.I.) is the correct interpretation of Hume’s theory of justice does some justice to a more systematic comprehension of Hume. Not only is it supported by robust textual evidence in Hume’s theory of justice, but also in Hume’s philosophy in general. Hume likes to think of himself as a kind of scientist or anatomist who seeks to distinguish, classify, and order the “powers of the mind” in a “mental geography” based on “observation” and “experiments” to discover the “secret springs” and “principles” that governs the “human mind” (EHU 1.13-15/SBN 13-14; T 2.1.12.2/SBN 325). This approach might suggest that Hume intended to develop a purely descriptive kind of philosophy.

However, even though Hume is not very fond of using normative vocabulary in his work, it does not follow that he completely dispenses with normative elements in his system. I believe that Hume could not have avoided making normative commitments. As I understand it, Hume makes normative claims, in particular, in his moral theory. Recall that, in the first chapter, I have pointed out that Hume identifies the moral standards by tracing common properties in the multitude of character traits in human behavior, whereby
moral sentiments of approbation and disapprobation are likely to be induced in impartial spectators. The moral standards are defined by the properties of character traits to be useful or agreeable to the individuals themselves or to others, as they either can satisfy their common necessities or can please them for psychological reasons. This kind of empirical survey seeks to describe how moral standards arise from the collective psychology of human beings. Thus, it is an evolutionary or descriptive account of the moral phenomenon. And a purely evolutionary or descriptive account of morality is not itself a normative one.

However, I argue that, for Hume, this evolutionary account of the moral phenomenon can also be understood as normative. I believe that the only strategy to justify this thesis is through a kind of reflective endorsement, as I have pointed out in the first chapter. The idea is that, to the extent that this empirical survey reveals the principles, ways of operation, and benefits of such moral standards to human beings, individuals can endorse them upon a reflection on the causal explanation of their origins. As Hume puts it, according to this “system”, when individuals “reflect” on the “sense of virtue”, they not only “approve” it, but they also approve the “principles” from “whence it is deriv’d”. A reflection on this empirical survey “finds nothing but great and good” in the “rise” and “origin” of this sense of virtue (T 3.3.6.3/SBN 619). According to this passage, it is possible to understand how a purely evolutionary account can entail normative evaluations. The kind of empirical survey conducted by Hume shows how moral standards arise from social psychology, which, in turn, can be reflectively endorsed by individuals and regarded not only as a result of evolutionary development but also as morally authoritative. This is how moral standards are justified in
Hume’s philosophy, as I understand it. And, clearly, Hume is committed to normative claims.

In Hume, the normative character of the moral theory stands out in the theory of justice as well. I have argued that, for Hume, the rules that specify private property rights are introduced by individuals based on a combination of two criteria. On the one hand, by imaginative associations, which are the result of the operations of psychological laws under the appropriate empirical circumstances. On the other hand, by considerations of public utility, which are the result of the moral standard suggested by Hume in his moral theory. It is by reference to these two criteria that individuals select a particular set of rules over another to define the particular distribution of resources among them. That is how the rules of immediate or present possession, occupation, prescription, accession, and succession are chosen to specify private property rights. And, clearly, this is also a normative claim.

These two cases represent clear examples of normative claims made by Hume. However, I make a stronger point here. As I understand it, normative claims underlie Hume’s theory of justice as a whole. Recall that, for Hume, the “sole origin of justice” is the “public utility”. When Hume considers the circumstances of justice, it is argued that justice could only exist if a particular set of empirical conditions were obtained, since, otherwise, it would be simply of no use for individuals. Under other circumstances, it would be irrational for individuals to establish rules of justice. Given a scenario of extreme shortage of goods or extreme selfishness, they would rather be “dictated” by “prudence” and would have “no other expedient than” to fight for their “self-preservation”. Given a scenario of plentiful resources or extensive sympathy, they would dismiss with rules of justice entirely, as they would be “totally useless” to them (EPM 3.3-9/SBN 183-187). So, Hume
supposes that individuals would rationally consider what is useful to them in a wide range of circumstances. By the same token, the decision to establish a convention of justice is a rational choice for them, based on their “understanding” and “experience”, to solve their “necessities” in light of their “inclinations” and “passions” (EPM App. 3.9/SBN 307). In the particular state and condition in which individuals find themselves, a “medium” of “extremes”, the establishment of a convention of justice is a rational choice for them (EPM 3.13/SBN 188). As a result, from the perspective of the individual choice, which is how Hume frames this discussion, each one would be willing to regulate their behavior according to a set of rules, provided that all others do the same. As such, the convention of justice expresses the disposition of individuals for conditional cooperative behavior. Since individuals are willing to benefit and not bearing losses from this social structure of interaction, they would be disposed to comply with the pattern of behavior only if others also do. After all, the convention of justice is agreed upon by individuals with the expectation that this set of rules of justice is able to yield socially advantageous outcomes. Hume’s theory of justice is based on the idea that individuals agree to establish rules of justice by considering what is beneficial to them. Thus, clearly, this theory of justice entails normative claims.

In conclusion, it should be pointed out that, (E.I.) may seem the most intuitively plausible line of interpretation of Hume’s theory of justice, as it satisfies both the theses of naturalism and evolutionism. Both theses are shared by Hume as well. However, it does not follow that (E.I.), as described by Hume scholars, is the correct line of interpretation, for Hume’s theory of justice is a normative account, while (E.I.), in such terms, is not. By restricting itself to a purely causal or descriptive perspective of Hume’s theory of justice,
(E.I.) is a thin and superficial interpretation of it. It does not capture Hume’s theory of justice to its full extent. Hume’s theory of justice is an example of an evolutionary account that is, nonetheless, normative as well. This point is entirely missed by (E.I).

6.3 The Public and the Private Utility in Hume

With regard to the normative interpretations of Hume’s theory of justice, the alternatives are (RU.I.) and (C.I.). While both (RU.I.) and (C.I.) are able to capture the normative character of Hume’s theory of justice, they are also problematic in other aspects. In fact, perhaps (RU.I.) and (C.I.) are the lines of interpretation of Hume’s theory of justice that misunderstand Hume’s philosophical approach the most. As I understand it, both (RU.I.) and (C.I.) do not take fully into account the thesis of naturalism in Hume. They neglect the full extent of the effects of the mental states of individuals in Hume’s philosophy. Thus, both (RU.I.) and (C.I.) provide descriptions of Hume’s theory of justice that do not fit with a more systematic comprehension of his philosophy in general.

The problem with (RU.I.) and (C.I.) lies in the concept of public utility in Hume’s theory of justice. Recall that, for Hume, individuals agree upon a set of rules of justice with the expectation that, by complying with the pattern of behavior defined by them, they can reap socially advantageous outcomes. In this theory, justice is based solely on the public utility that it is able to afford. The question is how one should interpret the concept of public utility in Hume. Both (RU.I.) and (C.I.) interpret Hume’s concept of public utility as a kind of function of individual utility functions. Still, they differ in how this function is supposed to be constructed in Hume. While (RU.I.) understands Hume’s concept of public utility as a kind of social welfarism,
(C.I.) interprets it as equally representing the individual utilities of all. More specifically, for (RU.I.), public utility amounts to the maximization of the overall sum of welfare in society, which is a function of the aggregation and ordering of all subjective utility functions in one single preference ranking. For (C.I.), public utility amounts to the satisfaction of each of all different subjective utility functions considered separately. There is textual evidence for both lines of interpretation in Hume.

On some occasions, Hume claims that the conventions of justice is agreed by individuals based on a sense of self-interest, which is supposed to be common to all of them. This entails that the set of rules of justice satisfies each of all different subjective utility functions. Hume reinforces this point in his theory of promises by claiming that all agree to establish a rule of the obligation of promises, which, together with the other rules of justice, constitute a system of actions designed for the benefit of each individual. Such points speak for (C.I.), for the concept of public utility is built as a function that assigns equal weight to the satisfaction of the interests and preferences of all.

On other occasions, Hume refers to possible incongruities between single acts of justice and the system of actions that comprises the practice of justice, which, in turn, suggests a conflict between the utility afforded by a single act of justice and public utility in general. Still, Hume argues that, by balancing the average losses and benefits, individuals should still regard themselves as gainers in the trade-off of the grand scheme of justice. Hume reinforces this point in his theory of property by claiming that rearrangements of the distributive pattern of society should take into account the interests and preferences of a number of individuals just sufficient to strengthen the state or public utility. Such points speak for (RU.I.), for the concept of public utility
is comprehended as a function of the average of individual utilities or, the greatest satisfaction for the greatest number of individuals.

This lack of clarity with regard to the concept of public utility is the reason why there is textual evidence to support both lines of interpretation in Hume. This ambiguity has induced scholars to dispute over (RU.I.) and (C.I.) as the most suitable line of interpretation of Hume’s theory of justice for a long time. I believe that the root of this ambiguity lies precisely in the two perspectives from which Hume addresses the convention of justice. In the course of this account, it can be observed that Hume shifts the focus from the perspective of the individual choices to the perspective of the scheme of justice itself. On the one hand, in most of the Treatise Book 3, Part 2, Section 2, and in most of the Enquiry Section III and Appendix III, the convention of justice is addressed from the perspective of the individual choices. When this is the case, the set of rules of justice is imagined as a result of a convergence of all the individual utilities, since it is agreed upon by individuals with the expectation of mutually advantageous outcomes. In other words, the convention of justice can be said to be established based on the interested recognition, as it is required by (C.I.). The concept of public utility in Hume could be, then, adequately reduced to the individual utilities of all, insofar as they are rationally considered. On the other hand, in paragraph 22 of the Treatise Book 3, Part 2, Section 2, and in paragraphs 3, 5, and 6 of the Enquiry Appendix III, the convention of justice is addressed from the perspective of the scheme of justice itself. When this is the case, Hume seems to disagree that the public utility coincides with the set of shared and common individual interests. The set of rules of justice are characterized by having the property of affording greater social utility by balancing out the eventual particular losses. This indicates that the concept of public utility is not identified by the
individual utilities of all, but by the greater social utility, which is more the result of a positive trade-off of utilities, as it is required by (RU.I.).

As I interpret it, Hume is simply not clear enough in his concept of public utility for it to be conclusively comprehended either according to (RU.I.) or (C.I.). This ambiguity and lack of clarity hamper the rational reconstruction of this concept in Hume’s theory of justice. And the fact that the concept of public utility in Hume is in such a way underdetermined has led scholars who support both (RU.I.) and (C.I.) to argue unsuccessfully about this point for a long time already.

However, there is another problem with (RU.I.) and (C.I.). This problem refers to the misunderstanding of the thesis of naturalism in Hume’s philosophy. I believe that scholars who support (RU.I.) and (C.I.) are guilty of this failure. The thesis of naturalism in Hume’s philosophy is characterized by the basic fact of nature that human beings are more likely to think and behave by feelings, habits, instincts, and propensities than by reason alone. This basic fact of nature defines a wide range of natural beliefs and inclinations that ultimately shape human thought and experience. Under this assumption, it is expected that natural propensities and instinctive imaginative associations can influence the deliberation process of individuals as well. This is the case of the rules that specify private property rights in Hume’s theory of justice, which are partially defined by imaginative tendencies of human beings to recognize salience to settle distributive problems concerning the acquisition of goods. And, strictly speaking, this is not a rational response in a sense required by (RU.I.) and (C.I.). It does not necessarily maximize the overall sum of welfare in society or satisfy the interests and preferences of all. This illustrates the role played by the psychological effects of the different
mental states of individuals in Hume’s theory of justice, which both (RU.I.) and (C.I.) seem to neglect.

As it is, both (RU.I.) and (C.I.) do not seem to provide appropriate descriptions of Hume’s theory of justice. Not only does the controversy between (RU.I.) and (C.I.) seem to end in a deadlock, but scholars that support both lines of interpretation seem to miss why it is the case. This is because they fail to fully understand the main concept of Hume’s theory of justice, the concept of public utility. Besides, both (RU.I.) and (C.I.) seem to overlook the full implications of Hume’s naturalism in his theory of justice. Hume’s theory of justice involves elements that can only be ascertained by examining more carefully the psychological effects of the different mental states of individuals. By neglecting Hume’s naturalism, both (RU.I.) and (C.I.) fall short of representing the overall picture of Hume’s philosophy. The two lines of interpretation of Hume’s theory of justice do not fully harmonize with a more comprehensive and systematic level of understanding of Hume’s philosophical project. Ultimately, both (RU.I.) and (C.I.) cannot be regarded as suitable interpretations of Hume’s theory of justice.

6.4 Degrees of Plausibility and the Interpretation of Hume’s Theory of Justice

After considering how the three main lines of interpretation of Hume’s theory of justice square with Hume’s philosophical project, it is possible to answer the three questions that were introduced in this chapter. First, why do the three main lines of interpretation fail to fully explain Hume’s theory of justice? The answer to this question is that each of the three lines of interpretation seems to fall short of capturing relevant elements of Hume’s theory of justice. While (E.I.), as described by Hume scholars, seems to neglect the normative
character of Hume’s theory of justice, both (RU.I.) and (C.I.) do not seem to correspond to a more comprehensive interpretative level. As Hume’s theory of justice must be comprehended as a part of Hume’s philosophy in general, it follows that the interpretation of the theory of justice must take into account not only the normative character of this theory but also the thesis of naturalism in Hume.

Second, why have the three main lines of interpretation been so influential among Hume scholars for so long? The answer to this question is that Hume scholars have long struggled to articulate the concepts of this theory in one single plausible description capable of matching the textual source. In addressing this theory of justice, scholars have not only made mistakes. They have also made important contributions to a better comprehension of it over time. They addressed Hume’s theory of justice from the different perspectives adopted by Hume in the course of his theory. And, indeed, the three lines of interpretation managed to accurately describe some of the features of Hume’s theory of justice, which has encouraged this interpretative dispute for such a long time. Still, they have failed to fully account for it eventually. That is why the three main lines of interpretation have been so influential for so long.

Third, how can the interpretative dispute about Hume’s theory of justice be most reasonably settled? First, the answer to this question depends on recognizing that neither line of interpretation, as described so far, is entirely correct to describe Hume’s theory of justice. Second, it should be noticed that even though none of them can provide the most accurate description of Hume’s theory of justice, they still capture relevant points of it. This conclusion can pave the way for developing a correct interpretation of Hume’s theory of justice. Therefore, to settle this interpretative dispute in
the most reasonable way, it is necessary to consider how the three lines of interpretations can contribute to the accurate description of Hume’s theory of justice.

Here, the best interpretative strategy seems to be to examine, compare, and assess the three main lines of interpretation by their degree of plausibility in face of the textual evidence in Hume. By comparing the strengths and weaknesses of each line of interpretation, it is possible to find out which one is the closest to describing Hume’s theory of justice and to which extent it should be corrected or developed in order for the correct interpretation to be obtained. Now, in terms of the degree of plausibility, the rank order of the lines of interpretations seems to be as such: first (E.I.), second (C.I.), and third (RU.I.). I will consider the reasons for this particular rank order.

(E.I.) stands out as the most plausible one to describe Hume’s theory of justice. It focuses on the kind of account favored by Hume, namely, to explain social phenomena by their origins and gradual development. This kind of account is used by Hume to explain the origins of morality, the virtue of justice, and the government. As such, this account requires a causal explanation of the effects of the mental states of human beings, given the standard empirical circumstances in which they find themselves, in the establishment of the rules of justice and the disposition to comply with the pattern of behavior defined by them. In this sense, (E.I.) addresses the rules of justice from a causal or descriptive perspective. And this perspective fits very well with Hume’s naturalism as well, which makes (E.I.) a very plausible candidate for the correct interpretation of Hume’s theory of justice. What speaks against (E.I.), as described by Hume scholars, is the fact that it fails to account for the normative character of Hume’s theory of justice.
(C.I.) is also a fairly plausible interpretation of Hume’s theory of justice. It addresses the rules of justice from the perspective of individuals who consider only their rationally considered interests and preferences. More specifically, it addresses the establishment of the rules of justice and the disposition to comply with the pattern of behavior defined by them as kinds of rational choices of each individual, which, in the case of a social structure of interaction involving multiple players, amounts to a game-theoretical account. Hume conducts most of his account of justice from the perspective of individual choices. What speaks against (C.I.) is the fact that Hume does not adopt only one single perspective in his theory of justice.

(RU.I.) is the least plausible interpretation of Hume’s theory of justice. It reiterates that the establishment of the rules of justice and the disposition to comply with the pattern of behavior defined by them are grounded on public utility. But, it interprets public utility as the overall sum of social welfare or the greatest satisfaction for the greatest number of individuals. In other words, Hume addresses the rules of justice from the perspective of the global welfare afforded by the grand scheme of justice. This line of interpretation is supported only by a few passages in Hume’s entire work. As the analysis carried out so far seems to suggest, (RU.I.) is the least plausible line of interpretation to describe Hume’s theory of justice, which entails that it must be cast aside here.

Now, by comparing the lines of interpretation, it should be noticed that both (C.I.) and (RU.I.) are incompatible with each other. The normative theories suggested by both lines of interpretation entail two very different conceptions of distribution of social advantages yielded by the rules of justice. This is because they interpret the concept of public utility in Hume in two different ways. And, in most of his theory of justice, Hume does not adopt the
kind of perspective that favors (RU.I.). Instead, (RU.I.) relies on very limited textual evidence in Hume’s works. Thus, it follows that (RU.I.) is less plausible than (C.I.) as a candidate for the correct line of interpretation of Hume’s theory of justice. (RU.I.) must not be considered an appropriate baseline for developing the correct interpretation of Hume’s theory of justice.

Thus, (C.I.) emerges as a more plausible line of interpretation to describe Hume’s theory of justice in comparison to (RU.I.). In terms of the degree of plausibility, (C.I.) is favored over (RU.I.) because Hume adopts the perspective of the individual choices in most of his theory of justice. The discussion is strongly framed by how individuals, motivated by their self-interest, would choose to behave to better satisfy their interests. Unsurprisingly, the idea of mutually beneficial outcomes becomes central for this line of interpretation. (C.I.) is supported by more robust textual evidence than (RU.I.). However, as tempting as it is to take (C.I.) as the baseline for developing a correct interpretation of Hume’s theory of justice, the fact that Hume adopts multiple perspectives in this discussion discourages this conclusion. This suggests that there may be a better line of interpretation to serve as a starting point to find out the correct interpretation of Hume’s theory of justice.

The case is different with (E.I.). This line of interpretation seems to provide a description that best explains Hume’s naturalism. And, from a structural point of view, (E.I.) is a far more comprehensive interpretation than both (C.I.) and (RU.I.). Strictly speaking, the former one is making a very different point from that made by the latter ones. Whereas for (E.I.), Hume intends to describe how this particular set of rules of justice could have emerged in society, for (C.I.) and (RU.I.), Hume intends to justify this particular set rules of justice in society. Nonetheless, I believe that some
overlap between evolutionary and normative theories is possible. At least, it does not necessarily seem to imply a logical contradiction that normative theories can take advantage of an evolutionary approach. Unlike the case of (C.I.) and (RU.I.), (E.I.) is not, in principle, incompatible with the other lines of interpretation. These features of (E.I.) make it a far more promising baseline for developing the correct interpretation of Hume’s theory of justice.

So, the clue to settle this interpretative dispute is (E.I.). Since (E.I.) seems to be a more comprehensive line of interpretation than the others, it could be considered, in terms of the degree of plausibility, the closest one to describing Hume’s theory of justice. (E.I.) seems to have more explanatory power than other interpretations. (E.I.) seems to take Hume’s naturalism fully into account, while both (C.I.) and (RU.I.) do not. And, while both (C.I.) and (RU.I.) are mutually exclusive, (E.I.) does not necessarily seem to conflict with neither one of them in particular. Still, (E.I.), as described by Hume scholars, does not explain the normative character of Hume’s theory of justice. But, it can serve as an appropriate baseline for developing the correct interpretation of Hume’s theory of justice. It seems that understanding how the correct interpretation of Hume’s theory of justice is possible depends on understanding how the evolutionary account that characterizes the baseline interpretation can be normative as well. After all that has been considered in this study, it should not come as a surprise that Hume’s moral theory could be regarded as entailing normative claims. However, Hume scholars who endorsed (E.I.) never really considered that Hume’s theory of justice could be normative in the first place. And this is a fundamental mistake.

The correct interpretation of Hume’s theory of justice can be defined as a kind of normative version of (E.I.). So, the causal or descriptive perspective that characterizes (E.I.) should be regarded as a kind of normative
account as well. This interpretation claims that the establishment of the rules of justice and the disposition to comply with the pattern of behavior defined by them in Hume should be interpreted as the result of the causal effects of a wide range of the mental states of individuals under standard empirical circumstances. Common necessities or desires, enlightened self-interest, imaginative associations, habits, are all kinds of mental states that, to varying degrees, causally influence individuals in establishing the rules of justice and creating the virtue of justice. Of course, the importance of each mental state varies depending on the extent of its causal influence. For instance, according to this account, enlightened self-interest is expected to play a much bigger role than the moral sentiments in the establishment of the rules of justice and the virtue of justice.

I argue that, for the sake of consistency, the strategy for interpreting (E.I.) as a kind of normative account should be the same for interpreting Hume’s moral theory as a normative account. After all, for Hume, the virtue of justice is just a kind of virtue like any other that makes up our sense of virtue. Thus, the strategy should resort to the idea of reflective endorsement, as discussed in the first chapter. Recall that, for Hume, justice is grounded on public utility. It is a social creation designed to promote socially advantageous outcomes. The account of the origins and foundations of justice must be able to be reflectively evaluated, as this study is capable of laying out the causes and circumstances under which justice, as a socially beneficial practice, is possible. More specifically, this study is intended to disclose the causal articulation between the psychological causes and the external circumstances that would enable both the establishment of the rules of justice and the disposition to comply with the pattern of behavior defined by them. The workings of the passion of self-interest, the faculty of reason, and the
imaginative associations, in the context of standard empirical circumstances, could be reflectively endorsed by individuals, as they would be aware of the role played by these mental states in the establishment of a practice that yields socially beneficial outcomes. To the extent that the outcomes generated in this way are good and beneficial and to the extent that individuals understand that these mental states are absolutely necessary to afford them, these mental states can be reflectively approved. The virtue of justice is clearly different from most of the other virtues in Hume, for it is based only on the utility it is able to afford. But, just like any other virtue in Hume’s moral theory, the virtue of justice could also be considered a normative practice, for it could be approved by a proper reflective activity on the principles that govern it. And they can only be unveiled by the kind of account developed by Hume.

This interpretation differs from the others. Recall that both (C.I.) or (RU.I.) rely on the same textual evidence and end up coming to conflicting conclusions. This is because they focus on two different perspectives adopted Hume in the course of his theory of justice. So, we should remain skeptical about the plausibility of either (C.I.) or (RU.I.) as suitable interpretations of Hume’s theory of justice. And, although the normative version of (E.I.) can strongly coincide with (C.I.) in most of Hume’s account of justice, the justification that both provide is quite different. (C.I.) claims that Hume’s theory of justice should be imagined as the result of a unanimous assent, as it expresses a convergence of the interests and preferences of all. The normative version of (E.I.) claims that Hume’s theory of justice should be interpreted as an account that, in describing all the causes and circumstances for the emergence of justice, allows this the set of causes to be endorsed and approved by a further reflection on their beneficial effects. Unlike with (RU.I.), both the normative version of (E.I.) and (C.I.) overlap most of the
time. Both emphasize the pivotal role of the rational self-interest of individuals in Hume’s theory of justice. And both regard Hume’s theory of justice as a normative account, as the practice of justice is able to satisfy our common interests. Still, they express two different kinds of justification. They express two different ways of assessing the rationality of a given account. As far as this interpretation of Hume’s theory of justice is concerned, I believe that it could be properly defined as a normative extension of (E.I.), which can only be formulated if one accepts that Hume’s moral theory can entail normative claims. This is a view that previous Hume scholars have either denied or neglected. As such, I believe that this interpretation best combines the thesis of naturalism and the normative character that Hume’s theory of justice seems to incorporate.

Now, there is another brief point, not strictly related to the analysis of the textual evidence in Hume, that may also reinforce my interpretive claim. This point addresses a historiographical argument. The historical period in which Hume lived and developed his philosophy, the mid-eighteenth century in Great Britain, is a very particular moment in the history of philosophy. About ninety years after the publication of the Leviathan by Hobbes and fifty years after the publication of the Second Treatise of Government by Locke, the Treatise of Human Nature is published by Hume divided by the first two volumes, that came out in 1739, and the third volume, that came out in 1740. Hume was well acquainted with the social contract theories that featured in the discussions about political philosophy in his time. This is because Hobbes and Locke had a significant influence not only on Hume but also on all other philosophers in this period. The social contract theory shaped most of the debate about political philosophy in the 18th century.
However, Hume did not really endorse any kind of social contract theory in his time, as it is evidenced by the criticism raised by him in some of his *Essays*. Instead, in the mid-eighteenth century, Hume’s main influences were sentimentalist philosophers like Joseph Butler and Shaftesbury. They developed a kind of ethical naturalism based on the moral sentiments of human beings. This is the approach that Hume embraces in his moral theory. So, even if Hume’s theory of justice might be described as a kind of social contract theory by a particular line of interpretation, it must be clear that Hume never really adopted the social contract theory as the main influence of his thought.

About 15 years after Hume’s death, *An Introduction to the Principles of Morals and Legislation* was published by Bentham. It was the first time that a fully developed utilitarian theory has been formulated. Previous philosophers have already been working on the basic intuitions for such a theory, but it was Bentham who has first articulated them in a truly systematic way. Bentham is known to have been influenced by Hume in recognizing the moral value of the property of things to have the tendency to promote certain interests or ends. The principle of utility is introduced by Bentham as a norm that commands actions to be oriented to maximizing the social utility they are able to afford. This principle is the central concept of the utilitarian theory. For this reason, Hume is sometimes characterized as holding a proto-utilitarian position. So, even if Hume has not developed a kind of utilitarian theory, his philosophy was an important influence for utilitarian philosophers, in particular, for Bentham.

Notice, then, that Hume is situated in a historical period that is characterized by a convergence of different trends in moral philosophy. It is plausible to argue that either (E.I.), or (RU.I.), or (C.I.) could qualify as
plausible candidates to describe Hume’s theory of justice. However, I argue that the most plausible assumption is that Hume has developed his theory of justice influenced by a plurality of different moral theories. This might explain why Hume adopts multiple perspectives in the course of his theory of justice. And some of Hume’s concepts were also incorporated by later philosophers in their moral theories, which has strengthened their association with Hume. I believe that this point speaks against the thesis that either (E.I.), or (RU.I.), or (C.I.), as described by Hume scholars, are suitable to describe Hume’s theory of justice. Rather, this point speaks for the thesis that Hume’s philosophy is a function of different influences of this particular historical period. I believe that this historiographical point should not be underestimated. It strengthens the case for the thesis that Hume’s theory of justice is better interpreted by a kind of interpretation that should be developed over the others so far suggested, as I am arguing here.

6.5 A Final Remark

This interpretation of Hume’s theory of justice is supposed to shed some light on the concept of virtue in Hume. It could be said that, in light of the interpretative claim made in the first chapter, namely, that the natural virtues are conceptually dependent on the artificial virtue of justice, this interpretation is of the utmost importance to clarify the essential concept of Hume’s moral theory. This interpretation is supposed to explain how a general sense of virtue in human beings is possible in the first place. It is supposed to explain the conditions that must be satisfied for an intersubjective social perspective to emerge. In turn, this intersubjective social perspective ensures the impartial stance of the general point of view, implied in the very concept of virtue and vice in general.
The advantage of this interpretation is that it allows us to consider the full extent of the effects of the mental states of individuals in the establishment of a set of rules that enables social coexistence. In contrast to the other lines of interpretation, the normative version of (E.I.) emphasizes most strongly Hume’s naturalism in his theory of justice. This interpretation highlights the role of all mental states of human beings in the establishment of the rules of justice. And it regards the influence of these mental states as justified in view of the socially favorable effects afforded by the rules of justice. For Hume, justice arises from a convention or agreement among individuals motivated by their enlightened self-interest in circumstances of confined generosity and moderate scarcity of provisions or goods. Individuals, motivated by their enlightened self-interest, agree to establish a set of socially advantageous rules, the rules of justice, to govern the terms of social coexistence. By implication, the disposition to regulate one’s conduct to the pattern of behavior defined by this set of rules, which may be properly referred to as the virtue of justice, is motivated by their enlightened self-interest as well, as I have argued in the third chapter. But, although the passion of self-interest is the most important element of this theory, it is not enough to account for the whole phenomenon of justice in Hume. For additional rules of justice to be established, the enlightened self-interest has to be supplemented by other motivational sources, such as, for instance, imaginative associations. For Hume, imagination plays a role in the selection of additional rules of justice for solving particular distributive problems. And, the virtue of justice, which is introduced and established by the enlightened self-interest, is gradually reinforced by other mental states such as the habit and the moral sentiments developed by individuals towards this social practice. Thus, it is this articulation of these mental states, in particular,
between the passion of self-interest and imaginative associations, that matters the most in the description provided by Hume’s theory of justice.

A general sense of virtue, which involves the disposition of individuals to behave and recognize as virtuous or vicious certain character traits of agents, whoever they may be, depends on a wide articulation of the various mental states, which are specifically aimed at establishing a particular set of rules. It depends on the enlightened self-interest, psychological propensities, imaginative associations, and so on. Other lines of interpretation hold that a general sense of virtue depends either on reasons for maximizing the overall sum of social welfare, or reasons for satisfying rational self-interest, or a pure description of the operation of these mental states. It seems to me that each of these lines of interpretation is too focused on just one single perspective adopted by Hume in the course of his exposition. I suggest that Hume’s theory of justice should be fully taken into account and evaluated as a whole. In this way, one can understand and evaluate how all mental states contribute to the emergence of justice. As a result, it is also possible to assess, in light of this description, which lines of interpretations are more plausible, such as (E.I.), or even (C.I.), and which are less plausible, such as (RU.I.). In a sense, it is not surprising to find out that the theory of justice parallels the moral theory in Hume if it is accepted that this normative version of (E.I.) is the correct interpretation of Hume’s theory of justice. On the one hand, Hume’s theory of justice is supposed to explain the gradual development of a set of rules that enables social coexistence. On the other hand, Hume’s moral theory is supposed to explain the gradual enrichment of the sentimental framework of each individual and the full flourish of a sense of virtue in human beings, which can only take place under a social coexistence governed by the rules of justice. And, as I interpret them, both accounts entail normative
claims. I believe that these similarities clarify the relationship between justice and a general sense of virtue.

In this sixth and last chapter, I have attempted to settle the interpretative dispute about Hume’s theory of justice. First, I have summarized the ways by which Hume’s theory of justice can be described in its different parts by the three main lines of interpretations suggested so far, even though such interpretations, as described by Hume scholars, are fundamentally inconsistent with each other. I have diagnosed that this interpretative debate could not be settled from the perspective of the interpretation of Hume’s theory of justice alone, because each of the three lines of interpretation conflicted, to some extent, with Hume’s text. So, I have suggested that this interpretative dispute could be solved by making reference to a higher and more general interpretative perspective, the perspective of Hume’s metaphilosophy. Second, I have proceeded to argue that Hume’s philosophy is committed to a kind of naturalism. A naturalistic approach underlies Hume’s philosophy as a whole. I have attempted to prove this point by showing that Hume’s philosophy is best characterized by the thesis of naturalism, for Hume’s main project, as I understand it, was to undertake a philosophical investigation about the fundamental principles of human nature. As a result, any kind of interpretation intended to accurately describe Hume’s theory of justice must harmonize with the thesis of naturalism that characterizes his philosophy in general. Third, I have attempted to outline the limits of the descriptive and the normative in Hume’s philosophy, which is particularly relevant for the plausibility of (E.I.). Fourth, I have analyzed the problematic concept of public utility in Hume in face of the plausibility of (RU.I.) and (C.I.). Fifth, in light of the conflicting interpretative conclusions so far, I have suggested that Hume’s theory of justice can be better
comprehended if the three main lines of interpretations are assessed by their degrees of plausibility in face of the textual evidence. By comparing the three main lines of interpretation, I have concluded that (E.I.) is the most plausible one to serve as an appropriate baseline for developing the correct interpretation of Hume’s theory of justice, as it comes the closest to accurately describing it. After having expanded on this particular line of interpretation, I have claimed that a normative version of (E.I.) could qualify as the correct interpretation of Hume’s theory of justice. It draws on the most appropriate baseline interpretation and provides the normative description entailed in Hume’s moral theory. To sum it up, the correct interpretation of Hume’s theory of justice is a normative version of (E.I.), which is able to combine the two most important elements of Hume’s theory of justice: his naturalism and his normative character.
Conclusion

This study showed that the biggest challenge in interpreting Hume’s theory of justice is to make sense, in a single and coherent account, of the multiple perspectives that Hume adopts in the course of his theory. And, depending on which perspective one decides to focus, one comes to endorse a particular line of interpretation. Consequently, I claimed that the three main lines of interpretation suggested by Hume scholars reflect the three different perspectives adopted by Hume in his theory. In this sense, there are as many lines of interpretation as there are perspectives that can be identified in Hume’s account. The three main lines of interpretation are: i) the evolutionary interpretation, ii) the rule-utilitarian interpretation, and iii) the contractarian interpretation. (E.I.) reflects a purely causal or descriptive perspective. (RU.I.) reflects the perspective of the global advantages afforded by the scheme of justice itself. And (C.I.) reflects the perspective of individual choice. So, this study comes down to the analysis of the rules of justice in Hume to find out from which perspective Hume considered the different parts of his theory. Consequently, in this study, I analyzed Hume’s theory of justice by assessing each of the three rules that constitute this theory in light of the three lines of interpretation.

In the first chapter, I started the discussion by examining the relationship between the moral theory and the theory of justice in Hume. I argued that this relationship is characterized by a kind of conceptual dependence that natural virtues hold in relation to the artificial virtue of
justice in Hume. This claim implies a radical interpretative change. As a result, this entire moral theory must be interpreted from the perspective of the artificial virtue of justice. As a result, the most important interpretative task in this moral theory is to clarify the nature of the artificial virtue of justice.

In the second chapter, I pointed out that three philosophical assumptions of Hume’s moral theory set the framework in which individuals agree to establish a set of rules of justice by an agreement or convention. These assumptions are: i) the instrumental conception of practical reason, ii) the artificial nature of the virtue of justice, and iii) the circumstances of justice. These three assumptions define the features of the Humean state of nature, which is the pre-conventional structure of interaction. I emphasized that Hume addresses this framework from the perspective of the individual choices, as the establishment of justice is characterized as the most useful alternative for individuals given the conditions in which they find themselves. Since this framework involves multiple individuals, this perspective invited a game-theoretic account of justice. And, by considering the similarities between the Humean state of nature and the Stag Hunt, I came up with the solution for the problem posed by this game theory model. This solution, which is a description of a given social structure of interaction, was supposed to orient the discussion in the third chapter.

In the third chapter, I addressed the origins of the convention and the virtue of justice in Hume. For this, I examined Hume’s convention of justice, which is the central concept of Hume’s theory of justice. After all, the virtue of justice, which is the pattern of compliance behavior with the rules of justice, can only be consolidated after the establishment of the convention of justice. I argued that the enlightened self-interest is the motive that induces individuals to establish a convention of justice and to behave according to the
pattern of behavior defined by its rules. This enlightened self-interest is nothing more than a restrained sense of the passion of self-interest, which is supposed to be redirected towards a more socially beneficial pattern of behavior, namely, the establishment of the rules of justice and the pattern of compliance behavior with the rules defined them. I also analyzed the nature of the convention of justice by examining the definitions provided by Hume. The conclusion is that the convention of justice is a kind of agreement that incorporates the most important elements described in the solution for the Stag Hunt game model suggested in the second chapter. Finally, I interpreted Hume’s account of the convention of justice in light of the three lines of interpretation of Hume’s theory of justice. I attempted to show that each line of interpretations derives from a specific perspective adopted by Hume in the course of his theory.

In the fourth chapter, I considered additional rules of justice. In particular, I addressed the rules of justice that specify private property rights in Hume, which are introduced by the first rule of justice, the rule of the stability of possessions. For Hume, these rules govern the distribution and acquisition of private property in a society. I claimed that the rules that define the initial distribution of possessions, the rules of occupation, prescription, accession, and succession, and the rule of the translation of property by consent are all determined, to varying degrees, by a combination of considerations of public interest and imaginative associations. I also argued that Hume’s theory allows for a redistributive rule to be introduced, which is designed to mitigate sensitive social inequalities that could emerge from the given distributive order. This rule enables changes in the distributive pattern of society. I pointed out that Hume introduces the rules that specify the distribution and acquisition of private property both from two perspectives:
the causal or descriptive perspective and the perspective of the global advantages afforded by these rules, which favors (E.I.) and (RU.I.). Like the rule of the stability of possessions, this set of rules comprises Hume’s theory of property, which is a central part of Hume’s theory of justice.

In the fifth chapter, I addressed the rule of the obligation of promises. This rule is related to Hume’s theory of promises, which comprises another central part of Hume’s theory of justice. I pointed out that Hume’s theory of promises is very relevant in the history of moral philosophy. I introduced this account in contrast to a more traditional account of promises in Hume’s time, which is formulated by the natural law theory, to illustrate the differences between both theoretical accounts of promises. I claimed that Hume explains the rule of the obligation of promises from the perspective of the individual choices. This frames the discussion in terms of the rational choice that individuals would make under these circumstances. As such, Hume goes as far as to describe promises in a way that it could be modeled as a game-theoretic account. For this reason, Hume’s theory of promises seems to be best described by (C.I.).

In this sixth chapter, I summarized and assessed how Hume’s theory of justice relates to each line of interpretation. I argued that this interpretative controversy can only be settled by considering each line of interpretation in terms of its degrees of plausibility in face of the textual evidence in Hume. This study concluded that, according to this criterion, a normative version of (E.I.) qualifies as the most plausible interpretation of Hume’s theory of justice. So far, the scholars who endorsed (E.I.) have never considered that this line of interpretation could entail normative claims, which, as I understand, is an indication that they have failed to comprehend Hume’s philosophy in a more comprehensive and systematic way. What distinguishes
my interpretation from the traditional (E.I.) is that, in resorting to the idea of a reflective endorsement, which finds textual support in Hume, I attempt to apply the normative logic underlying his moral theory to his theory of justice as well. As a result, the normative version of (E.I.) is the description of Hume’s theory of justice that best conciliates the naturalism and the normative character in Hume’s moral theory in general.
References:


