RMM Vol. 5, 2014, 47–63
Special Topic: Can the Social Contract Be Signed by an Invisible Hand?
http://www.rmm-journal.de/

Horacio Spector

Hume's Theory of Justice[*]

Abstract:
Hume developed an original and revolutionary theoretical paradigm for explaining the spontaneous emergence of the classic conventions of justice—stable possession, transference of property by consent, and the obligation to fulfill promises. In a scenario of scarce external resources, Hume's central idea is that the development of the rules of justice responds to a sense of common interest that progressively tames the destructiveness of natural self-love and expands the action of natural moral sentiments. By handling conceptual tools that anticipated game theory for centuries, Hume was able to break with rationalism, the natural law school, and Hobbes's contractarianism. Unlike natural moral sentiments, the sense of justice is valuable and reaches full strength within a general plan or system of actions. However, unlike game theory, Hume does not assume that people have transparent access to the their own motivations and the inner structure of the social world. In contrast, he blends ideas such as cognitive delusion, learning by experience and coordination to construct a theory that still deserves careful discussion, even though it resists classification under contemporary headings.

Keywords: Hume, justice, property, fictionalism, convention, contractarianism.

1. Introduction

In the *Treatise* (Hume 1978; in the following cited as T followed by page numbers) and the *Enquiry concerning the Principles of Morals* (Hume 1983; cited as E followed by page numbers)[1] Hume discusses the morality of justice by using a revolutionary method that displays the foundation of justice in social utility and the progression of mankind. Hume dismisses a rationalist explanation of justice by relying upon his thesis, defended in section I, part I, book III of the *Treatise* that "moral distinctions are not derived from reason". He also discards an explanation based on natural self-interest. His words in this respect leave no possible doubt: "[I]t is 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." (T, 480) If the source of justice is neither reason nor natural self-interest, what is it? To answer this question, Hume formulates a theory of justice that can be called

[*] I am grateful to Bernd Lahno and Maura Priest for comments on an earlier version.

[1] Spelling in the Hume texts has been modernized.

conventionalist in a rather unorthodox sense of 'convention', which he defines in the third appendix to the second *Enquiry*:

> "It has been asserted by some, that justice arises from *human conventions*, and proceeds from the voluntary choice, 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." (E, 95)

In the quoted paragraph Hume summarizes his theory about how the classic system of justice—i.e., the conventions of property or constant possession, its transference by voluntary agreement and the undertaking of promissory obligations—has arisen gradually as the unintended outcome of individual decisions guided by a sense of common interest. It is worth quoting a commentator in this respect: "To see justice in this way, as an unintended consequence of individual human actions, must be one of the boldest moves in the history of the philosophy of law." (Haakonssen 1981, 20)

Though the consistent reconstruction of Hume's theory is far from simple, it might be thought that little of value remains to be said in light of the abundance of classic and contemporary commentaries. I believe, however, that the complexities of Hume's thought leave room for further discussion. Hume constructs his theory in three discernible stages. First, Hume needs to display the social artificiality of the virtue of justice and the mechanisms through which this artificiality is generally masked. Second, Hume seeks to provide a naturalistic, nonpolitical explanation of the patterns of behavior that are required by the conventions of justice. Third, Hume wants to explain why we *call* those patterns of behavior morally virtuous and why we *judge* them as morally praiseworthy. The difference between the second and third stages is that, in the former, Hume focuses on agents' behaviors, whereas in the latter he is concerned with third-person moral judgment.

In this paper I will concentrate on the three stages of Hume's explanation of justice and I will largely leave aside his theory of government and political allegiance. In *section 2*, I discuss the famous 'is-ought' passage. In *section 3*, I turn my attention to Hume's argument for the artificiality of the virtue of justice. In *section 4*, I discuss his view about the spontaneous and progressive emergence of justice. In *section 5*, I address Hume's account of the consolidation of justice as a result of annexing the idea of moral praiseworthiness to the originally self-interested conventions of justice. In the last section I show the dissimilarities between Hume's theory, on the one hand, and contractarianism and indirect utilitarianism, on the other.

2.

Hume says that every moral theory he knows makes a surprising deduction from propositions connected with 'is' and 'is not' to conclusions connected with 'ought' and 'ought not'. "This change"—he goes on—"is imperceptible; but is, however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, "it is necessary that it should be observed and explained; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it." (T, 469) It has been common to think—perhaps as a consequence of G. E. Moore's account of the 'naturalistic fallacy'—that what Hume really claims is that the derivation of ought-statements from merely descriptive statements is fallacious. But it is plain that Hume does not share at all Moore's thesis that naturalism commits any 'fallacy', because he himself provides a naturalistic account of moral notions. For instance, Hume adopts a naturalistic analysis of 'natural obligation'. Thus, for Hume the proposition 'We lie under an obligation to do X' means something like 'The non-performance of X is vicious'. As Frankena (1939) has long ago shown, it would beg the question to claim that a naturalistic account is fallacious just because one assumes that naturalism is false.

Curiously, in a classic essay John Searle (1964) sought to show that 'institutional ought's' can be derived from 'institutional is's', but believed that this was a way of refuting Hume's 'is-ought' doctrine. Searle claimed that 'ought' can be deduced from 'is' within the context of institutionalized forms of obligation. In this way he intended "to demonstrate a counterexample" to Hume's thesis that "no set of statements of fact by themselves entails any statement of value". But Hume just says that the deduction *seems* inconceivable, not that it *is* inconceivable. How could he otherwise deduce conclusions establishing obligations and rights from factual observations concerning the basic circumstances of human society?

Commentators acknowledge today that Hume does not take the derivation of 'ought' from 'is' to be a fallacy. Therefore, the 'is-ought' passage should be read as an introduction to the discussion of the rules of justice. Thus, Alasdair MacIntyre (1969) says that Hume does not regard the inference as logically impossible. Hume requires that the inference "should be observed and explained". In fact, Hume cites psychological and sociological remarks as premises for his explanation of justice. Selfishness and confined generosity, on the one hand, and the scarcity and instability of possession of external goods, on the other, are the fundamental facts that explain why we have a basic common, shared interest in establishing the conventions of justice. These conventions are the stability of possession, its disposition by consent and the obligation to keep promises. MacIntyre thinks that Hume justifies the obligations of justice on the basis of those facts, thus engaging in the deduction he has emphatically denunciated. In support of his interpretation, MacIntyre quotes the following passage: "And even every individual person must find himself a gainer, on balancing the account; since, without justice, society must immediately dissolve, and everyone must fall into that savage and solitary condition, which is infinitely worse than the worst situation that can possibly be supposed in society." (T 497) Despite the quotation, I agree with Murray MacBeth (1992) that Hume's primary goal is not to offer a justification of the conventions of justice, but rather an explanation of their emergence and consolidation. He also seeks to *explain* the moral approbation of justice as a social fact, as something different from justifying this approbation. However, if Hume's explanation of the establishment of the convention of justice is sound, it must be true that the participants in the convention reason from factual premises to the normative conclusion that they ought to approve of the rules of justice. For the moral approval of the rules of justice to be consilient with Hume's explanatory account, these reasonings have to be premised on the same factual claims that he includes in the *explanans* of the rules of justice.

Annette Baier defends a position similar to MacIntyre's. She points out that, in the famous paragraph, Hume purports to attack the deficiencies of rationalist moral theories rather than to make a general impossibility claim. Baier says: "He [Hume] observes and explains the transition from facts about importance to claims about norms, from facts about conventions to conclusions about rights, and makes them reasonable." (Baier 1991, 177) In overlooking the explanation/justification distinction, Baier makes herself also liable to MacBeth's critique. But the rejoinder I offered previously is here also available.

My reading is in the MacIntyre/Baier tradition, but I emphasize the *internalist* sense of the paragraph. Hume's moral theory is internalist in that he understands moral notions as inherently motivational. On this view, moral propositions must have an adequate connection to internal motivational acts or states. What is baffling to Hume, I think, is how 'external' facts, that is, facts unrelated to our passions and motives, can create new moral obligations. Hume has no objection to the proposition that our natural passions and sentiments can warrant *natural obligations*, but he is reluctant to accept that an 'external' fact can create, in and of itself, an obligation. It is certainly puzzling to regard someone's promising or occupying a tract of land as obligation-generating facts. Therefore, I take Hume's 'is-ought problem' to be really that of explaining how claims on 'external' facts can entail new obligations. Indeed, what requires an explanation is not the inference of 'ought' from *any* 'is'. What demands an explanation is the derivation of ought-judgments from statements that denote 'external' facts, that is, facts unrelated to our natural inclinations.

The best interpretation of the 'is-ought' passage does not view it as the denunciation of a fallacy, but as the preamble to an explanation of *artificial obligations*. Artificial 'ought's' can be deduced from certain 'is's', even if the inference is not premised on the meaning of 'obligation'. But a fallacy is not necessarily committed. It can be avoided by just providing an explanation of justice, as Hume precisely does. Conversely, if the explanation is wanting or fails, "a very gross fallacy" is certainly committed. Hume says: "Those, therefore, who make use of the word property, or right, or *obligation*, before they have explained the origin of justice, or even make use of it in that explication, are guilty of a *very gross fallacy*, and can never reason upon any solid foundation." (T, 491; italics added) The fallacy is not just to deduce 'ought' from 'is', but rather to do so without a prior explanation about the origin of justice. When such explanation is given, the derivation is possible, and no fallacy is involved. So it seems that the explanation of 'artificial ought's', which Hume demands in the 'is-ought passage', is really the explanation of the origin of justice. Once this explanation is provided, 'ought' can be deduced from 'is'. The "gross fallacy" is to try to infer prescriptions about justice from 'externalist' factual claims in the absence of an explanation of the origin of justice.

3.

How does Hume prove that justice obligations are artificial rather than natural? He avails himself of a logical method for disclosing the artificiality of justice. He assumes that nature by necessity accords with logical principles, and that, therefore, the descriptions of natural phenomena cannot lead to inconsistencies or fallacies. If nature does not cause fallacies, the discovery of hidden fallacies in our reasonings is a mark of our own intellectual frailty, that is, of artificiality. Hume dissects our ideas of justice and shows that the rules of justice are contrived, invented and projected onto the world, rather than discovered. Hume's assumption is that human societies generate superstitions that present themselves as if they were realities. But superstitions are imperfect and, therefore, empirical observation and philosophical argument can discover their contrived character.

In the *Treatise* Hume's view about the merit and demerit of virtuous actions is a *subjectivist* one. The moral quality of benevolence does not lie in the external behavior but in the underlying motive. Hume asserts that "when we praise any actions, we regard only the motives that produced them, and consider the actions as signs or indications of certain principles in the mind and temper" (T, 477). He adds: "The external performance has no merit." (T, 477) A few pages later, he asks: "*Wherein consists this honesty and justice, which you find in restoring a loan, and abstaining from the property of others?*" (T, 480) And he answers in this way: "It does not surely lie in the external action. It must, therefore, be placed in the motive, from which the external action is derived. This motive can never be a regard to the honesty of the action. For it is a plain fallacy to say, that a virtuous motive is requisite to render an action honest, and at the same time that a regard to the honesty is the motive of the action. We can never have a regard to the virtue of an action, unless the action be antecedently virtuous." (T,

480) From this he concludes that “we have naturally no real or universal motive for observing the laws of equity, but the very equity and merit of that observance; and as no action can be equitable or meritorious, where it cannot arise from some separate motive, there is here an evident sophistry and reasoning in a circle. Unless, therefore, we will allow, that nature has established a sophistry, and rendered it necessary and unavoidable, we must allow, that the sense of justice and injustice is not derived from nature, but arises artificially, though necessarily from education, and human conventions.” (T, 483) Hume detects a circular reasoning embedded in the language of justice and regards it as a “mark of artifice and contrivance” (T, 528).

Hume needs a further premise to substantiate his assumption that the sense of justice cannot count as a natural motive. For Hume natural sentiments have the following two marks. First, natural moral sentiments have a reference to sympathetic engagement and, therefore, are directed toward someone else’s affections or sentiments, or to mankind or society as a whole. For instance, private benevolence is a natural virtue because it is directed to someone else’s suffering or needs, and public benevolence is a virtue because it is directed to mankind or society in general. Hume rejects the idea that the sense of justice can be a regard to either private or public interest (T, 480–2). Justice cannot be equated with public or private benevolence because neither the public nor the private interest is naturally or necessarily attached to all particular acts of justice. For instance, the sense of justice requires actions that often fail to be a form of sympathy with the person to whom we owe the just action, as when he is “a vicious man”, who “deserves the hatred of all mankind” (T, 482). Second, natural moral sentiments are *intrinsically good*. For instance, generosity is always morally praiseworthy and its moral praiseworthiness does not depend on the maintenance of general patterns of behavior. Hume says very clearly that the “social virtues of humanity and benevolence exert their influence immediately, by a direct tendency or instinct”, and he goes on to say that “the good, resulting from their benign influence, is in itself complete and entire”, and that it “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 of society” (E, 93).

In acute fashion Hume claims that, unlike the natural virtues, the social virtues of justice and fidelity are only *socially good*. He says: “They are highly useful, or indeed absolutely necessary to the well-being of mankind: But the benefit, resulting from them, is not the consequence 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.” (E, 94) This instrumental account of justice implies that when social convergence is absent, justice is useless and it is pointless to follow its constraints. The example of the society of ruffians seems like an application of the theory of the second best: “Suppose likewise, that it should be a virtuous man’s fate to fall into the society of ruffians, remote from the protection of laws and government; what conduct must he embrace in that melancholy situation?” (E, 23) Hume’s answer is consonant with his view that justice is not intrinsically good: "[H]is particular regard to justice being no longer of *use* to his own safety or that of others, he must consult the dictates of self-preservation alone [. . .]." (E, 23)

Hume's circularity argument is amenable to different interpretations. To start with, it is perplexing to hold that honest acts, such as restoring a lent sum of money (T, 479), are virtuous (though artificially so) only if they derive from a separate motive, and at the same time to deny that honest acts can have a separate motive. There are at least three possible ways out of this difficulty. The first way, suggested by Mackie (1980, 79f.), is that Hume relaxes the requirement of an independent motive for artificial virtues. Honest acts can be made out a regard to justice alone, even if a regard to justice is simply defined as a disposition to perform honest acts. Mackie suggests a reformulation of Hume's premise along this line: Actions count as naturally virtuous only in so far as they are signs of virtuous motives; actions count as artificially virtuous if they are made out of regard to a general scheme that promotes the public good.

An alternative interpretation finds textual support in the section of the *Treatise* in which Hume discusses the obligations created by promises (Section V, Second Part, Book III). Hume reproduces his general claim that there is no natural motive that causes a natural obligation to fulfil promises: "Now it is evident we have no motive leading us to the performance of promises, distinct from a sense of duty. If we thought, that promises had no moral obligation, we never should feel any inclination to observe them." (T, 518) Again, the problem is that the sense of duty does not generate an obligation to perform promises, but rather is a consequence of this obligation: "[A] sense of duty supposes an antecedent obligation." (T, 518) In order to accommodate Hume's axiomatic claim about the value of all meritorious actions, the artificiality of the obligation to do as promised is concealed under a sophistic metaphysics of the mind. Thus, Hume thinks that promising involves feigning a natural motive because "every new promise imposes a new obligation of morality on the person who promises" (T, 524). He says that this new obligation is thought to arise from the promisor's will: "Here, therefore, we *feign* a new act of the mind, which we call the *willing* an obligation; and on this we suppose the morality to depend." (T, 523) This imagined act of the mind is utterly fictional: "[I]t is one of the most mysterious and incomprehensible operations that can possibly be imagined, and may even be compared to *transubstantiation*, or *holy orders*, where a certain forms of words, along with a certain intention, changes entirely the nature of an external object, and even of a human creature." (T, 524)[2]

[2] Hume discusses the fictionalization in the discourse of justice even before dealing with the obligation of performing promises, in the section on the transference of property by consent. Rather than as an external object, Hume conceives of property as social construct: "The property of an object, when taken for something real, without any reference to morality, or the sentiments of the mind, is a quality perfectly insensible, and even inconceivable; nor can we form any distinct notion, either of its stability or translation." (T, 515) But it is the translation of property that requires still a greater deal of imagination. Hume says that, "in order to aid the imagination", we require the physical delivery of the sold object to the new owner in order to consider the transfer-

Mackie (1980, 103) stresses the fictional component of this explanation in holding that "promising involves the *fiction* that we create an obligation by a special act of the mind, the willing of that very obligation". This fictional interpretation makes sense if one does not relax the principle that all meritorious actions are signs of underlying motives that are constitutive of their essence. That is, if all meritorious actions are signs of prior motives, and the observance of promises is meritorious, we must postulate the existence of a motive that is prior to and independent of the existence of the obligation. Now, since the sense of duty does not fit into the definition of a natural motive, because it presupposes the existence of the relevant duty, it is necessary to feign the existence of a motive.

Knud Haakonssen adopts a similar reading. He claims that the psychological mechanism Hume envisages is a two-step one. First, "the natural tendency in men to see behaviour as an expression of motives, and motives as expressions of qualities of character, leads them to *imagine* that there is a natural motive (and thus a character trait), namely the willing of an obligation, behind promises [. . .]." (Haakonsssen 1978, 13). Second, agents morally approve of this feigned motive in oneself or others, and disapprove of its lacking in general or on particular occasions. When it is oneself who lacks the motive, self-hatred is addressed to this (imagined) "defect and imperfection in the mind and temper" (T, 518). This self-hatred has a motivational force but is nonetheless an irrational sentiment in that its factual assumption (i.e., the existence of a natural motive) is false. It is important to notice that the actuating force of promises derives on this account from a natural sentiment too. It is the natural sentiment to hate oneself for having a bad character. Compassion, for instance, is backed by this self-hatred: "Though there was no obligation to relieve the miserable, our humanity would lead us to it; and when we omit that duty, the immorality of the omission arises from its being a proof, that we want the natural sentiments of humanity." (T, 518)

I propose a third line of interpretation that maintains that the motive behind promises is a real one, though its appearance as a natural motive is feigned. Promisors who have a sense of duty really feel motivated to perform the promise, even when they have never entered into the state of self-hatred because the prior existence of the motive was not questioned. What promisors imagine is not a motive, which they already have in the form a sense of duty, but rather its naturalness. Feigned motives are not motives at all because they lack motivational force. But motives can be represented through the false image of a natural motive. The sense of duty is a motive artificially implanted in persons. It behaves as any other motive, namely, leading agents to act in the relevant ways. So it is not feigned motives what gives promises their motivational force, but rather real motives that pass off as genuine natural motives. Acts are naturally meritorious or obligatory if they stem from passions such as self-love, private beneficence or public interest. We do not encounter any of these natural passions in the case of

ence of property effectively perfected, and, when the physical delivery is not possible, "men have invented a *symbolical* delivery, to satisfy the fancy" (T, 515; italics in the original).

just actions. Therefore, the sense of duty must be implanted in agents by conventions, but once implanted it is a real motive, any confusion about its origin notwithstanding.

4.

Hume claims that the rules of justice are social inventions. But what kind of inventions? Inventions often are deliberately designed artifacts to serve various purposes. For instance, bridges, harbors, ramparts, canals, fleets and armies are deliberately invented (T, 539). It would be a serious mistake to think that the rules of justice have been invented in the same way (let alone engineered or planned). Like other authors in the Scottish enlightenment (e.g., Adam Ferguson), Hume maintains that the fundamental institutions of mankind are spontaneous and unintended consequences of people's self-interested actions. Hume means that they are evolved rather than designed inventions:

> "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 convention and agreement. 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." (E, 95)

Hume claims that the conventions of justice have arisen as a byproduct of collective patterns of behavior originally motivated by self-interest and self-love: "This system (of justice), therefore, comprehending the interest of each individual, is of course advantageous to the public; though it be not intended for that purpose by the inventors." (T, 529) However, as indicated at the beginning of this essay, it is not natural self-love or self-interest, but self-love coupled with judgment and understanding that is the original source of the conventions of justice: "The remedy, then, is not derived 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." (T, 489; latter italics added)

How does Hume conceive of the gradual evolution of the institution of justice? This is by no means an easy question (cf. Kliemt 1986, 74). The key to understanding Hume's thought is that he viewed society as a mutually advantageous invention. In this he echoes Hobbes's idea that civil society is a way to escape from the dark state of nature. The infirmities of mankind foredooms the human species to tragedy without an invention that makes up for mankind's infirmities: "Of all the animals, with which this globe is peopled, there is none towards whom nature seems, at first sight, to have exercised more cruelty than towards man, in the numberless wants and necessities, with which she has loaded him, and in the slender means, which she affords to the relieving these necessities." (T, 484)

He goes on to say that "society provides a remedy" in a three-fold way: "By the conjunction of forces, our power is augmented. By the partition of employments, our ability increases. And by mutual succor we are less exposed to fortune and accidents." (T, 485) Now, the main challenge to the maintenance of society, which originally grows from the 'natural appetites betwixt the sexes' and the new affection to company and conversation cultivated by early society, is the "insatiable, perpetual, universal and directly destructive of society avidity of acquiring goods and possessions for ourselves and our nearest friends" (T, 491f.). This destructive sentiment would prevent people from the peaceful "enjoyment of such possessions as we have acquired by our industry and good fortune" (T, 487) if all the members of society did not enter into a convention "to bestow stability on the possession of those external goods" (T, 489). Hume stresses that "the convention for the distinction of property, and for the stability of possession, is of all circumstances the most necessary to the establishment of human society, and that after the agreement for the fixing and observing of this rule, there remains little or nothing to be done towards settling a perfect harmony and concord" (T, 491). Therefore, the convention for the stability of possession is a sine qua non condition of the existence of society, which, in turn, is indispensable for rescuing mankind from the bleak prospects of a solitary and violent existence.

As the paragraph quoted at the beginning of this paper shows, Hume rejects the idea that the origin of justice rests on a promise or contract. The convention founding justice arises from "a general sense of common interest" (T, 490). Hume explains the operation of the convention in this way: "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 expressed, and is known to both, it produces a suitable resolution and behaviour. And this may properly enough be called a convention or agreement betwixt us [. . .]." (T, 490) In the *Treatise* Hume gives two well-known examples to illustrate the emergence of the rules of justice without an explicit convention.[3] The first example, already quoted, is the spontaneous cooperation between two rowers:

> "Two men, who pull the oars of a boat, do it by an agreement or convention, though they have never given promises to each other. Nor is the rule governing the stability of possession the less derived 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, 490; italics added)

[3] See Hardin 2007, 84–5, for a detailed list of coordination examples in Hume's works.

The second example is given in the section on the origin of government: "Two neighbours may agree to drain a meadow, which they possess in common; because it is easy for them to know each other's mind; and each must perceive, that the immediate consequence of his failing in his part, is, the abandoning the whole project." (T, 538)

The standard way of modeling the two examples is by means of a matrix that represents a perfect convergence of interests (*figure 1*). Both agents have a single individually and collectively advantageous strategy. Though Hume thinks that situation can be called 'properly enough' a 'convention' or 'agreement', they are hardly conventions or agreements in a proper sense, because both agents will predictably pursue their (weakly) dominant strategy. To be sure, since the game has two equilibria, the question of equilibrium selection can be formally raised. But coordination seems all too natural in this situation because individual failing guarantees the minimal payoff 0 while choosing Row/Drain offers a chance of one unit with no risk of any loss. So there is no coordination *problem* in these examples (Hardin 2007, 78–9). Though a modicum of understanding is certainly necessary to prevent failing, norms or conventions of coordination seem unnecessary (see also Ullman-Margalit 1977, 79f.). Under this analysis, it can be contested whether Hume's examples should be understood as exemplifying a proper theoretical model of the conventions of justice. Hume should be interpreted, instead, as giving just two examples of how a sort of agreement can be reached without the need for explicit consent. On this interpretation, the analogy between the emergence of stable possession and the cooperative patterns in the two examples would only run as long as one focuses on the common implicit character of otherwise different kinds of social cooperation.

<table>
<tr><td></td><td></td><td colspan="2">Player 2</td></tr>
<tr><td></td><td></td><td>Row/Drain</td><td>Fail</td></tr>
<tr><td>Player 1</td><td>Row/Drain</td><td>1 | 1</td><td>0 | 0</td></tr>
<tr><td></td><td>Fail</td><td>0 | 0</td><td>0 | 0</td></tr>
</table>

Figure 1: Convergence of interests

A preferable model for representing the essential properties of the convention of justice as characterized by Hume may well be the Stag Hunt Game (Skyrms 2004). This is a coordination game that also includes two equilibria. And, again, one equilibrium, in which both men row or drain, is the most advantageous, while the other, in which both opt for playing alone, is less advantageous. But now opting for the less advantageous equilibrium is safer because its success does not depend on the other partner (*figure 2*).[4] Thus, there is no guarantee that the players will follow their best (joint) strategy.

		Player 2	
		Row/Drain	Fail
Player 1	Row/Drain	2 | 2	0 | 1
	Fail	1 | 0	1 | 1

Figure 2: Stag Hunt game

The game is inspired in a parable of Rousseau in *A Discourse on the Origin of Inequality* (1993, 86f.): "If a deer was to be taken, every one 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." Rousseau's story is useful to highlight Hume's theory of justice (though perhaps the application is infelicitous given Rousseau's rancor for Hume; see Zaretsky and Scott 2009). The story makes it clear that partners can defect joint endeavors that are *both individually and collectively* advantageous out of risk-aversion, miscalculation of probabilities, high discount rate, or plain irrationality. In like manner, the stability of possessions and the other rules of justice can be threatened by defective behaviors even if the social union that such conventions allow has immense advantages in terms of conjunction of forces, division of labor, mutual succor and other forms of social cooperation. Hume thinks that, without the conventions of justice, "society must immediately dissolve, and every one must fall into that savage and solitary condition, which is infinitely worse than the worst situation that can possibly be supposed in society" (T, 497).

The tragedy of social dissolution can be modeled as the undesirable equilibrium in a stag-hunt game. But it could also be modeled as the undesirable equilibrium in a chicken game, or as the Hobbesian equilibrium in a prisoners' dilemma game. I think it is misguided to attribute to Hume one single model. In fact, the relevant texts in the *Treatise* can support various interpretations, and some of them are incompatible with the model of perfect convergence suggested by his two famous examples. Thus, some of Hume's paragraphs evoke a coordination game with multiple equilibria, or even a prisoners' dilemma game. For instance, Hume says that when the stability of possession is developing: "I observe, that it will be for my interest to leave another in the possession of his

[4] See Lewis 1989, 7; Ullman-Margalit 1977, 121–4. In *figure 2* I follow Ullman-Margalit's modeling of the game.

goods, *provided* he will act in the same manner with regard to me." (T, 490; italics in the original) This suggests that the cooperative player suffers a loss if the other player defects.

In any case, which model we choose is unimportant, because Hume had in mind a diversity of social situations that possess one common trait: a joint strategy is available that is in the best interest of all the participants. This can happen in interactions that can be modeled in various ways. The important point is that Hume regards the emergence of justice as the unintended outcome of an intelligent practice of understanding and judgment focused on a general plan of actions. The endorsement of this system carries with it the progressive emergence of a sense of justice that is different from natural self-interest. The perception of the potential gains arising from the conventions of justice is not immediate, and justice-generating patterns of behavior are strengthened by a "slow progression" and by "our repeated experience of the inconveniences of transgressing it". It is evident that standard game-theoretical tools are too simplistic to model this process because they do not consider the feedback mechanisms that operate when participants' motivational and cognitive systems change over time in response to new environments that are produced by the unexpected consequences of their prior decisions. "Justice establishes itself by a kind of convention or agreement" (T, 498) when men "have had experience enough to observe, that whatever may be the consequence of any single act of justice, performed by a single person, yet the whole system of actions, concurred in by the whole society, is infinitely advantageous to the whole, and to every part" (T, 497–8). Learning by experience, and not merely static self-interest, is the driving force of the evolution to a just society. Hume's elegant prose is worth quoting: "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." (T, 26)

5.

In a 'narrow and contracted society' a model of coordination could explain the emergence of stable rules of possession on the assumption that the parties are endowed with sound judgment and understanding. Yet, when society "becomes numerous, and increases to a tribe or nation", the interest in the preservation of justice "becomes remote" (T, 499). In these conditions, men do not "readily perceive, that disorder and confusion follow upon every breach of these rules, as in a more narrow and contracted society" (T, 499). In a large society, the currency of the conventions of justice requires that "we annex the idea of virtue to justice, and of vice to injustice" (T, 498). In other words, the *natural obligation* to justice must be transformed into a *moral obligation*. This can be done by means of *moral language* and *moral criticism* addressed to other people's just and unjust actions, regardless of whether such actions are close to or distant from us (both geographically or historically). Hume thinks that we naturally praise human behavior because of its utility: “It appears to be matter of fact, that the circumstance of *utility*, in all subjects, is a source of praise and approbation: That it is constantly appealed to in all moral decisions concerning the merit and demerit of actions [. . .].” (E, 50) “The esteem for justice as useful and indeed necessary to the support of society relies on a natural sentiment of agreeableness that serves to turn justice into a moral obligation, and, therefore, to give it a stronger foundation.” (E, 34)

How can moral judgment operate more effectively in our actions than the virtue of justice itself? The reason Hume gives is that there is an asymmetry in our attitudes toward our behavior and toward other agents’ behaviors. Basically, we see more easily the mote in somebody else’s eye than the beam in one’s own. Hume makes the point very clearly: “But though in our own actions we may frequently lose sight of that interest, which we have in maintaining order, and may follow a lesser and more present interest, we never fail to observe the prejudice we receive, either mediately or immediately, from the injustice of others; as not being in that case either blinded by passion, or bypassed by any contrary temptation.” (T, 499)

Hume states his theory of moral criticism in this way: “*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, 499f.) Whereas self-interest is the original motive to the establishment of justice, the natural sentiments of regard for public usefulness and sympathy for other people ground the moral approbation of justice. In turn, the moral approbation of justice generates the moral obligation to justice. When we contemplate injustice done by others, we understand immediately the prejudice done to the person who suffered the injustice and to human society in general. Our sympathy with public utility makes us condemn the injustice done by the actions of others and, progressively, also the injustice we make when we “sympathize with others in the sentiments they entertain of us”. Conversely, we morally approve of just actions because of their social utility.

The operation of natural sympathy is strengthened by the artifices of politicians, who “endeavor to produce an esteem for justice, and an abhorrence of injustice”, “in order to govern men more easily and preserve peace in human society” (T, 500). Though sympathy is a natural moral sentiment, the artifice of politicians assists nature. The artifice is conceptual. Politicians coin the words ‘honorable’ or ‘dishonorable’, ‘praiseworthy’ or ‘blamable’ in order to praise justice and blame injustice. Though these concepts are artificial, their motivational effectiveness is nurtured by our natural inclination to esteem public utility. What politicians do is to “extend the natural sentiments beyond their original bounds” (T, 500).

Hume underlines the role of private education and instruction. The inculcation of the principles of probity and of sentiments of honor by parents are critical for shaping the character of children in favor of justice. Educators teach children "to regard the observance of those rules, by which society is maintained, as worthy and honourable, and their violation as base and infamous" (T, 500f.).

The picture of how justice acquires solidity in numerous societies is only complete at the end of the section when Hume refers to the interest in our reputation: "There is nothing, which touches us more nearly than our reputation, and nothing on which our reputation more depends than our conduct, with relation to the property of others." (T, 501) In order to preserve his reputation a man "must fix an inviolable law to himself, never, by any temptation, to be induced to violate those principles, which are essential to a man of probity and honour" (T, 501).

6.

Hume's conventionalist theory of justice defies classification under contemporary headings. It is not fully contractarian, because the notions of voluntary choice, consent, and promise play no role in his theory. By the same token, the concept of an actual or hypothetical contract, which is the contemporary rendering of those notions, is also alien to Hume's thought. Contractarianism involves the (1) deliberate, (2) unanimous and (3) mutually advantageous (actual or hypothetical) choice of legal and political principles in a given baseline. Hume is at pains to clarify that the principles of justice are not the outcome of a deliberate convention (governed by unanimous or less-than-unanimous decision-making rules). For Hume the principles of justice do not derive from a single, synchronic, collective decision-making process. On the contrary, the system of actions that Hume calls the convention of justice arises as the byproduct of a gradual and progressive series of individual choices aided by understanding, judgment, and learning by experience. The first two defining features of contractarian choice (deliberate and unanimous choice) are then absent in Hume's theory because he rejects the idea that justice emerges from a contract, that is, from deliberate and unanimous choice. However, Hume does think that justice is a mutually advantageous institution (the third feature). In this respect, but only in this respect, Hume's theory may certainly be called contractarian. But perhaps we had better avoid this usage, given its potentially misleading implications. Hume's theory is conventional in the sense he properly clarified. This sense has only a partial overlap with contractarianism.

It might be also misleading to regard Hume's theory as a kind of indirect utilitarianism. Indirect utilitarianism purports to offer us a public policy program for redesigning our rules and conventions of justice to maximize social welfare. Such approach is alien to Hume's view. Hume thought that the rules of justice carry the accumulated practical knowledge of our civilization. A sense of practicality, rather than an idea of perfection or optimality, drives the evolution of justice. In fact, Hume alerts us against ideal rules of justice (e.g., perfect equality) that are impracticable and that would "reduce society to the most extreme indigence" (E, 28). Gradual and progressive evolution, judgment and understanding, learning by experience, and the transformation of people's cognitive and motivational dispositions are the hallmarks of the Humean theory of justice. These hallmarks stand in opposition to the assumptions of indirect utilitarianism—static preferences, possibility of one-dimensional maximization and institutional engineering.

Whereas indirect utilitarianism presents itself as a public blueprint for social reform, Hume offers his theory in order to explain a spontaneous process of evolution to a cooperative and mutually advantageous society. Society is too complex for proposing all-encompassing blueprints. Hume would have rejected proposals that sought to maximize one single collective goal. History teaches us that free human beings pursue various goals and that this constellation of various goals is the true basis of civilization. On Hume's view justice is an evolved conventional system that allows human beings to carry on their lives in a prosperous and peaceful society. No more, no less.

References

Baier, A.C. (1991), *A Progress of Sentiments, Reflections on Hume's Treatise*, Cambridge/MA: Harvard University Press.

Frankena, W. K. (1939), "The Naturalistic Fallacy", *Mind* 48, 464–477.

Haakonssen, K. (1978), "Hume's Obligations", *Hume Studies* IV(1), 7–17.

— (1981), *The Science of a Legislator, The Natural Jurisprudence of David Hume and Adam Smith*, Cambridge: Cambridge University Press.

Hardin, R. (2007), *David Hume: Moral and Political Theorist*, Oxford: Oxford University Press.

Hume, D. (1978), *A Treatise of Human Nature*, 2nd edition, Oxford: Clarendon Press, cited as T.

— (1983), *An Enquiry Concerning the Principles of Morals*, Indianapolis: Hackett, cited as E.

Kliemt, H. (1986), *Las instituciones morales*, Barcelona–Caracas: Alfa.

MacBeth, M. (1992), "'Is' and 'Ought' in Context: MacIntyre's Mistake", *Hume Studies* XVIII(1), 41–50.

MacIntyre, A. (1969), "Hume on 'Is' and 'Ought'", *The Philosophical Review* 68, 451–68.

Mackie, J. L. (1980), *Hume's Moral Theory*, London: Routledge.

Roussseau, J.-J. (1993) *The Social Contract and Discourses*, London: Everyman.

Searle, J. R. (1964) "How to Derive 'Ought' from 'Is'", *The Philosophical Review* 73, 43–58.

Skyrms, B. (2004), *The Stag Hunt and the Evolution of Social Structure*, Cambridge: Cambridge University Press.

Ullman-Margalit, E. (1977), *The Emergence of Norms*, Oxford: Clarendon Press.

Zaretsky, R. and J. T. Scott (2009), *The Philosophers' Quarrel, Rousseau, Hume, and the Limits of Human Understanding*, New Haven: Yale University Press.