Australasian Journal of Philosophy
Vol. 72, No. 4; December 1994

HOBBES, SPINOZA, KANT, HIGHWAY ROBBERY AND GAME THEORY

Louis Marinoff

I. The Problem

You are abducted by a highway robber, who seeks to hold you for ransom. By arguing continuously for several hours, you convince the robber that you are a philosopher. The robber becomes understandably upset, as the realization dawns of the unlikelihood that anyone will ransom you at any price. In fear for your life — whose worth your captor now seriously doubts — you promise to deliver the ransom yourself, at an appointed place and time, in exchange for your release. To your relief, your abductor is a constrained maximizer (Gauthier [5]), who reasons empirically that it would cost him less effort to release you than to feed you or do away with you; and who reasons dubiously that you, having some concept of ethics, are as likely to keep your promise as to break it. So he releases you. Should you subsequently keep your promise and deliver the ransom?

II. Four Solutions

(i) Hobbes' Solution

An unembellished version of this problem is posed and solved by Hobbes, in the heart of his *Leviathan*. Hobbes, of course, is concerned with delivering us from our wretched state of nature, succinctly described in chapter 13 as 'a war, as is of every man, against every man'. Such a condition, Hobbes asserts, is one of 'continual fear, and danger of violent death; and the life of man, solitary, poor, nasty, brutish, and short' [6, p.82].

The Hobbesian state of nature is neither moral nor lawful:

> The notions of right and wrong, justice and injustice have there no place . . . Force, and fraud, are in war the two cardinal virtues[1] . . . It is consequent also to the same condition, that there be no propriety, no dominion, no *mine* and *thine* distinct; but only that to be every man's, that he can get: and for so long, as he can keep it. [6, p.83]

We understand from this argument that Hobbes finds no moral or legal fault with highway robbery in a state of nature. In such a state, anyone who ventures abroad (which formerly meant merely setting one's foot out-of-doors) should fully expect to be accosted, assaulted, waylaid, kidnapped, robbed, murdered, or otherwise left for dead. The highway robber enjoys a perfect right to relieve us of our valuables and even of our lives, since his claim to these things — as well as ours — is predicated solely upon the power of possessing them.

[1] Notwithstanding his fierce polemics against Aristotle and Scolasticism, Hobbes seems to have 'borrowed' this latter principle from Aristotle [1].

It might seem to follow from this argument that, if the robber has a perfect right to abduct us by force, then we have as perfect a right to secure our release by fraud. But according to Hobbes, this is not necessarily so, not even in a state of nature. Inasmuch as Hobbes' contractarian philosophy aims at leading us out of the infamous state of nature and into an internally peaceful commonwealth, replete with systems of ethical and judicial process, Hobbes must restrict the universal acceptability of force and fraud alike.

He accomplishes this by first introducing a fundamental right of nature, which is a statement of the principle of self-preservation:

> The right of nature . . . is the liberty each man hath, to use his own power, as he will himself, for the preservation of his own nature; that is to say, of his own life; and consequently, of doing any thing, which in his own judgement, and reason, he shall conceive to be the aptest means thereunto. [6, p.84]

On the one hand, this principle can neither be undermined nor overshadowed by any of Hobbes' subsequently-introduced 'laws of nature', mere normative prescriptions which, if followed, conduce to the establishment of a commonwealth. In fact, the right of nature will serve to mitigate and even abrogate these so-called 'laws', whenever their application appears to infringe upon it. On the other hand, this principle knows no upper bound, nor any condition of constraint. Thus it can be stretched to accommodate the illimitable reaches of human desire.

Hobbesian contractarianism is founded upon the natural right of self-preservation. People band together into groups, the better to protect themselves. In so banding, they form tacit if not overt agreements, which most generally consist either in renunciation or in transfer of right.[2] Both the renunciation and the transfer of right, Hobbes maintains, are voluntary acts, in which the actor's object 'is some good to himself':

> And lastly the motive, the end for which this renouncing, and transferring of right is introduced, is nothing else but the security of a man's person, in his life, and in the means of so preserving life, as not to be weary of it. [6, p.87]

And what specific right is to be renounced or transferred? It is the natural right of all men to all things, which underlies and perpetuates the state of nature. The renunciation or transfer of natural right is none other than the instantaneous recognition of a 'mine' and 'thine' distinct. The mutual transference of right is called, by Hobbes [6, p.87], a 'contract'. Contracts are necessary to but insufficient for the maintenance of social order, in that an actor's decision to recognize another's right to something in an earlier instant can easily be superseded by a decision not to recognize it in some later instant, particularly if a new circumstance arises in which the actor's security (or felicity) is heightened expressly by de-recognition.

Thus Hobbes is obliged to introduce the notion of a pact, or covenant, by which 'both parts may contract now, to perform hereafter' [6, p.87]. A covenant is therefore a con-

[2] Hobbesian renunciation is an act wherein the actor 'cares not to whom the benefit thereof redoundeth'; Hobbesian transfer is an act wherein the actor 'intendeth the benefit thereof to some person, or persons' [6, pp.86-87].

and justly caused to make restitution to his victims, all subject to lawful process. But Hobbes does argue, in effect, that if a robber chooses to defy the laws of the land, and returns to a unilateral state of nature, then law-abiding subjects of the commonwealth are still not justified in breaking promises to the robber on the mere grounds that he is a robber.[3] Hobbes seems to be implying simply that two wrongs do not make a right. He therefore prescribes that the victim seize the moral high ground, pay the ransom, and seek compensation by recourse to lawful process.

(ii) Spinoza's Solution

Spinoza takes up this problem in his *Tractatus*, but solves it quite differently:

> . . . suppose a robber forces me to promise that I will give him my goods whenever he wishes. Now since my natural right is determined solely by my power, as I have already shown, obviously, if I *can* get out of the robber's clutches by making a counterfeit promise to do anything he wishes, I have a natural right to do this . . . From this I conclude that a contract can have no binding force but utility; when that disappears it at once becomes null and void. [17, p.131]

An appreciation of Spinoza's solution is most readily gained by comparing the principle features of his consequentialism with those of Hobbes' contractarianism, and thus by identifying the points upon which the two philosophers' systems diverge. On the whole, Spinoza appears to have been profoundly influenced by Hobbes. Hobbes (an inveterate geometer) adopted Euclidean methodology in the exposition of his natural, nominalistic philosophy, so that his arguments are concatenations of carefully defined terms. Spinoza ventured much further than this. His *Ethics* employs not only the methodology but also the literal framework of Euclid's *Elements*; his ethical arguments are set out as so many propositions, replete with demonstrations and corollaries thereof.

First, we have seen the amorality of the Hobbesian state of nature, which is predicated upon the natural right of all men to all things. And what Hobbes calls natural right, Spinoza terms the 'supreme law of nature'; which is,

> . . . that everything does its utmost to preserve its own condition, and this without regard to anything but itself . . . Thus man's natural right is not determined by sound reason, but by his desire and his power. [17, pp.125-127]

[3] This can be contrasted with Locke, who distinguishes between States of Nature, of War, and of Society [10, *passim*]. In the Lockean State of Society, just as in Hobbes' commonwealth, the subject relinquishes the prerogative of exacting personal retribution, in order '. . . to assist the Executive Power of the Society, as the Law thereof shall require' [10, p.371]. But in Locke's system, a highway robber has in effect declared war on society: '. . . he who attempts to get another Man into his Absolute Power, does thereby put himself into a State of War with him' [10, p.297]. Thus, unlike Hobbes, Locke argues

> This makes it Lawful for a Man to *kill a Thief*, who has not in the least hurt him, nor declared any design upon his Life, any farther than by the use of Force, so to get him in his Power, as to take away his Money, or what he pleases from him: because . . . I have no reason to suppose, that he, who would *take away my Liberty*, would not, when he had me in his Power, take away every thing else. [10, pp.297-298, Locke's emphasis]

Moreover,

> To act absolutely in conformity with virtue is, in us, nothing but acting, living, and preserving our being . . . as reason directs, from the ground of seeking our own profit. [16, p.198]

Now let us compare their notions of good and evil. Hobbes' classic relativist definition is set forth in the *Leviathan*:

> But whatsoever is the object of any man's appetite or desire, that is it which he for his part calleth *good*; and the object of his hate and aversion, *evil*. [6, p.32]

In the Scholium to Proposition XXXIX in Third Part of his *Ethics*, Spinoza states:

> By *good*, I understand here every kind of joy and everything that conduces to it . . . By *evil*, I understand every kind of sorrow. For we have shown . . . that we do not desire a thing because we adjudge it to be good, but, on the contrary, we call it good because we desire it, and consequently everything to which we are averse we call evil. [16, pp.139-140]

In effect, these nominalistic conceptions of good and evil are identical.

Finally, compare their views on power. According to Hobbes [6, p.56], 'The power of a man, to take it universally, is his present means, to obtain some future apparent good.' As for Spinoza, the power of a being is '. . . the effort by which it endeavours to persevere in its being' [16, p.114]; and 'By virtue and power, I understand the same thing' [16, p.181]. Given that we persevere in our beings primarily by seeking to satiate our needs, and fulfil our desires, then Hobbes' and Spinoza's notions of power are, for these purposes, equivalent.

For Spinoza, it follows that

> Anything, then, that an individual who is considered as subject only to nature judges to be useful to himself — either through the guidance of sound reason or through the impetus of passion — he has a perfect natural right to desire and indeed to appropriate by any means in his power — by force, fraud, entreaty, or however he finds it easiest . . . It follows that the right and law of nature . . . forbids nothing save what nobody desires and nobody can do . . . [17, p.127]

Now, Hobbes' egress from this state of nature is accomplished by mutual renunciation of natural right, which is the entrance into a contract, followed by mutual promise to abide by the contract, which is the enacting of a covenant. At this juncture, however, Spinoza is faced with a problem. He has already pre-empted the possibility of a persistent covenant by having affirmed that the only binding force on a contract is its momentary utility. Given Spinoza's universal tenet of human nature, namely the straightforward maximization principle

> that no one forgoes anything he thinks good save from hope of a greater good or fear of a greater loss, or tolerates any evil save to avoid a greater, or from hope of a greater good [17, p.129]

then how, if at all, can Spinoza impose a binding condition on a contract? In order to do so in a way consistent with his consequentialism, Spinoza is obliged to introduce a Draconian measure, and to enshrine it as sovereign policy. Spinoza asserts that

> . . . it is foolish to require a man to keep faith with you for ever unless you also try to ensure that breach of the contract will bring him more loss than gain. Now this precaution must be given pride of place in the state. [17, p.131]

The implication is clear, and Spinoza [17, p.133] wastes no time in spelling it out:

> But since everyone's natural right is determined solely by his power (as I have already shown), it follows that in so far as he transfers his power to another — whether voluntarily or by compulsion does not matter [this like Hobbes] — he necessarily surrenders to the other his right as well; and that the man who has supreme power to coerce all, and to restrain them by the threat of a supreme penalty which is universally feared, has also supreme right over all [this unlike Hobbes].

Hobbes adjudges that the ultimate criterion by which one decides whether or not to enter into a covenant and, once having entered into it, decides whether or not to abide by it, is, as usual, fear. Hobbes argues that 'the force of words' is 'too weak to hold men to the performance of their covenants', and that the sole criterion of persuasion is either 'a fear of the consequence of breaking their word [which anticipates Spinoza's position]; or a glory, or pride in appearing not to need to break it' [6, p.92]. Prudently, Hobbes dismisses the latter as 'a generosity too rarely found to be presumed on' [6, p.92], which leaves us once again at Draco's door.

While the similarities between Hobbes' and Spinoza's positions are undeniable, the differences between them are crucial. In Hobbes' system, transfer of one's natural right to all things does not entail transfer of one's right to self-defence; in Spinoza's system, it does. Recall that in Hobbes' commonwealth, a subject may justifiably refuse to obey a command from his sovereign, if such obedience jeopardizes the subject's natural right to self-defence. Of course, at the same time, the sovereign may, with equal justification, have the subject executed for disobedience. Hobbes' moral point is that such a death is not dishonourable, so that while the execution itself would be justified, accusations of cowardice or treason would not be justified. But in Spinoza's society,

> It follows that the sovereign is bound by no law, and that all citizens must obey it in all things, since they must have contracted to do so, either tacitly or expressly, when they transferred to it all their power to defend themselves; i.e. all their right. [17, p.133]

Thus for Spinoza, while disobedience is also punishable by death, that death itself is indeed dishonourable.

Differences between these systems are manifest not only with respect to capital crimes, but also with regard to highway robbery. For if Spinoza relies upon a sovereign to promulgate laws against such conduct, then Spinoza also depends upon the necessarily Draconian teeth in those very laws. Hypothetically then, from the instant a robber abducts Spinoza, the robber is in effect under sovereign sentence of death, provided that credible evidence of the robber's crime be presented. Since the hypothetical victim is also the best possible witness in this case, no sane sovereign would have him executed for having made a counterfeit promise to the robber, for the dual purpose of saving himself and bringing the criminal to justice. One understands from Spinoza that the subject's right to self-defence is transferred to the sovereign, not obliterated by the sovereign, and therefore one allows that the sovereign can mete this right back to a subject if the occasion warrants — can pay it out to him as though it were a coil of rope — and moreover can do so retroactively. One infers, therefore, that Spinoza would break his promise to the robber irrespective of the state in which the promise was made, be it natural or societal. That, in sum, is Spinoza's solution to the problem at hand.

(iii) Kant's Solution

Kant's treatment of, and solution to, this problem is revolutionary and compelling, yet also incomplete and unsatisfactory. To begin with, as Sullivan [18, p.xi] points out, '. . . it is distressingly easy to misunderstand individual strands of Kant's moral thought apart from the context of his entire moral theory'. We will mention in passing an ostensible context of Kant's entire moral theory, but will necessarily focus on individual strands thereof in this discussion, in order to furnish sufficient justification — as opposed to exhaustive specification — of his apparent solution to our problem.

A significant difference between Hobbes and Spinoza on the one hand, and Kant on the other, is that the formers' meditations took place between the twilight of the Dark Ages and dawn of the Enlightenment; the latter's, between the dusk of the Enlightenment and the dawn of Modernism. The materialist and nominalist philosophy of Hobbes, along with the scientific experiments of Galileo, Boyle, and Harvey, can be construed as cogs in the Goldberg contraption of European progress, which erratically lurched western humanity, kicking and screaming, out of that agrarian feudalism parented by Monarchy and Theocracy, and eventually into post-modern feudal-industrialism, adopted by Multinationality and Technocracy. During this process, humanity became vastly more knowledgeable but not altogether wiser; its follies and excesses tended to increase with the availability of the science and technology necessary to perpetrate them. Morality became caught up in the associated machinery more often than did suspenders, and incalculably more lives were thereby mangled. And therein, as Sullivan suggests,[4] may lie the problem to which Kant's moral theory addressed itself most generally.

Above all, Kant is committed to deriving good and evil neither from nominalism, as did Hobbes, nor from consequentialist preponderance of ends over means, as did

[4] Sullivan [18, p.8]:
> Kant's solution to our problem is but a specific case of his general resolution of this paradox of the two viewpoints.

Spinoza. Thus, from the outset, Kant will be able to avoid the charges of moral scepticism, nihilism and hedonism that became corollaries to his predecessors' propositions. What, then, does Kant deem to be 'good'?

> Nothing can possibly be conceived in the world, or even out of it, which can be called good without qualification except a good will [9, p.10] . . . This will, although not indeed the sole and complete good, must be the supreme good, and the condition of every other. [9, p.14]

So, in order to avoid impaling his moral theory on any of the received horns (such as rationalism and superstition as of old; scepticism and nihilism as of new), Kant resorts to impredicative definition. While Kant's concept of will as 'faculty distinct at once from feeling and knowing' [19, p.35] dissociates it effectively from both Hobbesian passion and Socratic examination, Kant has yet to deal clearly with the troublesome impredication — that is, with his notion of goodness itself.[5]

It appears that Kant has freed the will from obfuscating passion, empirical reason, and teleological presupposition alike in order to bind it to moral obligation. Thus, according to Kant, we do not behave morally because we will to do so, where the will is previously conditioned by desire or knowledge or belief; rather, we behave morally because the will is free to recognize a higher governing principle, namely that of duty, where 'Duty is the necessity of acting from respect for the [moral] law' [9, p.19]. In consequence, as for example Teale [19, pp.99-100] points out, 'What makes an action good or bad, right or wrong, is not the end it serves, nor the results it achieves, but the principle by which it is determined'. And while Kant's remarkable deontology is theoretically underived and arguably underivable,[6] it is nonetheless empirically grounded. For Kant (apparently influenced by his Pietist roots) perceived a manifestation of his transcendental ethic in the morality of the ordinary person.[7] By grounding that ethic in ordinary moral consciousness, Kant was able to offer apparent empirical evidence of its universality.[8]

[5] It is Teale's [19, p.5] contention that Kant scrupulously avoids such dealings:

> No further account of the 'sole and complete good' is vouchsafed. Indeed, no further reference is made to this latter conception throughout the book, and the reader is left to infer the distinction between and the relation of 'the sole and complete good' and 'the supreme good' from an argument which at every vital point turns upon assertions which never receive the explanation or defence they evidently require.

[6] 'It seems clear enough that Kant's argument has failed as a deduction of the supreme principle of morality . . . we cannot by inference derive morality from the presupposition of freedom, and still less can we by inference derive the necessity of presupposing freedom . . .' [13, p.244].

[7] In Sullivan's [18, p.4] view,

> Kant had enormous respect for the prephilosophical moral convictions of ordinary people and he in fact based his entire analysis of morality on what he refers to as 'ordinary moral consciousness'. He believed that people do have a fundamentally reliable, if not always clear, grasp of what morality is about, certainly better than most philosophers.

[8] E.g. '. . . to a humble plain man, in whom I perceive righteousness in a higher degree than I am conscious of in myself, my mind bows whether I choose or not, however high I carry my head that he may not forget my superior position. Why? His example holds a law before me which strikes down my self-conceit when I compare my own conduct with it . . .' [8, p.77].

However, Kant's categorical imperative does not depend upon this previous ratiocination, which admits of circularity. His categorical imperative (exactly like Hobbes' so-called laws of nature) is a purely normative claim, intended to complete the full circle of his impredicative moral theory. For thus far we find an ethical system in which goodness means a good will, in which the will is unconditioned by passion and reason alike in order to be guided by dutiful thoughts and deeds, and in which duty is the necessity of acting out of respect for a transcendental moral law, of which the well-meaning common folk have an uncommonly profound intuition (unlike philosophers, whose thoughts are apparently tainted by contemplation). Kant [9, p.46] completes his circular theory by means of his categorical imperative, which implicitly links his moral law with goodness itself: 'Act only on that maxim whereby thou canst at the same time will that it should become a universal law.'

In order to make the linkage explicit, Kant [9, pp.21-22] subjects his imperative to a practical test, which is none other than the problem at hand:

> Let the question be, for example: May I when in distress make a promise with the intention not to keep it? I readily distinguish here between the two significations which the question may have: Whether it is prudent, or whether it is right, to make a false promise.

Kant immediately admits [9, p.22] that 'The former may undoubtedly often be the case'. But he continues by making an important observation, to the effect that just as it may on occasion be prudent to make a false promise, it may also, on occasion, be wrong to make a true promise.[9] Here Kant has given an answer to both Spinoza and Hobbes. To Spinoza, Kant replies that it is prudent but wrong to lie to the robber, because the lie is told in anticipation of immediate consequences. To Hobbes, Kant replies that it is prudent but morally indifferent to tell the truth to the robber, if the truth be told in anticipation of future consequences. Kant then concludes [9, pp.22-23]:

> The shortest way, however, and an unerring one, to discover the answer to this question whether a lying promise is consistent with duty, is to ask myself, Should I be content that my maxim (to extricate myself from difficulty by a false promise) should hold good as a universal law, for myself as well as for others? and should I be able to say to myself, 'Every one may make a deceitful promise when he finds himself in a difficulty from which he cannot otherwise extricate himself'? Then I presently become aware that while I can will the lie, I can by

[9] Kant [9, p.22]:
I see clearly indeed that it is not enough to extricate myself from a present difficulty by means of this subterfuge, but it must be well-considered whether there may not hereafter spring from this lie much greater inconvenience than that from which I now free myself, and as, with all my supposed cunning, the consequences cannot be so easily foreseen but that credit once lost may be much more injurious to me than any mischief which I seek to avoid at present, it should be considered whether it would not be more prudent to act herein according to a universal maxim, and to make it a habit to promise nothing except with the intention of keeping it. But it is soon clear to me that such a maxim will still only be based on the fear of consequences. Now it is a wholly different thing to be truthful from duty, and to be so from apprehension of injurious consequences.

> no means will that lying should be a universal law. For with such a law there would be no promises at all, since it would be in vain to allege my intention in regard to my future actions to those who would not believe this allegation, or if they over-hastily did so would pay me back in my own coin. Hence my maxim, as soon as it should be made a universal law, would necessarily destroy itself.

So Kant makes a true promise to the robber, ostensibly not out of regard for consequences, but because he cannot will that lying should become a universal law. And for Kant, that which we can will to be universal, must be good.

(iv) Kan's Solution

Partly for the sake of symmetry, and partly for the sake of articulating an objection to Kant's 'unerring' method, we can conceive of a fourth solution to this problem. Consider an anti-Kantian philosopher — let's call her Kan — whose moral theory resembles Kant's in salient respects, save that Kan is able to tell the truth, but is unable to will that the maxim of truth-telling become universal law. In sum, whereas Kant can't lie, Kan can. Kan's argument follows.

Although Kant finds that he is quite capable of willing a lie himself, he concludes that he cannot universalize this willingness, because, he claims, either no-one would believe his alleged intentions (since there could be no true promises made), or anyone gullible enough to believe them 'over-hastily' would inevitably attempt to deceive him in turn. Kant concludes that such a maxim must necessarily contradict itself. Suppose that such a maxim were in effect, so that everyone necessarily told untruths. Now suppose that Kant wished others to believe some assertion A; to achieve this end, Kant would have to assert $\sim A$. Others would necessarily assume Kant to be lying; that is, would assume $\sim A$ to be false, and thus infer A to be true. But Kant has thereby possibly violated his own maxim; for, if Kant happens to know that A is actually false, then $\sim A$ must be true — thus Kant would have asserted a truth. The alternative is that Kant assert the falsehood A to begin with; but then everyone would disbelieve him, and assume $\sim A$ to be the case, in which case Kant would have failed to deceive. Thus no-one could actually deceive another, at least not without contradicting the maxim itself.

But surely, counters Kan, this argument too is ventured with respect to consequences. Kan argues thus: in reply to Spinoza, Kant asserts that making a false promise to the robber is prudent but morally wrong, since its appeal is consequentialist; i.e. it is to immediate self-preservation. Moreover, in reply to Hobbes, Kant asserts that making a true promise to the robber is similarly prudent but morally indifferent, just in case its appeal is also consequentialist; i.e. it is to long-term societal preservation. However, in reply to Kant, Kan asserts that making a true promise to the robber without regard to prudence but in conformity with a maxim of universal truth-telling is also morally wrong, because the maxim of universal truth-telling has been invoked specifically in order to avoid the meta-consequence of invoking the alternative maxim (namely that the maxim becomes unworkable in practice). Thus the Kanian claims that, while Kant's moral theory espouses an optimistic transcendentalism that eclipses Spinoza's naked individualism and Hobbes' constrained societism alike, Kant's ethical position is based nonetheless on meta-consequentialist reasoning, albeit of a different degree than the others', but mani-

festly of the same kind.[10]

Kan thus concludes that the only maxim which utterly disregards its own consequences, and meta-consequences alike, is that of lying. So the Kanian makes a counterfeit promise to the robber, and moreover wills that the making of counterfeit promises becomes a universal law. According to her lights, the Kanian's deceit is both prudent and morally right.

III. The Problem as a Game

This problem is amenable to game-theoretic modelling. Each of the above solutions readily lends itself to a matrix representation of choices and their associated payoffs. But before examining the model, we note that there is a striking taxonomic difference between this game and other classes of games heretofore treated by that theory.

Von Neumann and Morgenstern [20, pp.49-50] stipulated that a game must have at least one player (and may indeed have two or finitely many more). In the case of a one-player game, the sole participant usually plays against a state of nature; for example, against probabilistic states of a deck of cards in games of solitaire, or against a deistic versus a non-deistic universe in Pascal's wager, or against a state of the boxes' contents in Newcomb's problem. These are all one-person games against a state of nature because it is understood that the player's choice exerts no causal influence on the possible states of nature themselves.

In our problem, Hobbes, Spinoza, Kant and Kan are not playing two-person games against the robber. Rather, each philosopher plays a two-person game in which he or she becomes both players at two different times. At an earlier time the player either promises to pay the robber, or does not promise to do so. For the sake of definiteness, we now strengthen the latter disjunct, so that the player either promises to pay the robber, or promises not to do so. At a later time, the player either keeps the promise or does not keep it. The payoffs of this game are not simple; rather, compound. Three things are at stake, to be gained or forfeited in combination, depending on the outcome of the choices made: the player's virtue, the player's liberty, and the player's money. Let these be represented by V, L and M respectively. We assume that our four philosophers understand L and M in the same way, at least in terms of possible choices in this game, and thus the gain or loss of physical liberty or money is represented identically in each of their decision matrices. Then again, each philosopher understands V in a distinctively different way, and thus the maintenance or forfeiture of virtue, again in terms of choices, is reflected quite differently in their respective matrices.

For each of these philosophers, a transitive ordering of the compound payoffs is effected precisely by his or her ethical system. Thus, in theory, the players' present and

[10] To those who cavil at this treatment of Kant, let it be said that the spirit of his moral theory is admittedly beneficent. If everyone were Kantian, then arguably there would be no highway robbers. But not everyone is Kantian, and moreover Kant's ethic does not and cannot compel everyone to be Kantian. Thus one may speculate as to why Kant rarely (if ever) ventured beyond Koenigsburg: he thereby avoided highway robbers, and other embarrassments to his moral theory, and thus avoided having to submit his theory to difficult empirical tests. In any case, we have attempted to present a minimally plausible 'textbook' account of the Categorical Imperative, if admittedly for the consequentialist end of fleshing out our matrix with his — and Kan's — solutions.

future choices are predetermined by their respective ethics. Then again, in practice, a player may be tempted to ignore or alter previous ethical prescriptions. This suggests another interpretation of this game: instead of viewing the present player as playing a game against himself in a future state, we can view the present player as playing a game against his own system of ethics. Moreover, since we emphatically reject a central claim of human sociobiology, namely the reducibility of ethics to genetics,[11] we conclude that our problem belongs to a class of one-person games that are played not against nature, but rather against nurture.

Figure (1): Hobbes' Decision Matrix

Hobbes' preferences: $(L \,\&\, V) > M$

Hobbes	pays	doesn't pay
promises to pay	$\underline{V, L, \sim M}$	$\sim V, L, M$
promises not to pay	$\sim V, \sim L, \sim M$	$V, \sim L, M$

Figure (2): Spinoza's Decision Matrix

Spinoza's preferences: $L \supset V, (L \,\&\, M) > (L \,\&\, \sim M)$

Spinoza	pays	doesn't pay
promises to pay	$V, L, \sim M$	$\underline{V, L, M}$
promises not to pay	$\sim V, \sim L, \sim M$	$\sim V, \sim L, M$

Figure (3): Kant's Decision Matrix

Kant's preferences: $V > (L \,\&\, \sim M) \vee (\sim L \,\&\, M)$

Kant	pays	doesn't pay
promises to pay	$\underline{V, L, \sim M}$	$\sim V, L, M$
promises not to pay	$\sim V, \sim L, \sim M$	$\underline{V, \sim L, M}$

Figure (4): Kan's Decision Matrix

Kan's preferences: $V > (L \,\&\, M) \vee (\sim L \,\&\, \sim M)$

Kan	pays	doesn't pay
promises to pay	$\sim V, L, \sim M$	$\underline{V, L, M}$
promises not to pay	$\underline{V, \sim L, \sim M}$	$\sim V, \sim L, M$

Figures (1) through (4) illustrate the four decision matrices, with the respective ethically-prescribed choices underlined in each case.

Hobbes' system demands both that he free himself from the robber, and that he not break his covenant. So Hobbes' transitive ordering of payoffs is $(L \,\&\, V) > M$. Since there is only one outcome that satisfies Hobbes' preferences, he makes the associated choices. That is, he both promises to pay, and pays.

Spinoza's system demands both that he free himself from the robber, and that he maximize his power. For Spinoza, liberty itself is virtuous. Hence $L \supset V$. Spinoza must then choose between liberty with money and liberty without it. Since Spinoza is an unconstrained maximizer, his transitive ordering is $(L \,\&\, M) > (L \,\&\, \sim M)$. Hence Spinoza promises to pay and does not pay.

Kant's system demands only that he be virtuous. Nowhere does Kant prescribe that one ought to free oneself from the robber; indeed, his ethical theory is founded precisely upon an utter disregard for consequentialist payoffs, such as physical liberty or money. Kant demands only virtue, and Kant's virtue requires only that he both will and subscribe to a universal maxim of truthfulness. Thus Kant either both promises to pay, and pays; or he both promises not to pay, and does not pay. Kant's transitive ordering is therefore $V > (L \,\&\, \sim M) \vee (\sim L \,\&\, M)$. Kant has two ways of keeping his promise, and is presumably indifferent between them.

[11] E.g. see Wilson [21, p.3] for a statement of that claim, and Marinoff [11], for a specific repudiation thereof (in the context of structural Prisoner's Dilemmas).

Kan's system also demands only that she be virtuous. However, Kan's utter disregard for consequences prescribes that she adopt the maxim of universal untruthfulness. Hence Kan's transitive ordering is $V > (L \,\&\, M) \;\vee\; (\sim L \,\&\, \sim M)$. Thus Kan also has two ways of breaking her promise, and is indifferent between them.

IV. The Iterated Game

A plausible but not immutable assumption was smuggled into this problem at the outset; namely, that the robber himself is a constrained maximizer, a Hobbesian highwayman who keeps all covenants with his victims. Suppose now that this is not the case. Let the robber be an unconstrained maximizer, a thoroughly Spinozan bandit whose motives are purely consequentialist. How, if at all, does this modification affect the game-theoretic model?

Plausibly, this would be a sufficient condition for the possible iteration of the game. If the Spinozan robber releases his hostage with the expectation that the ex-hostage will return next day with the ransom, and the ex-hostage does so, then the Spinozan robber might well decide to re-abduct him, with the inductive expectation of earning a steady income thereby. And if the ex-hostage does not return next day with the ransom, a determined robber might seek him out, and re-abduct him, in the hope of extracting what was promised but not rendered. In either of these cases, the game would become iterated.

Moreover, even if the robber is of the Hobbesian variety, so long as he continues to rob, the possibility remains that he may inadvertently re-abduct a former hostage; in which case the game between these two would become iterated as well.

Finally, no matter what the ethical predispositions of the robber and hostage alike, either payment or non-payment of the ransom by the hostage, subsequent to his release, might induce the robber to continue robbing. Regardless of whether the robber succeeds in extracting or fails to extract ransom from a given hostage, he may thereby feel encouraged or impelled to persist in his career of crime. Thus, either way, one admits the possibility of an iterated game.

And while the game would be strictly iterated only if the same robber were to abduct the same hostage on more than one occasion, it is also possible to loosen this restriction, and to conceive of a weakly iterated game. If a given robber successively abducts different hostages, or conversely if a given hostage is abducted successively by different robbers, then the game-theorist can surely describe either succession as a weak iteration.

So let our philosophers be confronted by an iterated highway robbery game, whether of the strictly or weakly iterated variety. Do their ethical prescriptions change in light of this modification?

Neither Kant's nor Kan's ethical theories prescribe different behaviour in the iterated case. For both Kantian and Kanian alike, moral virtue consists solely in disregarding what each deems to be the consequences of possible actions, and in adhering to a predetermined universal maxim. To remain respectively consistent, both Kant and Kan must disregard not merely a single set of consequences stemming from one pair of choices, but also multiple sets of consequences stemming from multiple pairs of choices. That is, Kant adheres to his maxim of truthfulness, and Kan adheres to her maxim of untruthfulness, just as often as they are required to make such choices.

In fact, the iterated case shows that they are both functionally reminiscent of Socrates as well as the late Roman Stoics, who extolled the virtue of principle above all else, and who refused to value anything that could be appropriated by others, such as liberty and money, up to and including one's very life.[12] Viewed as such, Kant and Kan value only their respective virtue, because neither can be forced to behave in such a way as to relinquish it. Having made and kept one promise or the other, Kant will either regain his liberty and forfeit his money, or lose his liberty and retain his money. Similarly Kan will either regain her liberty and retain her money, or forfeit her liberty and her money together. In the iterated case, Kant stands to become either a wealthy prisoner or a penniless freeman; Kan, a penniless prisoner or a wealthy freewoman. Since Kan stands to lose or gain much more than Kant, and yet remains indifferent withal for the sake of a principle, Kan's Stoicism is stronger than Kant's.

And neither does Spinoza's ethical theory prescribe different behaviour in the iterated case. Spinoza's rule of straightforward maximization applies each time he is required to make a choice. No new ethical property, either intrinsic to or implied by an iteration of such choices, obliges Spinoza to alter his pure strategy, which is nothing but the iteration of his pure principle. Spinoza's willingness to lie repeatedly to attain his ends confers upon his ethic, if nothing else, the property of reliability.

When we come to Hobbes, however, we discover that his ethical theory furnishes him with ample scope for modifying his strategy in the iterated case. We understand that Hobbes retains his right of nature as an inner perimeter of defence, outside which his other so-called laws of nature are operative, but inside which they may not intrude. At the same time, Hobbes loosely specifies an outer perimeter of tolerance, beyond which his laws of nature should not be expected to hold. The first two laws are most relevant to our account. They are normative appeals which nonetheless contain escape-clauses introduced to prevent the exploitation of the would-be law-abiding subject by any unscrupulous co-subjects.

Hobbes' fundamental law prescribes

> . . . that every man, ought to endeavour peace, *as far as he has hope of obtaining it*; and when he cannot obtain it, that he may seek, and use, all helps, and advantages of war. [6, p.85, emphasis added]

Hobbes' second law prescribes

> . . . that a man be willing, *when others are so too*, as far-forth, as for peace, and defence of himself he shall think it necessary, to lay down this right to all things; and be contented with so much liberty against other men, as he would allow other men liberty against himself. [6, p.85, emphasis added]

[12] E.g. see Seneca: 'All that is best for a man lies beyond the power of other men . . .' [15, p.441]. One finds praiseworthy remarks about Stoicism, particularly but not exclusively in contrast with hedonism, in Kant [8, p.60, p.116, p.127]. Kant's approbation of Stoic morality bears striking similarity to that of Augustine [2, pp.214-215, pp.546-565]; moreover, both hold that the Stoic's unflinching adherence to principle is necessary but insufficient, since its inspiration is not theistic. Like Augustine, Kant ultimately subsumes his entire moral theory under a Divine ontology: ' . . . it is morally necessary to assume the existence of God' [8, p.126].

Although Hobbes lacked appropriate mathematical structures and decision-theoretic axioms for fleshing out the bones of what he has just proposed, it has not escaped the notice of contemporary scholars (e.g. Rawls [14, p.269]) that Hobbes has expressed his conditional willingness to co-operate in the form of an *N*-person prisoner's dilemma. Hobbes is willing to seek peace with his fellows *provided that* he has sufficient grounds to believe they are not planning to war against him; he is willing to lay down his natural right to all things (insofar as this enhances, and does not jeopardize, his right to self-defence) *provided that* he finds others to be so willing; and even then, the measure of his permissiveness remains defensive: Hobbes' tolerance is limited not by what he seeks to gain through the liberties he retains himself, but, rather, by what he seeks not to forfeit through the liberties he accords to others.

Relating this to our problem, it is clear that, even in the one-shot case, Hobbes is not bound to co-operate with any robber whose intentions he suspects. If Hobbes deems the robber to be a covenant-breaker, then Hobbes will not keep his own covenant either. And the iterated case, by definition, is one in which the robber is demonstrably unwilling to lay down his natural right to all things, or at least to as many things as he can get; and thus Hobbes' [6, p.85] mitigating condition comes into full effect:

> But if other men will not lay down their right . . . then there is no reason for any one, to divest himself of his: for that were to expose himself to prey, which no man is bound to, rather than to dispose himself to peace.

Hobbes further cites the Old Testament as a corroborating moral authority; contemporary scholars now cite the principle of cooperative maximization of expected utilities as a corroborating rational authority (e.g. Marinoff [12]).

Thus, in the iterated case, Hobbes adopts (or reverts to) a strategy — that of making a counterfeit promise to escape the robber's clutches — which is functionally equivalent to Spinoza's. Hobbes' ethical theory is relativist, his moral prescriptions are self-preservative, and his social norms are contractarian when not non-self-preservative. While Spinoza's ethical theory is also relativist, his moral prescriptions are consequentialist, and his social norms are Draconian when non-contractarian. Here the instrumentality of the game-theoretic model appears to obtrude, because it seems to blur the important philosophical distinctions between these functional equivalents. In the iterated case, both Hobbes and Spinoza lie to the robber without hesitation, compunction, or remorse, but do so for very different reasons.

Or do they?

In chapter 17 of the *Leviathan*, wherein Hobbes sketches the foundations of his commonwealth, his argument undermines the very rationale of his own contractarianism:

> For the laws of nature . . . without the terror of some power, to cause them to be observed, are contrary to our natural passions . . . And covenants, without the sword, are but words, and of no strength to secure a man at all. [6, p.109]

Hobbes recognizes the necessity of a sovereign power, and argues that covenants are ultimately enforceable only by threat of the sword, or by the sword itself. Thus Hobbes

has certainly paved the way for Spinoza's Draconianism.

We also note that, as Hobbes anticipated Spinoza, so Spinoza anticipated Kant. In the Scholium to Spinoza's proposition LXXII, as White [16, p.lxxxiii] points out,

> Spinoza supposes himself asked whether, under any circumstances, breach of faith is permissible; whether a man might, for instance, tell a lie to save his life. His answer is that the true test is the possibility of making the permission general. This, in substance, is exactly Kant's rule.

As it stands, this claim appears to contradict the whole of Spinoza's ethical theory. The apparent contradiction is speedily resolved, however, by consulting Spinoza upon the point. Spinoza's argument refers not to what a 'man' might do; rather, to what a 'free man' might do [16, pp.238-239]. And since Spinoza previously defines a 'free man' as 'a man who lives according to the dictates of reason alone' [16, p.235], we understand his subsequent discussion to be purely counterfactual.

V. Conclusions

Five conclusions can be briefly stated.

First, it is possible to identify a new class of games; namely, those played against a state of nurture, in contrast to those played against a state of nature.

Second, it is perhaps surprising, but otherwise encouraging, that the game-theoretic model at hand preserves and reflects the more significant philosophical attributes of and differences between the actors concerned. Thus the model is simple (as models ought to be), but does not appear pejoratively instrumental.

Third, in game-theoretic terms, it is clear that Hobbes' strategy is the most robust of those considered. Most generally, a robust strategy is one which tends to perform as well as, or better than, other competing strategies in a given game, especially across a spectrum of variations in attributes of the strategic population or in the rules of the game itself (e.g. see Axelrod [3], [4], Marinoff [12]). Of the four strategies considered in this highway robbery game, Hobbes' alone entails the possibility of choosing whether to keep one's promise or not, in both the one-shot and the iterated case, where the choice is influenced by what the hostage believes concerning the robber's designs. But, ironically, this robustness cannot guarantee that a theoretically prescribed outcome will be realized in any scenario, particularly one in which both abductor and abductee are Hobbesian agents. For, in such a case, the very possibility of choosing may indeed lead to an inappropriate choice, resulting from some misconstrual of the other's true intentions. Hypothetically, the robber who is able to identify a hostage as Spinozan, Kantian or Kanian, may predict what each will do. For Kantian or Kanian hostages, the robber also needs to know what they *say* they will do. And for any hostage, of course, actual (as opposed to hypothetical) identification itself is by no means a given thing to the robber. But note that the robber who positively identifies a hostage as Hobbesian, cannot predict with certainty what that hostage will do.

Fourth, one may also ask whether strategic robustness entails any moral virtues. From a meta-game-theoretic viewpoint, it has been asserted by Howard [7, p.xx] that

rationality means merely choosing the alternative that one prefers. (More rigorously stated: rationality is a dyadic relation, demanding consistency between a previously-ordered preference and a subsequently-chosen option.) It is therefore not rational to employ a robust strategy if the most probable outcome of that strategy is not preferred. Given this, one cannot assert — as does Gauthier [5, pp.2-3 *et passim*] — that it is rational to be moral. For in the game of highway robbery, the philosophers' different preferences are ordered by their respective ethical theories. If the philosophers are said to be rational because they choose the outcomes they prefer, and if their preferences are strongly transitively ordered by their ethical theories, then it is clearly moral to be rational. But is it rational to be moral? In this context, not necessarily. Our value-free conception of rationality here depends upon consistency between preferences and choices, where preferences must first be ordered by value-laden normative ethics. If our four philosophers accepted this premise of absolute rationality and relative morality (their respective metaethics should allow each to admit minimally that his or her conception of the good is not unique), then each of them would deem the others to be rational, but immoral. Thus each would find that 'It is moral for me to be rational, and rational for the others to be immoral'. But none would argue that it is rational for anyone to be moral.

Fifth, highway robbers — be they constrained or unconstrained maximizers — should not attempt to earn their keep by abducting philosophers. They would find that crime does not necessarily pay, because virtue does not contingently spend.

City College
City University of New York

Received December 1992
Revised February 1994

REFERENCES

1. Aristotle, *Politics* (Oxford: Clarendon Press, 1966).
2. Augustine, *City of God* (Harmondsworth: Penguin Books, 1984).
3. R. Axelrod, 'Effective Choice in the Prisoner's Dilemma', *Journal of Conflict Resolution* 24 (1980) pp.3-25.
4. R. Axelrod, 'More Effective Choice in the Prisoner's Dilemma', *Journal of Conflict Resolution* 24 (1980) pp.397-403.
5. D. Gauthier, *Morals By Agreement* (Oxford: Oxford University Press, 1986).
6. T. Hobbes, *Leviathan* (Oxford: Blackwell, 1957).
7. N. Howard, *Paradoxes of Rationality: Theory of Metagames and Political Behaviour* (Cambridge, MA: MIT Press, 1971).
8. I. Kant, *Critique of Practical Reason* (New York: The Liberal Arts Press, 1956).
9. I. Kant, *Fundamental Principles of the Metaphysic of Morals* (London: Longmans, Green and Co., 1916).
10. J. Locke, *Two Treatises of Government* (Cambridge: Cambridge University Press, 1967).
11. L. Marinoff, 'The Inapplicability of Evolutionarily Stable Strategy to the Prisoner's Dilemma', *British Journal for the Philosophy of Science* 41 (1990) pp.461-472.
12. L. Marinoff, 'Maximizing Expected Utilities in the Prisoner's Dilemma', *Journal of Conflict Resolution* 36 (1992) pp.183-216.
13. H.J. Paton, *The Categorical Imperative* (London: Hutchinson, 1947).
14. J. Rawls, *A Theory of Justice* (Oxford: Clarendon Press, 1972).
15. L. Seneca, *Moral Essays* (London: Heinemann, 1928).
16. B. Spinoza, *Ethics* (London: Fisher Unwin, 1894).
17. B. Spinoza, *The Political Works* (Oxford: Clarendon Press, 1958).
18. R. Sullivan, *Immanuel Kant's Moral Theory* (Cambridge: Cambridge University Press, 1989).
19. A. Teale, *Kantian Ethics* (Oxford: Oxford University Press, 1951).
20. J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behaviour* (New York: Wiley, 1946).
21. E. Wilson, *Sociobiology* (Cambridge, MA: Belknap Press, 1975).