Philos Stud (2010) 147:37–58
DOI 10.1007/s11098-009-9449-8

The Invisible Foole

Peter Vanderschraaf

Published online: 6 October 2009

Abstract I review the classic skeptical challenges of Foole in *Leviathan* and the Lydian Shepherd in *Republic* against the prudential rationality of justice. Attempts to meet these challenges contribute to the *reconciliation project* (Kavka in *Hobbesian moral and political theory*, 1986) that tries to establish that morality is compatible with rational prudence. I present a new *Invisible Foole* challenge against the prudential rationality of justice. Like the Lydian Shepherd, the Invisible Foole can *violate justice offensively* (Kavka, *Hobbesian moral and political theory*, 1986; *Law and Philosophy*, 14:5–34, 1995) without harming his reputation for justice. And like the Foole, the Invisible Foole dismisses the possibility that being just preserves goods intrinsic to justice, and will be just only if he fears that others will punish his injustice by withholding the external goods like labor and material goods that he would otherwise receive for their performance in covenants. I argue that given a plausible *folk-theorem interpretation*, Hobbes' response to the Foole's challenge is inconclusive, and depends crucially upon common knowledge assumptions that may or may not obtain in actual societies. I present two analogous folk-theorem arguments in response to the Invisible Foole's challenge, one using the idea that the Invisible Foole's power of concealment might be transitory, and the other using the idea that members of society might stop performing in covenants with anyone, thus punishing the Invisible Foole indirectly, if the Invisible Foole commits sufficiently many injustices.

Keywords Folk-theorem · Indirect punishment · Invisible Foole · Justice-conventionalism · Justice-Platonism · Offensive violation · Reconciliation project

P. Vanderschraaf (✉)
Department of Cognitive and Information Sciences, University of California, Merced, 5200 N. Lake Road, Merced, CA 95343, USA
e-mail: pvanderschraaf@gmail.com

The moral for Hobbesian political theory is that a great deal rests on a distinction that Hobbes himself did not make—and that Hume (1998, 2000) perhaps only glimpsed. This is the distinction between a society composed of rational, self-interested agents and one where the rational self-interest of all agents is continuing common knowledge. For the latter case, the Foole gives a compelling analysis. In the former case, the Hobbesian shadow of the future can operate. How it operates depends on the nature of the uncertainty of the members of society.

Brian Skyrms, 'The Shadow of the Future'.

1 The Foole

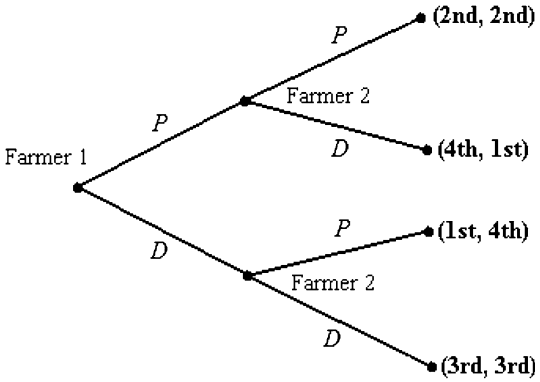
One of Thomas Hobbes' monumental contributions in *Leviathan* is an account of natural law morality based upon a single idea: personal survival. When earlier natural law theorists developed their accounts of the moral law, they relied upon a variety of controversial theological and philosophical doctrines. Hobbes gives a very simple definition of a natural law: A law of nature is a general rule that forbids one to act in a manner that is self destructive or that deprives one of the means necessary to preserve one's life (*Leviathan* 14:3).[1] One of the oldest and most difficult problems of philosophy is what Gregory Kavka dubs the *reconciliation project*, which is the attempt to show that acting morally or justly is compatible with rational prudence (Kavka 1984, p. 297). For moral theorists working in traditions such as Thomism, Kantianism and utilitarianism, reconciling the moral requirements of their systems with rational prudence is a very difficult task.[2] Kavka believes that most philosophers in our time think the reconciliation project is doomed to failure (Kavka 1984, p. 297). But Hobbes has an elegant and perhaps startling proposal for completing the reconciliation project. By adopting his particular definition of a natural law, Hobbes makes the requirements of his moral theory compatible with rational prudence *by definition*.

As Hobbes develops his account of the natural law, he considers a very important challenge to one of its fundamental precepts. Hobbes introduces this challenge through the words of a fictitious interlocutor, the Foole. Hobbes' third law of nature requires one to honor the promises one makes in a covenant. The Foole claims there is no God (*Leviathan* 15:4). The Foole also claims that in certain circumstances, he acts against rational prudence if he obeys Hobbes' third law.

[1] I give the chapter and paragraph number when citing a passage in *Leviathan*, section and paragraph number when citing a passage in *An Enquiry Concerning the Principles of Morals* and book, part, section and paragraph number when citing a passage in Hume's *A Treatise of Human Nature.* For instance, *Leviathan* 13:8 refers to the 8th paragraph of Chapter 13 of *Leviathan* and *Treatise* 3.2.2:22 refers to the 22nd paragraph of Book 3, Part 2, Section 2 of *A Treatise of Human Nature*.

[2] Indeed, some moral theorists in these traditions might dispense with the reconciliation project by denying its importance. For example, a Kantian might argue that a moral choice must not be motivated by considerations of rational prudence, and hence one has no need to reconcile morality with rational prudence. I disagree with this kind of inference, because I concur with Hume who maintains that one cannot expect people to obey the requirements of any moral system that requires them to act against their self-interest (*Enquiry* 9.2:16).

Fig. 1 The farmers' dilemma



> the question is not of promises mutuall, where there is no security of performance on either side; as when there is no Civill Power erected over the parties promising; for such parties are no Covenants: But either where one of the parties has performed already, or where there is a Power to make him performe; there is the question whether it be against reason, that is, against the benefit of the other to performe, or not (*Leviathan* 15:5).

The Foole is aware that if he enters into an covenant with another who honors her promise and performs first, then he has already received whatever benefits were promised him by the terms of the covenant. In this instance, the Foole claims he should not keep his promise. For if he did so, he would incur a cost to himself with no expectation of any additional benefit.

What happens if a party to a covenant who is to perform first agrees with the Foole's reasoning? David Hume has a response to this question in *A Treatise of Human Nature*. Hume has us consider an example:

> Your corn is ripe today; mine will be so tomorrow. 'Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains on your account; and should I labour with you upon my own account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security (*Treatise* 3.2.5:8).

Skyrms (1998) and Vanderschraaf (1998) summarize the situation the two farmers confront as an extensive form game, characterized by the tree diagram in Fig. 1.[3]

[3] Following the usual convention, in a 2-player extensive form game the first (second) component of the ordered pair at an outcome is the first (second) player's preference or payoff. For example, at the *D*, *H*-outcome Farmer 1 is at her least preferred outcome in the game and Farmer 2 is at his most preferred outcome in the game.

Using the Fig. 1 game, one can paraphrase Hume's argument as follows: Farmer 1 reasons that if she were to choose *P*, then Farmer 2's best response would be to choose *D*. So Farmer 1 rules out the *P*, *P*-outcome, and must decide between choosing *P* and arriving at the *P*, *D*-outcome or choosing *D*. Since the *P*, *D*-outcome is Farmer 1's worst outcome, to avoid this outcome Farmer 1 will choose *D*. Farmer 2 will then choose *D*, his best response. According to this *backwards induction analysis*, the two farmers will follow the unique equilibrium of this game where each chooses *D*. Skyrms (1998) and Vanderschraaf (1998) credit Hume with recognizing the basics of the backwards induction analysis of this extensive form game.

In fact, Hobbes and Hume both maintain that self-interested agents can have good reason to perform in covenants. They think that those who reason as do the Foole and the farmers make the mistake of analyzing the costs of performing in a covenant and the benefits of others' performance as if this covenant were the only covenant relevant to their interests. In ordinary circumstances, a single interaction like the commitment problem Hume's farmers face is embedded in a complex sequence of social interactions that occur over time. If one reasons that she can expect many opportunities for mutually beneficial cooperation with others in her community, then if one honors her promise made in a covenant and performs on a given occasion, she indicates to others that she will perform on other occasions, so that they will be willing to make and perform in covenants with her in the future. As Hume puts it,

> I learn to do a service to another, without bearing him any real kindness; because I forsee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me or with others. And accordingly, after I have serv'd him, and he is in possession of the advantage arising from my action, he is induc'd to perform his part, as foreseeing the consequences of his refusal (*Treatise* 3.2.5:9).

What about the Foole's claim that he should not reciprocate, but should instead fail to perform when the time comes? Such failure constitutes what Kavka terms an unexpected unilateral or *offensive violation* of the third law (1986, p. 139; 1995, p. 8). Hobbes and Hume give similar warnings against offensive violation. Hume declares that

> When a man says he promises any thing, he in effect expresses a resolution of performing it; and along with that, by making use of this form of words, subjects himself to the penalty of never being trusted again in case of failure (*Treatise* 3.2.5:10).

In his response to the Foole, Hobbes argues that

> He therefore that breaketh his covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any society, that unite themselves for Peace and Defense, but by the errour of them that receive him; nor when he is received, be retayned in it, without seeing the danger of their errour; which errours a man cannot reasonably reckon upon as the means

> of his security: and therefore if he be left, or cast out of Society, he perisheth; and if he live in Society, it is by the errours of other men,… (*Leviathan* 15:5).

Hobbes and Hume recognize that one who follows the Foole's advice and offensively violates the third law might enjoy an immediate gain by taking advantage of others' compliance. However, they also maintain that prudentially rational and well-informed agents will not perform in covenants with one who has offensively violated a covenant in the past. Hobbes and Hume maintain that if someone like the Foole is guilty of a past offensive violation, the others in society who know of his guilt will violate *defensively* against him.[4] The expected loss of future benefits of others' performance in future covenants far outweighs any immediate gain from a single offensive violation, think Hobbes and Hume.

This kind of rebuttal to the Foole is a *justice-conventionalist* rebuttal (Vanderschraaf 2003). The Foole specifically challenges the prudential rationality of following Hobbes' third law of nature, which requires one to honor one's end of a covenant. As Kavka notes, all the Hobbesian laws of nature have the same two-part structure: a main clause that summarizes a moral requirement one must not violate offensively, and a secondary clause that specifies when one may violate the requirement defensively, that is, when one expects others to violate the requirement and wants to avoid exploitation (1986, pp. 344–349; 1995, p. 8). So one can apply the basic logic of the Foole's argument to all the laws of nature and view the Foole's challenge as a challenge to the entire reconciliation project. Justice-conventionalists argue that to follow moral requirements is generally to one's benefit because this encourages the others in society to act cooperatively towards oneself. Justice-conventionalism is the tradition of Hugo Grotius, David Hume and Thomas Hobbes. This tradition is very well represented in our time by Ken Binmore, David Gauthier, Gregory Kavka, Brian Skyrms and Robert Sugden. Justice-conventionalists argue that one must be just in order to help generate and maintain a system of reciprocal expectations that the members of society will follow mutually beneficial practices such as respecting property rights. Indeed, this argument forms part of an analysis of justice for the justice-conventionalist, that is, the particular norms of justice in a society are precisely those rules its members follow because they expect to enjoy mutual benefits by following them.

2 The shepherd

In *Republic*, Plato (1992) considers several alternative accounts of justice, including a proto-contractarian account Glaucon proposes early in Book II.

> They say that to do injustice is naturally good and to suffer injustice bad, but that the badness of suffering it so far exceeds the goodness of doing it that those who have done and suffered injustice and tasted both, but who lack the power to do it and avoid suffering it, decide that it is profitable to come to an agreement with each other to do injustice, not to suffer it. As a result, they

[4] As with 'offensive violation', the term 'defensive violation' is due to Kavka (1986, p. 139; 1995, p. 8).

> begin to make laws and covenants, and what the law commands they call lawful and just. This, they say, is the origin and essence of justice. It is an intermediate between the best and the worst. The best is to do injustice without paying the penalty; the worst is to suffer it without being able to take revenge. Justice is a mean between these two extremes (*Republic* 358e–359b).

According to Glaucon's discussion, one's best outcome evidently occurs when one does injustice against others who cannot punish the injustice. One's worst outcome evidently occurs when one is the helpless victim of injustice. An intermediate outcome occurs when all act justly. Suppose that one can adopt one of three strategies: Be *unjust* (*U*) and try to exploit others, be *just* (*J*) and act cooperatively with others who are not unjust and retaliate against those who are unjust, and be a *victim* (*V*) and render oneself helpless, say by taking a drug. Then one arrives at one's best outcome when one follows *U* and the others follow *V*. One arrives at one's worst outcome when one follows *V* and the others follow *U*. An intermediate outcome occurs when all follow *J*. Glaucon assumes that all prefer the state where all are just over the state where all are unjust, so I will assume that all strictly prefer the all-*J* outcome over the all-*U* outcome. Finally, I will assume that a just person strictly prefers that others are just over others being victims, and both just and unjust persons prefer that others are victims over others being unjust. That is, I suppose that a just person would rather enjoy the benefits of the active cooperation of other just persons over the passivity of victims, and that both just and unjust persons rank the outcome of conflict with unjust persons lower than the outcomes of meeting victims. Then one can summarize Glaucon's proposal with the Fig. 2 matrix.[5] The Fig. 2 matrix characterizes a game in strategic form.[6] According to this interpretation, justice emerges and remains stable because the all-*J* outcome is the unique optimal equilibrium. The all-*U* outcome is also an equilibrium, but this equilibrium is evidently "fragile". If the others follow *U*, then *U* is a best response, but so is *J*. So if one deviates from the all-*U* outcome and follows this other best response *J*, then *J* becomes the others' unique best response to the deviation.

Glaucon does not merely assume that people regard being unjust against helpless victims as the best outcome, being a helpless victim of injustice as the worst, and all being just as somewhere in between. He defends his claim that one's best outcome is to do injustice against helpless victims with one of philosophy's most celebrated thought experiments. Glaucon repeats a fable of a Lydian shepherd referred to as the ancestor of Gyges. The Lydian shepherd finds a ring, and soon discovers that when he wears the ring he can render himself invisible to others at will. With his newfound power, the shepherd kills the king of Lydia and takes over the kingdom (*Republic* 359d–360b). Glaucon thinks the moral of the fable is clear: Anyone who possesses the power to act as if he were a god among humans, and in particular can

[5] The Fig. 2 matrix oversimplifies the structure of the encounter somewhat because it summarizes the outcome where one's counterparts in the encounter all follow the same strategy.

[6] Again following the usual convention, in a 2-player strategic form game the first (second) component of the ordered pair at an outcome is Player 1's (Player 2's) player's preference or payoff.

Fig. 2 Justice as a mean between extremes

		Others		
		V	*J*	*U*
Myself	*V*	(3rd,3rd)	(3rd,3rd)	(5th,1st)
	J	(3rd,3rd)	(2nd,2nd)	(4th,4th)
	U	(1st,5th)	(4th,4th)	(4th,4th)

$V =$ victim, $J =$ just, $U =$ unjust

commit any injustices against others with perfect impunity, has no reason to be just (*Republic* 360b–360c).

Socrates does not challenge the conclusion Glaucon draws from this thought experiment, though he might have done so. For the invisibility ring did not give the Lydian Shepherd the power to commit injustices with perfect impunity. With the ring, shepherd had the power of invisibility, but not the power of invulnerability. Indeed, a blind assassin with heightened senses of hearing and smell might have been able to dispatch the shepherd even while the shepherd was using his ring. In any event, Glaucon proceeds to argue that the real good of justice is the good reputation one typically gains from being just. Glaucon suggests that the outcome best for oneself is to do the greatest injustice while maintaining the greatest reputation for justice (*Republic* 362a–362b). Glaucon's discussion of the importance of reputation suggests an alternate moral of the fable of the Lydian shepherd: One has no reason to be just if one has the power to perfectly conceal one's identity when committing injustices. For then one cannot damage one's reputation when one commits injustices. Presumably the shepherd had a reputation for being just before he found the ring. The shepherd might have used his ring differently. Rather than setting himself up as the tyrant over the Lydians, he might have committed injustices against various Lydian people, exploiting them while invisible, but never revealing that it was he who possessed the ring. Then the Lydians apparently would have been helpless victims, but not because the shepherd was acting like a god among humans. The Lydians would simply not have known who to punish when the shepherd exploited some of them while invisible.

Glaucon, and later Adeimantus, challenge Socrates to show that someone who is bound to have a perfect reputation for injustice still has a good reason to be just and that someone who is bound to have a perfect reputation for justice still should refrain from being unjust. And Adeimantus anticipates the sort of argument he thinks Socrates will have to give.

> Socrates, of all of you who claim to praise justice, from the original heroes of old whose words survive, to the men of the present day, not one has ever blamed injustice or praised justice except by mentioning the reputations, honors and rewards that are their consequences. No one has ever adequately described what each itself does of its own power by its presence in the soul of

> the person who possesses it, even if it remains hidden from gods and humans. No one, whether in poetry or in private conversations, has adequately argued that injustice is the worst thing a soul can have in it and that justice is the greatest good. If you had treated the subject in this way and persuaded us from youth, we wouldn't now be guarding against one another's injustices, but each would be his own best guardian, afraid that by doing injustice he'd be living with the worst thing possible (*Republic* 366d–367a).

Adeimantus rightly notes that in order to meet the challenge he and Glaucon have laid out, Socrates must show what Adeimantus thinks no one has ever shown. Socrates must show that there is some good intrinsic to justice that is so valuable that one is better off being just even if one has a perfect reputation for injustice and suffers the consequences of having such a reputation. Such a *justice-Platonist* defense of the prudential rationality of justice must maintain that justice brings with it a good or goods unlike goods such as money, material goods or labor that one can obtain by means other than being just (Vanderschraaf 2003, p. 145). A justice-Platonist may allow that some combination of the intrinsic and the external goods of justice gives one good prudential reason to be just. But Glaucon and Adeimantus turn the screws on Socrates, insisting that Socrates defend the prudential rationality of justice in terms of its supposed intrinsic goods only. Justice-Platonism is the tradition of Aquinas, Aristotle and of course Plato himself. In somewhat different ways, John Finnis, Alasdair MacIntyre and John Rawls are prominent justice-Platonists in our own time.

Adeimantus even identifies in advance the good that Socrates will argue is both the greatest human good and the good intrinsic to justice, namely, a soul free from the corruption that injustice causes. To be sure, this is not the only good one might claim is intrinsic to justice. For example, Aquinas gives a justice-Platonist argument for the prudential rationality of respecting property rights when he argues that theft is a mortal sin (Aquinas II.II Q. 66). For Aquinas, if one commits this kind of injustice then one is liable to be excluded from Heaven in the next life. Socrates' key problem, as Adeimantus knows, is that if he accepts the challenge, then he must give some sort of justice-Platonist argument in favor of justice since he may not appeal to any of the goods that come with having good offices with others, as do Hobbes and Hume.

Adeimantus seems to think that Socrates has little prospect of mounting a successful justice-Platonist argument. Some recent authors who discuss Glaucon's thought experiment also appear to take a dim view of justice-Platonist arguments (Gauthier 1990; Copp 1997). Of course, Socrates accepts the challenge and gives exactly the sort of justice-Platonist argument Adeimantus expects. One of the core positions Plato defends in *Gorgias* and *Republic* is the following: If one commits an injustice and thus becomes unjust, this causes one's soul to become corrupt, and to have such a corrupt soul is the worst misfortune one could suffer. Socrates insists that rational prudence requires one to be just in order to maintain one's greatest good, namely, a healthy soul. I will not consider the merits of Socrates' arguments here. Instead, I will consider what, if any, moral Hobbes and Hume should draw from the fable of the Lydian shepherd.

3 Enter the Invisible Foole

Neither Hobbes nor Hume explicitly acknowledges the challenge Adeimantus and Glaucon raise in *Republic*. Perhaps they think they have no good reason to consider this challenge. After all, Hobbes and Hume are concerned with analyzing social life in the actual world, not the world of mythology. They might claim that no invisibility ring or any other technology that would enable one to perfectly conceal one's identity has ever existed, and consequently Glaucon's thought experiment is irrelevant to their analyses of the prudential rationality of justice. They might say that one should concentrate on the skeptical challenge the Foole raises, and not worry about the Lydian shepherd and Glaucon's thought experiment.

But perhaps one should not dismiss Glaucon's challenge on these grounds so quickly. True, no technology we know of can render someone invisible. However, one can take a variety of measures that will make one's identity exceedingly hard to detect in a variety of settings. One might attack and rob a pedestrian on a badly lit sidewalk while wearing a disguise, and leave the victim nursing his wounds and wondering, "Who was that masked man?" One might purchase goods using a stolen credit card number, and leave it to the credit card owner to deal with the surprise additional charges that will appear on his bill. And so on. One might regard the power of the shepherd's ring as a limit no one can actually reach, but also a limit many can approach quite closely. So Hobbes and Hume should regard Glaucon's challenge as more than an irrelevant thought experiment. Given how many ways one can render oneself in effect almost invisible, an argument for why one who possesses the invisibility ring is still better off if he honors the requirements of justice would be a valuable weapon in their arsenals.

Moreover, given Hobbes' assumptions regarding his Foole, such an argument will have to be a justice-conventionalist argument. Suppose the Foole acquires the ring that once belonged to the Lydian shepherd. Now he can turn invisible whenever he wishes, so I will rename him the *Invisible Foole*. And suppose further that after some experimentation the Invisible Foole learns that the ring has another power the shepherd did not notice. Using the ring, the Invisible Foole can appear to others as some person other than himself. With the ring, the Invisible Foole can enter into covenants appearing to be someone else and then turn invisible and abscond when it is his turn to perform. He can commit injustices at will against others all the while concealing his identity perfectly. After he completes his direct reply to the Foole, Hobbes states that the only way one can gain the felicity of Heaven is to obey the law of nature that requires one to keep covenants (*Leviathan* 15:6). But the Invisible Foole still does not believe in God, so he would scoff at Hobbes' statement, just as he would scoff at a Thomist who gives him a similar warning against unjust conduct. If we suppose that the Invisible Foole accepts Hobbes' materialism and mechanistic account of life in *Leviathan*, then we may surmise that the he would regard Socrates' talk of a corrupt soul as meaningless, same as "if a man should talk to me of a *round quadrangle*, or *accidents of bread in cheese*,… (*Leviathan* 5:5)". Indeed, the Invisible Foole appears to have no use for any talk of goods allegedly intrinsic to justice. Before he had the ring, the Foole seemed interested in only external goods such as money and labor he would willingly seek by being just if he had to, but would

seek by other means that would spare him the costs of being just if he could. If we suppose that the Invisible Foole cares for only these sorts of external goods, then of course no justice-Platonist argument will make any impression upon him. Hobbes' Foole claimed that injustice is the better course of action if he receives the benefits of the others' justice no matter what he does. The Invisible Foole appears to have the power to ensure that he is indeed guaranteed to receive the benefits of others' justice whatever he may do. Since many real people possess powers that approximate those of the Invisible Foole, Hobbes and Hume face a serious new challenge to their justice-conventionalist reconciliation of just conduct with rational prudence.

4 The folk-theorem interpretation of Hobbes' response to the Foole

Before dealing with the Invisible Foole further, I will take a step back and reconsider Hobbes' Foole as he was in *Leviathan*, with no invisibility ring. Hobbes maintains that the Foole's reasoning is flawed, and that the Foole should not offensively violate a covenant. Many interpretations of Hobbes' response to the Foole appear in the literature. However, I think the most natural way to interpret Hobbes' response is also one of the simplest. I will call this interpretation the *folk-theorem interpretation*. Robert Sugden (1986, pp. 165–169) and Brian Skyrms (1998) were perhaps the first to clearly recognize and articulate this interpretation.[7]

Hobbes, and later Hume, attribute to those who interact with individuals like the Foole and the corn farmers a policy I summarize as the *Humean strategy* (Vanderschraaf 2007, p. 177)[8]:

> *H*: Perform (*P*) in a covenant with innocent counterpart parties of the covenant, and do not perform (*D*) with guilty counterparts.

where a counterpart is guilty in case he has offensively violated a covenant in the past, and is otherwise innocent.[9] Perhaps the simplest nontrivial example of the Humean strategy in practice takes place in a community of two, where each must decide whether or not to perform in covenants with the other without knowing in advance what the other decides. The Fig. 3 Prisoners' Dilemma summarizes the structure of such covenants.

[7] One might also call this interpretation the *shadow of the future interpretation* since Skyrms presents a particularly fine presentation of the folk theorem interpretation in his essay 'The Shadow of the Future' (1998).

[8] I call this strategy the Humean strategy because I think Hume articulates the supposed consequences of offensive violation more clearly than does Hobbes.

[9] The Humean strategy is unambiguous if one assumes that an *H*-follower follows this strategy perfectly. If one allows that individuals can fail to follow *H* perfectly, then one needs to specify the conditions for innocence and guilt in case one deviates from *H* unilaterally with an innocent or guilty party. Hobbes and Hume do not consider this possibility. If individuals can fail to follow *H* perfectly, then I prefer to define a counterpart as guilty in case he has offensively violated *or* has performed in a covenant with another guilty counterpart in the past. This condition gives individuals greater incentive to bear the costs of punishing the guilty even when they are not the victims of past offensive violations.

Fig. 3 Prisoners' dilemma

		Party 2 P	Party 2 D
Party 1 P		$(1,1)$	$(-l_1, 1+g_2)$
Party 1 D		$(1+g_1, -l_2)$	$(0,0)$

P = perform, D = do not perform

$g_i > 0, l_i > 0, g_i - l_i < 1, i \in \{1,2\}$

In the Fig. 3 game, g_i is the extra gain Party i achieves if Party i successfully exploits his counterpart Party j by following D when Party j follows P. An exploited Party j suffers a loss $-l_j$ when Party j follows P and counterpart Party i follows D. This Prisoners' Dilemma is a close relative of the Farmers' Dilemma game of Fig. 1. However, the Prisoners' Dilemma is structurally simpler than the Farmers' Dilemma because the former has no sequential structure. If the two parties are to enter into a single covenant only, then neither will perform because D is the strictly dominant strategy for each. But if the parties are to engage in a sequence of Prisoners' Dilemma games, then they have available to them a great many new strategies other than following P all of the time or D all of the time. Under certain conditions, they can also settle into a variety of different equilibria of the indefinitely repeated Prisoners' Dilemma, some of which they prefer over others. John Nash first discovered this in 1950, and his proof of this result established the first *folk theorem* in game theory (Flood 1958).[10] The folk theorems are a growing body of results that identify the set of discounted average payoffs in an indefinitely repeated game for each of the players that can be sustained by some equilibrium of this repeated game.[11] In the indefinitely repeated Fig. 3 Prisoners' Dilemma, if each party follows the strategy of choosing D in each round of play, then they follow an equilibrium of the indefinitely repeated game. Vanderschraaf (1998) and Skyrms (2004, pp. 4–6) show that if the parties can also adopt the Humean strategy, then their game is transformed into a *Stag Hunt* game.[12] Figure 4 summarizes the structure of this Stag Hunt when the Prisoners' Dilemma has the Fig. 3 payoffs.

[10] Flood (1958) reports that Nash shared his proof with him in a private correspondence. Nash never published his folk theorem.

[11] Game theorists today refer to these results as folk theorems because early game theorists knew of and discussed some of these results informally for years before the first published proofs of these results appeared in the literature.

[12] Stag Hunt gets its name from Rousseau's example in Part II, Paragraph 9 of *Discourse on the Origins of Inequality* where a party of hunters can catch a deer if all work together, but each hunter is tempted to break from the deer hunt to capture a less desirable but immediately available hare. David Lewis proposes modeling Rousseau's Stag Hunt as this kind of coordination game (1969, p. 47). Gauthier (1979) and Skyrms (2004, Chapter 1) argue that Hume gives even earlier examples of coordination problems with a Stag Hunt structure in *A Treatise of Human Nature*.

Fig. 4 Stag hunt: prisoners' dilemma repeated m times

		Party 2 H	Party 2 A
Party 1	H	(m, m)	$(-l_1, 1+g_2)$
	A	$(1+g_1, -l_2)$	$(0,0)$

H = Humean strategy, $A = D$ always

Fig. 5 Foole's payoffs in a prisoners' dilemma repeated m times with others in a community

		Party $i(t)$ H	Party $i(t)$ A
Party i	H	(m, m)	$(-l_i, 1+g_{i(t)})$
	A	$(1+g_i, -l_{i(t)})$	$(0,0)$

H = Humean strategy, $A = D$ always

Suppose the parties do not know how many times they will enter into covenants with each other, that is, nether knows the value of m, but each expects that $m > 1 + g_i$, where again g_i is the extra gain (and $-l_j$ is the counterpart's loss) if Party i successfully exploits his counterpart in one Prisoners' Dilemma game. Then (H, H) and (A, A) are both equilibria of the indefinitely repeated Prisoners' Dilemma, and both parties would prefer to follow (H, H) over following (A, A). Note that as this indefinitely repeated Prisoners' Dilemma is played out by a fixed pair of counterpart parties, the Humean strategy in this case is equivalent to the strategy for playing repeated Prisoners' Dilemma with a fixed counterpart known as *grim trigger* (Axelrod 1984).

Of course, the Foole is not likely to live in a community of only two. The Foole can rightly observe that the game relevant for analyzing his challenge is likely to resemble a sequence of Prisoners' Dilemma games he plays out with members of a larger community where he might interact with different counterparts over time. If we continue to assume that the pure strategies community members might follow are A and H, this repeated game is summarized by Fig. 5. Here the Foole is Party i. At each period t of play, the Foole meets a counterpart Party $i(t)$, who might be a different counterpart than those the Foole meets at other periods. Using the Fig. 5 diagram, one can summarize the folk-theorem interpretation of Hobbes' response to the Foole this way: At the initial round of play, if the Foole follows D, then he violates offensively and gains the payoff $1 + g_i$ that includes the extra gain g_i of exploiting Party $i(1)$, who

follows P. But afterwards the Foole receives no positive payoff because the other counterparts he meets violate defensively and follow D when they meet the Foole, given that they follow the Humean strategy H. So if the Foole offensively violates at the initial round, then the best the Foole can do afterwards is to continue to follow A by following D in each subsequent round of play. Since all in the community expect that $m > 1 + g_j$ for each community member j, the Foole should expect to do worse if he offensively violates in the beginning than he would if he follows P, which he would follow if he were to adopt the Humean strategy like his fellow community members. So the Foole should not offensively violate at the initial round of play. And if the state where everyone in the community follows H is an equilibrium of the indefinitely repeated game, then by an exactly similar argument one can show the Foole that he should expect to fare strictly worse if he offensively violates at any round of play than he fares if he obeys Hobbes' natural law and consistently performs with his Humean counterparts. This possibility depends upon how likely each community member believes a continuation of the indefinitely repeated game to be after a given round of play. The probability a community member assigns to the event that the repeated Prisoners' Dilemma will be played an additional round after the current round is her *discount factor*. Using a random matching model similar in spirit to those Kandori (1992) and Ellison (1994) develop in their seminal essays on community enforcement, I proved the relevant folk theorems that show that if the community members' discount factors are all sufficiently high, the Humean strategy H indeed characterizes an equilibrium of this indefinitely repeated Prisoners' Dilemma (Vanderschraaf 2007). So if the members of a community are sufficiently confident that their opportunities to engage in covenants with the Fig. 3 Prisoners' Dilemma structure will continue in the future, then everyone including the Foole does better to follow H than to offensively violate.

Did Hobbes refute the Foole, as Hobbes seems to have thought? Given the folk-theorem interpretation, the answer to this question is plainly: "It depends." The Prisoners' Dilemma is the simplest nontrivial game that summarizes the actions associated with a covenant, but one can prove the relevant folk theorems for games that summarize covenants with more complex structures, such as covenants having a Farmers' Dilemma structure. If the Humean strategy characterizes an equilibrium for a community whose members regularly must choose whether or not to perform their respective ends of covenants, and if the Foole expects his fellow community members to all follow the Humean strategy, then the Foole's best response is to consistently perform. But these are rather big ifs! The Foole might challenge the claim that the Humean strategy characterizes an equilibrium of the repeated game. What if some of the members' discount factors are not sufficiently high, as might be the case if they expect the community to break up in the near future? Hobbes will have to hedge his reply and assert that the Foole should expect to live in a stable community whose members expect to interact with each other for a long time, so that their discount factors will be sufficiently high. Perhaps this is a plausible assumption, but we see already that the persuasive force of Hobbes' response to the Foole depends upon context. If the Foole interacts with others in a sufficiently unstable community, then offensive violation might be the Foole's best response to others' performance after all, given that the Foole thinks it sufficiently unlikely that the sequence of covenants in this community will continue.

The Foole can raise other serious objections to Hobbes' defense of his third law of nature. Kavka (1983, 1986) supplied the Foole with one such objection, an objection Skyrms subsequently refined in his essay 'The Shadow of the Future' (1998). Kavka and Skyrms argue that if one modifies in a seemingly innocuous way the assumption in the folk-theorem interpretation that interactions continue for an indefinitely long time horizon, then it is never rational to perform. Here is a summary of their objection: Suppose it is common knowledge in the community that for some value u, no one lives to see his uth opportunity to enjoy the benefits of others' reciprocal cooperation.[13] (A plausible upper bound on the number of possible interactions is $u = 1,000,000$.) If the $u-1$st round of repeated covenant games like Prisoners' or Farmers' dilemmas were somehow to be reached, the parties should analyze this game as a single-shot game, there being no subsequent round of play. Thus parties familiar with the reasoning of Hobbes' Foole and Hume's farmers would not perform at this round. But if this is the case, then the parties can disregard the future play should the $u-2$nd round be reached, and so this potential round should be analyzed as a single-shot game, and thus the parties would not perform for the same reason they would not perform at the $u-1$st round. And the same reasoning applies to the $u-3$rd potential round, and the $u-4$th potential round, and so on. By this backwards induction argument, or an appropriately strengthened version of it, the Foole can conclude that rational parties never perform, in which case the Foole would act against reason were he to obey Hobbes' third law. Skyrms observes that the force of Hobbes' response to the Foole depends crucially upon common knowledge conditions that Hobbes does not consider. If the community members are rational, know the structure of the indefinitely repeated game, but lack common knowledge of an upper bound to the number of covenant games they might play, then they will be unable to anchor the just described backwards induction argument. Consequently they might follow some equilibrium better for all than the all-follow-A equilibrium. But, perhaps paradoxically, if the community members have common knowledge of the game, their rationality and an upper bound to the number of interactions, then the Foole can conclude that the only possible outcome is the all-follow-A equilibrium (Skyrms 1998, pp. 16–19).

As just noted, the backwards induction argument of the previous paragraph relies upon an ideal common knowledge assumption that might not obtain in actual communities. Yet even if we return to the original assumption that from the perspective of the community members, their interactions are played out over an indefinitely long time horizon, and even if the Humean strategy defines an equilibrium of the repeated interaction, the Foole can mount still another objection. Skyrms recognizes the gist of this objection. Skyrms notes that the folk-theorem argument he attributes to Hobbes shows only that consistently refraining from offensive violations of covenants can be consistent with rational prudence. But Hobbes is evidently trying to show the Foole that refraining from offensive

[13] A proposition A is common knowledge among a group if each member of the group knows that all in the group know A and knows that all in the group can infer the consequences of this mutual knowledge (Lewis 1969, pp. 56–57).

violations *must* be consistent with rational prudence. The folk-theorem argument does not vindicate this stronger claim, because the indefinitely repeated game has equilibria other than the equilibrium of the Humean strategy, and at some of these equilibria the parties do not always perform with innocent counterparts. For example, if all follow *A*, then they are at one of these uncooperative equilibria (Skyrms 1998, p. 16). The Foole can develop this objection as follows: If parties can follow a history dependent strategy like *H*, then why should they not be able to follow other history dependent strategies that Hobbes, and Hume after him, never consider? If the parties can choose from a sufficiently rich set of history dependent strategies, then a whole continuum of different equilibria of the indefinitely repeated Prisoners' Dilemma game become possible. Indeed, there are even equilibria of the indefinitely repeated Prisoners' Dilemma that are optimal like the equilibrium of the Humean strategy, but where some of the parties exploit other parties some of the time and hence fare even better than they would fare at the all-*H* equilibrium (Kreps 1990; Vanderschraaf 2007). Which, if any, of all these equilibria should the Foole expect the members of his community to follow?

In his reply to the Foole, Hobbes assumes that the Foole interacts with the members of a community who all follow *H*. In fact, each community member has literally infinitely many different strategies to choose from in the indefinitely repeated Prisoners' Dilemma. Suppose the Foole enters into a covenant with another party, and this other party performs first. If, as Hobbes supposes, the Foole interacts in a community where the others all follow the Humean strategy, then he had better perform if his expectations that he will continue to interact with these others is sufficiently high. But if the Foole is in a community where most of the others reason as Hume's farmers reason, then in this particular covenant the Foole may have an opportunity to exploit another and actually achieve a net gain. For if the Foole lives in a community whose members are mostly following the all-*A* equilibrium to begin with, then why should the Foole not take advantage of any rare occasions that might come up where the other party actually performs? Or suppose the Foole is in a community whose members consistently perform in covenants. Their behavior is compatible with their following the all-*H* equilibrium, in which case *H* is a best response for the Foole if he expects to remain in this community for a sufficiently long time. However, their behavior is also compatible with their all following a simpler strategy than *H*, namely, follow *P* always. If the Foole interacts with others who follow *P* unconditionally in each round of the indefinitely repeated game, then the Foole's best response is to follow *D* in every covenant. For now the Foole can "fleece" the community of "suckers". Hobbes' reply to the Foole assumes crucially that the Foole interacts with others who follow some strategy that rewards performance in covenants with reciprocal performance and punishes offensive violation with defensive violation. Hobbes gives no argument for why this should be the case.

Still, perhaps one can at least partially rescue Hobbes' response to the Foole if one allows for another possibility Hobbes does not consider, namely, that norms of reciprocity might evolve in societies. Sugden (1986, pp. 116–117) examines another conditionally cooperative strategy for playing the indefinitely repeated Prisoners' Dilemma:

T_n: If innocent, follow D in a covenant with a guilty counterpart party of the covenant, and follow P otherwise.

where here a party is innocent if, and only if, she has never offensively violated a covenant or she has performed n times after an offensive violation. Sugden's strategy generalizes the strategy for playing repeated Prisoners' Dilemma with a fixed counterpart known as *tit-for-tat* (Axelrod 1984). Unlike the Humean strategy, T_n is forgiving, since an offensive violator can regain his innocence by performing n times after she becomes guilty. Sugden proves that if all in a community follow T_n, then they are at an equilibrium of the indefinitely repeated Prisoners' Dilemma if their discount factors are sufficiently high. This equilibrium yields each community member the same expected payoff she would achieve by following the all-follow-H equilibrium. Unlike the equilibrium of the Humean strategy, the equilibrium of the T_n strategy is *self-correcting*, that is, the community members can return to the state where all follow P after a community member offensively violates. Sugden also proves that T_n is *error stable*, that is, T_n is the unique best response if the others also follow T_n, with the caveat that at each period each community member with some positive probability deviates from T_n by mistake (1986, pp. 117–118).[14] Sugden argues that equilibria of *brave reciprocity* like the all-follow-T_n equilibrium are likely to evolve in communities (1986, pp. 119–125). If Sugden is right, then we have an evolutionary explanation for the emergence of a strategy similar to the Humean strategy Hobbes attributes to community members in his response to the Foole. This would plug the gap in Hobbes' folk-theorem argument, at the negligible cost of replacing the Humean strategy with the more forgiving T_n strategy. Indeed, perhaps replacing H with T_n in Hobbes' argument is advantageous to Hobbes. In real human societies, everyone makes mistakes. If it were literally true that real people follow the Humean strategy, then in time in a community of fallible individuals no one would ever perform in covenants because ultimately everyone would be guilty because of a past offensive violation, either by choice or by mistake. T_n is still punitive, but eventually forgives transgressions, which is likely to reflect the way real people react to offensive violations of covenants.

However, Sugden does not address one crucial question. Hobbes and Hume also fail to address this question. How do the community members know who are innocent? If community members cannot reliably distinguish innocent from guilty counterparts, then they simply cannot follow strategies such as H or T_n. In his community enforcement model, Kandori assumes that when parties meet, they exchange credentials that accurately indicate to each other whether the counterpart is innocent or guilty. Kandori likens this to presenting a merchant a credit card. But as he acknowledges, Kandori does not consider the possibility that parties might try to present fraudulent credentials (Kandori 1992, p. 71). Gaylord and D'Andria (1998) explore the possibility that a community can successfully enforce the Humean strategy in repeated Prisoners' Dilemma via informal communication. They test this claim using a computer model of agents who meet in an indefinitely

[14] Error stability is not the same as the evolutionary stability of Maynard Smith (1982). As is now well known, no strategy for playing indefinitely repeated Prisoners' Dilemma is evolutionarily stable in Maynard Smith's sense.

repeated Prisoners' Dilemma and who exchange with each other their "blacklists" of individuals they know to be guilty when they meet. Gaylord and D'Andria show that over time, the agents who follow the Humean strategy fare better than the agents who offensively violate. This would suggest that community members can know how to follow strategies like H and T_n simply via informal gossip. However, Gaylord and D'Andria assume in their computer model that each agent's "blacklist" includes the identities of only agents who are offensive violators. In effect, Gaylord and D'Andria assume that everyone tells the truth always. Elsewhere I show that if one extends their computer model to allow for the possibility that agents include the identities of innocent counterparts on their "blacklists", then offensive violators who add the identities of their innocent victims to their "blacklists" now fare better than Humean strategy followers (Vanderschraaf 2007). For once H-followers meet these offensive violators, they start to apply the Humean strategy incorrectly, punishing innocent victims when they should punish the guilty. This suggests that if a community relies upon informal gossip alone, then a Foole who not only offensively violates but lies and claims his victims offensively violated against him can successfully exploit his community.

So how can the members of a community successfully follow strategies like H or T_n? They can follow such strategies if their communities have certain institutions that facilitate common knowledge. For example, merchants and consumers successfully exchange services and goods using credit cards because it is common knowledge in the community of credit users that passing certain screening programs indicates that the owner of the credit card is in good credit standing. Credit grantors and merchants follow a strategy similar to the Humean strategy, and extend credit or accept payment with a credit card exactly with those they identify as being in good credit standing. The common knowledge that certain screening programs reliably indicate good credit standing is possible in part due to the existence of credit bureaus that record the reports of credit grantors regarding their customers. Obviously, the system of credit exchange could break down if too many of the reports submitted by credit grantors are inaccurate, or if the online credit reporting program is compromised. But this only underscores the more general point that communities can follow equilibria of brave reciprocity only under certain circumstances. Skyrms argues that if the community members have too much common knowledge, then they will be unable to follow the all-H equilibrium. I have argued here that community members might also be unable to follow equilibria of brave reciprocity like all-H or all-T_n if they have too little common knowledge. In any event, Skyrms' and my alternate arguments lead to the same conclusion. Hobbes' reply to the Foole is inconclusive.

5 Responses to the Invisible Foole

At first blush, one might think that the challenge of the Invisible Foole is unanswerable. As noted in Sect. 3, the Invisible Foole will not accept any justice-Platonist argument that purportedly shows that he should refrain from injustice. And as he possesses the ring of Gyges, the Invisible Foole seems to have no good

justice-conventionalist reasons to refrain from injustice. So the Invisible Foole appears to have no prudential reason to be just.

In this concluding section I will argue that here appearances are deceiving, and that the Invisible Foole can have good justice-conventionalist reasons to perform in covenants as Hobbes' third law of nature requires. The arguments in this section draw upon ideas from the folk-theorem interpretation of Hobbes' response to the Foole in *Leviathan* discussed in Sect. 4. The leading idea here is that the members of a community might follow somewhat more complex and more punitive strategies than the forgiving strategies of brave reciprocity discussed above. In particular, an Invisible Foole who offensively violates using the ring might spark a cycle of completely unforgiving behavior among the others that will cost the Invisible Foole in the long run. If this is the case, then the Invisible Foole can have good reason to refrain from offensive violations in order to minimize the risk of launching such a cycle of unrelenting punishment.

A first such justice-conventionalist response to the Invisible Foole is motivated by another objection Socrates might have raised against Glaucon's challenge, namely, that the power of concealment the invisibility ring gives the Invisible Foole could be transitory. After all, one can lose a ring. And the power of an invisibility ring might expire unexpectedly. Additionally, if one has in the past taken advantage of absolute or near absolute power to exploit others and then loses this power, one is liable to suffer a particularly severe punishment for these past offenses. In their classic exchange in Thucydides' *The Peloponnesian War*, the Melians warn the Athenians that should the Athenians treat them unjustly while they remain helpless, the Athenians may well suffer a terrible punishment should they one day find themselves helpless. In time, the Melians were proved right. One could present the Invisible Foole with these and other examples, and advise the Invisible Foole against offensive violation on the grounds that should the Invisible Foole lose the concealment power of the invisibility ring, he will be liable to suffer a similarly severe punishment.

To show how this response to the Invisible Foole can have bite, suppose the members of a community are engaged in the indefinitely repeated game described in Sect. 4 where at each period, one is matched in a Prisoners' Dilemma with a counterpart in the community and the counterparts can vary over time. Suppose further that the relevant institutions exist that facilitate common knowledge, so that the community members can follow equilibria of brave reciprocity. The Invisible Foole will be tempted to use the ring to enter into a given covenant appearing to be someone else, and then become invisible when it is his time to perform. If the community follows the all-T_1 equilibrium, then they will identify the individual the Invisible Foole impersonated as guilty and punish him instead of punishing the Invisible Foole. The hapless victim will have to suffer the punishment and perform with a counterpart who does not perform to regain his innocence, since to do otherwise would leave this victim even worse off. But suppose the community members all follow a variant of the T_1 strategy:

> c^*: Always follow D with a counterpart known to have used an invisibility ring in an offensive violation, and otherwise follow T_1 with one's counterpart.

In other words, community members will forgive an offensive violator who does not conceal his identity and who serves his punishment, but they never forgive one they discover to have used an invisibility ring in offensive violations. A community might have such a draconian rule for dealing with a known ring user partly because ring users are so dangerous to deal with in individual covenants, and partly because ring users lead others to mistakenly punish members who did not offensively violate covenants. Let q denote the probability that in a given period the community members discover the identity of one who has used an invisibility ring to commit injustices, perhaps because the power of the ring expires. Then one can prove a folk theorem showing that if q and the discount factors are sufficiently high, then the community members follow an optimal equilibrium if they all follow c^*. If the community follow the all c^*-equilibrium, then the Invisible Foole can expect to do worse if he deviates from c^* and offensively violates using the ring than if he performs in his covenants with the innocent (Vanderschraaf 2008). So if one relaxes the assumption that the invisibility ring is certain to conceal a user's identity perfectly for all time, then the Invisible Foole can have good justice-conventionalist reasons to be just, after all.

The assumption that the Invisible Foole could lose the concealment power of the ring might be plausible, but one might object that this assumption allows me to obtain a desired conclusion on the cheap. After all, in his exchange with Glaucon and Adeimantus, Socrates never helps himself to this assumption. Socrates accepts the challenge of the Lydian Shepherd in its strongest form, and does not question the assumption that the ring of Gyges conceals one's identity perfectly all of the time. Consequently, Socrates presses a justice-Platonist defense of the prudential rationality of justice because he thinks this is the only kind of defense he can press. What if the Invisible Foole presses his case and claims he is confident he will never lose the power of concealment the ring gives him? Then by hypothesis the community will never be able to identify the Invisible Foole as the guilty party when the Invisible Foole offensively violates. So it would seem that the community members cannot punish the Invisible Foole, in which case he has no good reason to perform in his covenants. But the justice-conventionalist has one final card to play. The community members now cannot punish the Invisible Foole directly for his injustices, because they can never discover who commits these injustices. However, the Invisible Foole should consider the possibility of suffering indirect punishment. If the Invisible Foole offensively violates covenants while concealing his identity, then his conduct might tend to undermine the willingness of others to perform in covenants with anyone. And if the community members stop following the norm that requires them to perform in covenants with the innocent, then no one will be able to enjoy the benefits of others' performance in covenants, including the Invisible Foole. So one can develop a second justice-conventionalist response to the Invisible Foole based upon indirect punishment.

Hume hints at the possibility of indirect punishment in his discussion of conventions in *A Treatise of Human Nature* (*Treatise* 3.2.5: 9) and of the Sensible Knave at the end of *An Enquiry Concerning the Principles of Morals* (*Enquiry* 9.2: 22). Again, one can use game theory to show how this response to the Invisible Foole can have persuasive force. Suppose the members of a community are once

more engaged in the indefinitely repeated Prisoners' Dilemma game described above, but this time they follow a strategy that does not always distinguish between the guilty and the innocent:

> d^*: Follow T_1 at the initial period $t = 1$, and for each period $t > 1$, continue to follow T_1 or change to A (always follow D) according to the following rule: (*i*) If at period t one follows T_1 and one's counterpart $i(t)$ does not offensively violate, then at period $t + 1$ continue to follow T_1, (*ii*) If at period t one follows T_1 and one's counterpart $i(t)$ offensively violates, then at period $t + 1$ with probability p continue to follow T_1, and with probability $1 - p$ follow A, and (*iii*) If at period t one follows A, then at period $t + 1$ follow A.

Now each community member continues to follow T_1 until he meets an offensive violator, in which case with probability p he continues to follow T_1 but with probability $1 - p$ changes irrevocably to following A. Intuitively, each time a d^*-follower who has been forgiving meets an offensive violator, there is some chance that this d^*-follower loses patience and starts to punish everyone she meets all of the time. If the Invisible Foole uses his ring to commit injustices, he might start a *contagion* of nonperformance in covenants that will lead to a collapse of social cooperation, in which case the Invisible Foole will lose the gains of free riding on others' justice. One can prove another folk theorem similar to the community enforcement folk theorems of Kandori (1992) that shows that if the discount factors are sufficiently high, then the community members follow an optimal equilibrium if they all follow d^*. The Invisible Foole can expect to fare strictly worse if he deviates from this all d^*-equilibrium and offensively violates (Vanderschraaf 2008). In this case, the Invisible Foole should perform in his covenants as justice requires even if he is sure he can conceal his identity in covenants, because he prefers to avoid the indirect punishment that he will suffer should he trigger a contagion of A-following with an offensive violation.

Plainly, the two justice-conventionalist responses to the Invisible Foole proposed here are inconclusive, and for much the same reasons Hobbes' response to the Foole of *Leviathan* interpreted as a proto-folk-theorem response is inconclusive. In particular, the Invisible Foole can note that he has no *a priori* reason to assume that those he encounters will ever start following such relentlessly punitive strategies that are the backbone of the direct and indirect punishment arguments of this section. And historical examples of the terrible punishments that follow after some mighty individual or state falls from power are at best thin anecdotal evidence for the claim that these punitive strategies would tend to evolve in societies in which some members had the concealment power of the invisibility ring. The most the analysis here shows is that the Invisible Foole can have good justice-conventionalist reasons to refrain from injustice. If conditions are such that the members of society can and do follow sufficiently punitive strategies of direct or indirect punishment, then the Invisible Foole should obey Hobbes' third law of nature even if he has the ring of Gyges. Under different conditions, where the members of society cannot or will not be so punitive, then to commit injustices using the ring might be the Invisible Foole's prudentially rational choice. So the justice-conventionalist refutation of the Invisible Foole is no more decisive than Hobbes' reply to the

Foole of *Leviathan*. But perhaps this is not so bad. I believe that none of the justice-Platonist or justice-conventionalist arguments philosophers have proposed that purportedly reconcile justice with rational prudence succeed perfectly. I have argued here that an Invisible Foole can have some good justice-conventionalist reasons for performing in covenants as justice requires. This conclusion appears to differ from what Plato thought and what some of his contemporary successors (Gauthier 1990; Copp 1997) have thought. Perhaps philosophers will never fully complete the reconciliation project. But perhaps the response to the Invisible Foole proposed in this essay is a small step towards doing so.

Acknowledgments My thanks to Brad Armendt, Richard Kraut, Kevin Zollman and the participants of the Laguna Workshop 2008 in Honor of Professor Brian Skyrms, hosted by the University of California at Irvine Department of Philosophy, for their many fine comments on early versions of this essay.

References

Aquinas, T. (c1274, 1948). *Summa Theologiæ*, Fathers of the English Dominican Province (trans.). New York: Benzinger Brothers.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books, Inc.

Binmore, K. (1994). *Game theory and the social contract volume I: Playing fair*. Cambridge, MA: MIT Press.

Binmore, K. (1998). *Game theory and the social contract volume II: Just playing*. Cambridge, MA: MIT Press.

Copp, D. (1997). The ring of Gyges: Overridingness and the unity of reason. *Social Philosophy and Policy, 14*, 86–106.

Curley, E. (1994). Introduction to Hobbes' *Leviathan*. In E. Curley (Ed.), *Leviathan* (pp. viii–xlvii). Indianapolis: Hackett Publishing Company.

Ellison, G. (1994). Cooperation in the Prisoner's Dilemma with anonymous matching. *Review of Economic Studies, 61*, 567–588.

Flood, M. (1958). Some experimental games. *Management Science, 5*, 5–26.

Gauthier, D. (1979). Thomas Hobbes: Moral theorist. *Journal of Philosophy, 76*, 547–559.

Gauthier, D. (1982, 1990). Three against justice: The Foole, the sensible Knave and the Lydian shepherd. *Midwest Studies in Philosophy, 7*, 11–29. Reprinted in *Moral dealing: Contract, ethics, and reason* (pp. 129–149). Ithaca and London: Cornell University Press.

Gaylord, R., & D'Andria, L. (1998). *Simulating society: A mathematica toolkit for modeling socioeconomic behavior*. New York: Springer Verlag.

Hobbes, T. (1651, 1991). In R. Tuck (Ed.), *Leviathan*. Cambridge: Cambridge University Press.

Hume, D. (1740, 2000). In D. Norton and M. Norton (Eds.), *A treatise of human nature*. Oxford: Oxford University Press.

Hume, D. (1777, 1998). In T. Beauchamp (Ed.), *An enquiry concerning the principles of morals: A critical edition*. Oxford: Clarendon Press.

Kandori, M. (1992). Social norms and community enforcement. *Review of Economic Studies, 59*, 63–80.

Kavka, G. (1983). Hobbes's war of all against all. *Ethics, 93*, 291–310.

Kavka, G. (1984). The reconciliation project. In D. Copp & D. Zimmerman (Eds.), *Morality, reason and truth* (pp. 297–319). Totowa, NJ: Rowan and Allanheld.

Kavka, G. (1986). *Hobbesian moral and political theory*. Princeton: Princeton University Press.

Kavka, G. (1995). The rationality of rule-following: Hobbes's dispute with the Foole. *Law and Philosophy, 14*, 5–34.

Kreps, D. (1990). *Game theory and economic modelling*. Oxford: Oxford University Press.
Plato. (1992). *Republic* (trans: Grube, G. M. A.), (Rev: Reeve, C. D. C.). Indianapolis: Hackett Publishing Company.
Skyrms, B. (1998). The shadow of the future. In J. Coleman & C. Morris (Eds.), *Rational commitment and social justice: Essays for Gregory Kavka* (pp. 12–22). Cambridge: Cambridge University Press.
Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge: Cambridge University Press.
Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Oxford: Basil Blackwell, Inc.
Vanderschraaf, P. (1998). The informal game theory in Hume's account of convention. *Economics and Philosophy, 14*, 215–247.
Vanderschraaf, P. (2003). Justice-conventionalism, justice-Platonism and the social contract. In P. Heugens, H. van Oosterhout, & J. Vromen (Eds.), *The social institutions of capitalism: Evolution and design of social contracts* (pp. 141–163). Cheltenham: Edward Elgar.
Vanderschraaf, P. (2007). Covenants and reputations. *Synthese, 157*, 167–195.
Vanderschraaf, P. (2008). Indefinitely repeated prisoners' dilemma with information concealing technology. Unpublished manuscript.