2013

Reversible Rewards

Omri Ben-Shahar
omri@uchicago.edu

Anu Bradford
Columbia Law School, abradf@law.columbia.edu

Reversible Rewards

Omri Ben-Shahar, *University of Chicago Law School* and Anu Bradford, *Columbia Law School*

Send correspondence to: Omri Ben-Shahar, University of Chicago Law School, 1111 East 60th Street, Chicago, IL 60637, USA; E-mail: omri@uchicago.edu.

This article offers a new mechanism of private enforcement, combining sanctions and rewards into a scheme of "reversible rewards." The enforcing party sets up a pre-committed fund and offers it as reward to another party to refrain from violation. If the violator turns down the reward, the enforcer can use the money in the fund for one purpose only—to pay for punishment of the violator. The article shows that this scheme doubles the effect of funds invested in enforcement and allows the enforcer to stop violations that would otherwise be too costly to deter. It argues that reversible rewards could be used to bolster the enforcement of rights in selective areas of private and international law and could also be applied strategically in litigation in contexts where compliance incentives are otherwise weak. (*JEL*: K42)

1. Introduction

There are two general ways of inducing people into action. One is to reward them for the desired behavior; the other is to punish them for undesired behavior. The typical normative inquiry in the compliance literature focuses on the carrot-versus-stick selection and asks when one device is more effective than the other (Wittman, 1984; Levmore, 1986; Dari-Mattiacci, 2009; Dari-Mattiacci and Geest, 2010). A basic insight derived from this literature is that sanctions are superior to rewards when

Helpful comments were provided by Oren Bar-Gill, Lee Fennell, Pete Leeson, Ronald Mann, Richard McAdams, Ariel Porat, Eric Posner, Lior Strahilevitz, two referees, and workshop participants at Harvard, Chicago, Michigan, Stanford, and the ALEA 2010 Annual Meeting.

American Law and Economics Review
doi:10.1093/aler/ahs018
Advance Access publication December 18, 2012

they are credible: a credible threat of sanctions does not need to be carried out and thus costs nothing. An effective reward, in contrast, needs to be paid. The typical descriptive inquiry seeks to explain legal patterns: Does the law's (frequent) use of sanctions and (less frequent) use of rewards conform to the theory of efficient enforcement?

This article offers a new insight into how to create incentives to comply when conventional enforcement methods are unable to generate compliance. The idea is simple: instead of choosing between sanctions and rewards, an efficient enforcement scheme could *combine* the two. If sanctions and rewards are interlocked together in a particular way, compliance can be obtained at a lower cost. Specifically, we develop a novel concept of *reversible rewards*; a reward that is offered to the recipient who complies with its donor's demands, reinforced with a threat that the same reward could be reversed against this recipient in case of non-compliance and be used to pay for a sanction. We show that reversible rewards can induce compliance in situations where simple sanctions or simple rewards are too costly and therefore fail.

Reversible rewards work as follows. One party—the "Enforcer," who is seeking to influence the behavior of another party, the "Violator"—sets up a reversible reward fund. The money in the fund is pre-committed and cannot be recovered by the Enforcer. This money is offered to the Violator as a reward for choosing a course of conduct that the Enforcer prefers. If the Violator turns down the reward and does not change its behavior, the Enforcer can use the money in the fund for one purpose only: as reimbursement for the cost of sanctioning the Violator.

We demonstrate that this scheme *doubles* the incentive effect that money spent on enforcement generates. This doubling effect is desirable—in fact, crucial—in situations in which either simple sanctions or simple rewards are ineffective. A sanction or a reward might be ineffective because they are too costly. Sanctions could be too expensive for the Enforcer relative to the harm they save. Similarly, rewards sufficient to buy off the Violator's compliance might be too high. In these scenarios, where ordinary inducements to comply are too costly, reversible rewards can fill the void. Under the conditions that we specify, reversible rewards can achieve deterrence at (roughly) half the cost. This leads Violators to comply in a greater number of circumstances than conventional compliance schemes suggest.

To illustrate intuitively the double effect of reversible rewards, imagine a corporation or an individual trying to influence the policies of election candidates through a campaign contribution of a fixed sum. The money can be offered to candidate A or to candidate B for some quid pro quo—some policy that a candidate is willing to support in return for this contribution. But the contributor can do better than offering one candidate the reward. It can instead offer a reversible reward by setting up a fund, depositing the contribution into it, and offering candidate A the choice between taking this sum and having it directed to candidate B. If candidate A declines, he loses twice: he loses the spending advantage over his rival that this money could buy and he allows his rival to gain a spending advantage by receiving this sum. The "wedge" that the reversible reward creates (the difference in the spending capacity of the two candidates) is doubled to twice the face value of the reward. This means that the contributor can get the same quid pro quo that any simple contribution buys at half the cost by using a reversible contribution.

Reversible rewards could be used in private enforcement of rights: a rightholder could offer a Violator a smaller reward for ending the violation, coupled with the threat that declining the reward would lead to a certain (pre-committed) sanction. For example, a property owner can offer a neighbor a sum of money to cease a nuisance, with the threat that if such a reward were declined, this money would be directed toward some form of retaliation. Reversible rewards could also be used in litigation: a defendant could offer a plaintiff a smaller settlement, coupled with the threat that declining the settlement offer would free this money for trial and attrition. Similarly, reversible rewards can be used in contractual relations. When performance of a service becomes costlier than envisioned, the incentives to renege on the contract arise. Here, the client can offer a modest reward for the completion of the contracted service, coupled with the threat to use the same money to finance an enforcement action, if necessary. Importantly, states could resort to reversible rewards to enhance compliance with international law, which suffers from weak or inexistent supranational enforcement institutions. To induce action by another state—for example, to reduce carbon emissions or to end a nuclear program—a fund can be established and offered as a reversible reward. If turned down, this fund would then finance what is otherwise a costly and non-credible threat to impose economic or

even military sanctions. Compliance can thus be bought at a substantially lower cost.

The concept of reversible rewards contributes to the rewards-versus-sanctions inquiry. It also advances the literature on credible enforcement by private parties. This literature has wrestled with the question how legal rights can be protected when rightholders' threats to enforce those rights are not credible. The solutions to the credibility problem often build on some nuanced understanding of the costs of enforcement and on various mechanisms to commit to expending these costs (Bebchuk, 1988, 1996; Bebchuk and Guzman, 1996; Bar-Gill and Ben-Shahar, 2009). In this paper we offer a solution that has not been previously proposed or applied. Our task, therefore, is to argue that reversible rewards are plausible, promoting private enforcement in areas where the reach of ordinary sanctions and rewards is limited.

Before proceeding, it is important to understand the intended scope of the reversible rewards and the central limits to its application. A crucial assumption underlying the idea of reversible rewards is that sanctions are costly. When that is not the case, simple sanctions always dominate reversible rewards: a credible threat to sanction the Violator can deter the violation at no cost to the Enforcer. Our inquiry is, therefore, limited to situations where the enforcement of rights is costly and no credible threat to sanction exists as a result.

However, we must also ask whether the pre-commitment device that we rely on to bolster the reversible rewards could similarly be harnessed to enhance the credibility of simple sanctions. If this were the case, reversible rewards would lose their advantage over sanctions. The intuition behind this argument is the following. If the Enforcer can pre-commit to imposing the sanction when the reversible reward is declined, it should similarly be able to make an irrevocable commitment to simply sanction the Violator. With such a commitment in place, the Violator would be deterred and the costly sanction would not have to be expended.

Yet this logic fails upon a closer scrutiny. Such a pre-commitment to a simple sanction would not be superior to a pre-commitment to reversible rewards because it would be more expensive than the mechanism we propose. If pre-commitment is achieved by sinking the costs of enforcement into an irrevocable fund, the advantage of a reversible reward in comparison

with a simple sanction is that it requires a smaller fund. To be sure, in equilibrium the reversible reward fund ends up being depleted (as the reward has to be paid), whereas the simple sanction fund remains unspent (yet the violation is deterred). But, importantly, even if unspent, the money sunk upfront to create a credible threat to inflict a simple sanction may not be recouped in any way, or else the pre-commitment would be compromised and the incentive to inflict the sanction undermined.

The reversible rewards mechanism is cheaper than a mechanism relying on sanctions alone, but its practical application may be limited. First, as we noted, simple sanctions are always superior to reversible rewards when a threat of sanctions is credible and thus need not be carried out. Simple sanctions are also superior if the same pre-committed fund can be reused to address sequential violations by multiple actors. In equilibrium the money in a simple sanction fund remains unspent. Accordingly, it can be reused to induce compliance of additional potential violators. In contrast, the funds necessary to finance multiple reversible rewards would have to be sequentially replenished. Accordingly, we conclude that reversible rewards are potentially valuable in settings of unique, specific violations and are not valuable in the case of systematic, generic violations.

One last methodological clarification is in order. Reversible rewards help parties enforce private rights, regardless of the social desirability of such enforcement. The next section illustrates our argument using a scenario in which the violation of the private right is inefficient, and thus the improved enforcement through reversible rewards is socially desirable. But we note that private parties may at times use rewards to deter socially desirable conduct as well. In this sense, the analysis is descriptive, not normative.

2. An Illustration of the Argument

Party A engages in an activity that is harmful to Party B, which Party B seeks to stop. Denote Party A as the "Violator" and Party B as the "Enforcer." There are two ways in which the Enforcer can induce the Violator to discontinue its harmful behavior. It can either sanction the Violator for causing the harm or it can reward the Violator, conditional on the Violator ceasing the violation. Both sanctions and rewards are assumed to be

costly for the Enforcer. Rewards are costly because they have to be paid out in full. Sanctions are costly for various reasons: they often entail litigation costs necessary to collect compensation, resources spent on retaliation, and the possibility of counter-retaliation by the Violator, to name a few. The challenge is to devise an enforcement scheme that stops the violation at the minimum private cost to the Enforcer.

To make the problem concrete, assume that the Violator's activity causes a harm of $100 to the Enforcer. Assume, also, that the gain enjoyed by the Violator is only $80. It is, therefore, socially optimal to cease the violation. To achieve this, the Enforcer can inflict a sanction s on the Violator, but let us assume that the cost of inflicting such a sanction is greater than s. Specifically, assume that the cost is $1.5s$. For example, to inflict a loss of $100 on the Violator the Enforcer would have to bear a cost of $150. Alternatively, the Enforcer can offer a reward r, and let us assume that the cost of such reward is r. That is to say, sanctions cost the Enforcer more than the pain they inflict on the Violator, whereas rewards consist of simple transfers of cash.

2.1. Simple Sanctions

The Enforcer can impose any level of sanction on the Violator. In order to deter the violation, the sanction has to be at least $80, which equals the Violator's gain from violation. Thus, such a sanction costs the Enforcer at least $120.

A simple sanction is not effective in this situation. Often sanctions are levied after the harm has already occurred and thus have only a retaliatory effect. But assuming that the sanction has an incapacitating effect, forcing the Violator to cease the harmful activity, it would only be rational for the Enforcer to impose it if the cost of such a sanction does not exceed the harm it eliminates. Since the cost of such a sanction for the Enforcer is at least $120, it exceeds the harm of $100 that the Enforcer suffers from the violation. Thus, the Enforcer's threat to inflict even an incapacitating sanction is not credible. A Violator, recognizing this, is not deterred by the threat of a sanction. Thus, simple sanctions fail to stop the violation.

The simple sanction would not work even under an assumption that the Enforcer could pre-commit the funds to later pay for the sanction. The sum necessary to fund such a pre-commitment would still be $120, more than

the cost of the harm. With such pre-commitment, the violation would be deterred and the fund would not have to be used. But since the money is committed, it may not be recouped by the Enforcer, once the violation is deterred, or else it would not be regarded as a sunk cost. And if the money is not considered truly sunk, the threatened sanction loses its deterrent effect. Thus, the dilemma remains: The Enforcer must spend $120 to stop the harm, or endure the $100 harm.

2.2. Simple Rewards

Alternatively, the Enforcer can induce the Violator to cease its violation by offering a reward. Since the Enforcer has more to lose from the violation than the Violator has to gain—recall that we assume that the violation is inefficient—there is room for a Coasian bargain, a "bribe." Any reward of at least $80 and of no more than $100 would make both parties better off. Let us assume that the Enforcer can successfully offer a reward of slightly more than $80 in return for the Violator ceasing the violation. Under some conditions, rewards would cost more (up to $100). But the question we explore here is whether the Enforcer can do even better. Can compliance be induced at a cost lower than $80?

2.3. Reversible Rewards

The core contribution of this paper is to devise an enforcement mechanism that reduces the cost of enforcement. We call it "reversible rewards" because it uses a reward to lure the Violator, but also reverses the reward against the Violator if the Violator continues its harmful conduct. The reversible rewards scheme has three simple elements:

(1) The Enforcer deposits money in an irrevocable fund, which could be used for two alternative purposes, as follows.
(2) If the Violator ceases the harmful activity, the entire money in the fund is given to the Violator as reward.
(3) If the Violator does not cease the harmful activity, the money in the fund is used to reimburse the Enforcer for the cost of sanctioning.

Under this scheme, the reward offered to the Violator is backed up by an explicit threat: if the violation continues, not only will the Violator forfeit the reward, but it will also bear a sanction. Other than for sanction

reimbursement, the Enforcer cannot recoup the money in the fund. This renders the Enforcer's threat to impose the sanction credible.

A reversible reward would be significantly lower than the \$80 that was necessary for the simple reward. In fact, we show that the lowest sufficient reversible reward is \$48. To see why, consider first the maximum sanction that the Enforcer would be willing to impose if the violation continues. Expecting to be reimbursed up to \$48 from the fund, the Enforcer would have an incentive to inflict a sanction as high as $s = 32$. This sanction would cost the Enforcer $1.5s$ (namely $1.5 \times 32 = 48$), exactly the amount available in the fund. A sanction higher than \$32 would not be fully reimbursed and thus the threat to impose it would not be credible.

Recognizing the credibility of the threat to inflict a sanction of $s = 32$, the Violator has to choose between two options: (i) a violation, which would entail a net payoff of \$48 (that is, a gain of \$80 from continuing violation minus a sanction of \$32) or (ii) no violation, which would entitle him to the reward of \$48. Thus, endowing the fund with a little more than \$48 would be enough to make the Violator strictly prefer compliance to violation. A reversible reward of at least \$48 can therefore lead to full compliance.

2.4. Reversible Reward: Why It Works

The example above illustrated that the reversible reward scheme can succeed where simple sanctions fail and that it costs less than a simple reward. Two distinct factors interact to explain why the success of this scheme is general and not merely an artifact of the particular example we chose: (1) the *double effect* of the expenditure—using the same money to fund both the reward and the punishment and (2) solving the problem of the credibility of threats to sanction through a *pre-commitment* device.

2.4.1. *The double effect.* A reversible reward uses the same money twice. In Section 1, we illustrated this double effect through an example of how a campaign contribution is offered to two opposing candidates, operating once as a carrot and another time as a stick. There, a non-complying candidate loses twice: first by forgoing the offered campaign contribution and second, by having his or her opponent gain the advantage that the contribution buys. Put differently, the enhanced incentive to comply is generated

by a "wedge" between the payoffs available from violation and from compliance. The greater this wedge, the stronger the incentive. This wedge can be "stretched" in two directions: the Enforcer can offer a higher payoff for compliance or a lower payoff for violation. A simple reward operates in the first direction by offering a higher payoff for compliance. A simple sanction operates in the second direction by offering a lower payoff for violation. A reversible reward operates in both directions, doubling the absolute size of the wedge.

The idea of resorting to both rewards and sanctions to influence parties' incentives is not novel. But conventionally sticks and carrots are presented as alternative ways of enhancing compliance. Under the law of restitution, a party who commits a desirable act—rescue, salvage, enhancement of property value—can in some circumstances collect a reward from the beneficiary. Under tort law, a party who commits an undesirable act—injury, damage, destruction of property value—is in most circumstances liable to pay compensation to the injured party. Saul Levmore (1986, 2000) has identified cases of joint use of sanctions and rewards. For example, some jurisdictions provide rewards for rescue and penalties for failure to rescue. Or stores incentivize sales staff by offering commissions for high sales and penalties for low sales. Furthermore, construction contracts sometimes include "risk versus reward" terms, with penalties for delayed completion and rewards for ahead-of-schedule completion. The reversible reward mechanism is similar in utilizing the two-sided incentives, but it is designed to address an additional concern—costly enforcement—by linking the funding of the sanctions and rewards.

We use the term "double" effect loosely. More precisely, the multiplier effect of reversible rewards would depend on the cost function of the precise sanctions. In the example above, we used a cost multiplier of 1.5s, and as a result the reward fund of \$80 was reduced to somewhat less than half, \$48. A multiplier of 2s, for example, would weaken the double effect, requiring a reversible reward to at least \$54.[1] A more effective sanctioning mechanism, by contrast, would more than double the effect. For example, a cost multiplier of less than 1, say 0.5s, would require a reversible reward of only

1. Such a reversible reward fund could finance a sanction of \$27. The Violator would prefer to take the reversible reward of \$54 rather than commit a violation and net \$80 − \$27 = \$53.

$27.[2] Thus, the cost structure of the sanction affects the exact size of the necessary reversible reward fund, but its qualitative effect of delivering a "dual punch" is maintained in all settings.[3] Simply, the more effective the sanctions, the smaller the necessary reversible reward.

2.4.2. *Credible commitment.* Reversible rewards can be used with or without a pre-committed fund. We mentioned the risk-versus-reward clauses in performance contracts and bonus-versus-fine schemes in employment contracts or incentive schemes. Such devices allow the principal to motivate an agent using smaller magnitudes of bonuses and fines. The reversible reward scheme we developed here adds another aspect—a pre-commitment of the fund—which bolsters the credibility of the threat to reverse the reward into a sanction. Because of this pre-commitment of funds, the Enforcer has nothing to lose by carrying out its threat to sanction. This is particularly helpful in situations where the Enforcer has to inflict the sanction *after* having already suffered the harm, and thus would have no incentive to do so.

The problem of pre-commitment to enforcement has been studied before, and this paper offers no new theory on how to resolve it. Contracts, irrevocable trusts, agency, reputation, sunk costs, and various combinations of these mechanisms have been identified as commitment devices.[4] Reversible rewards are not a new commitment device, but rather a method to take advantage of a particular pre-commitment strategy: the

2. Such a reversible reward fund could finance a sanction of $54. The Violator would prefer to take the reversible reward of $27 rather than commit a violation and net $80 − $54 = $26.

3. In some settings, a reward could also entail excess cost—the money paid by the Enforcer is more than the money received by the Violator. This factor would reduce the efficacy of rewards generally, but it would not undermine the advantage of reversible rewards. In our original numerical example, if a reward of r now costs $1.25r$, a simple reward of $80 would cost $100. A reversible reward fund, when the sanction has no excess cost, would cost $45. Such a fund could finance a sanction of $45 and a reward of $36. The Violator would prefer to take the reversible reward of $36 rather than commit a violation and net $80 − $45 = $35.

4. Another mechanism that couples a pre-commitment with a double effect is Ian Ayres' StickK concept. See www.stickk.com. *See also* Ayres (2010). There, the pre-commitment is accomplished by contract with a third-party website, and the double effect is created by directing the deposited bond to a charity least favorable to the designator, in the event that the designated obligation is not fulfilled.

irrevocable fund. Thus, the crux of our argument is that to the extent that pre-commitment can be accomplished by pre-funding the cost of sanctions, reversible rewards reduce the necessary fund size.

Simple sanctions may be thought of as superior to reversible rewards under different commitment mechanisms. Consider, for example, a commitment to impose a sanction that is created by *contract.* The Enforcer may enter into a contract with a third party Administrator (similar to a "retainer" contract with an attorney), who, for a prepaid fee, would then be contractually obligated to impose a sanction on the Violator. Knowing that the Administrator is thus bound, the Violator would be deterred. Here, the Administrator would incur no actual cost in equilibrium, and would therefore not have to be paid much ex ante. The problem with such contractual commitment is the incentive of the Administrator to renege at the time when an obligation to impose the costly sanction is triggered by the contract. Knowing this, the Violator would not be deterred. To restore the Administrator's incentive to punish as required by the contract, either the cost of sanctioning has to be pre-funded or the Enforcer has to threaten the Administrator with a sanction for breaching the contract. Either way, the problem of costly sanctions reappears—either for the Administrator (vis-à-vis the Violator) or for the Enforcer (vis-à-vis the Administrator). This suggests that the problem of costly enforcement has to be solved in a way that goes beyond a contractual commitment. Whether the sanctioning is done by the Enforcer himself or outsourced to a third party, it remains costly, and the Enforcer needs to create a mechanism that renders the threat to inflict the sanction credible. The mechanism we rely on is pre-funding the cost of sanctioning, and the advantage of reversible rewards as opposed to simple sanctions is in cutting the size of the fund by roughly one half.

3. Formal Analysis

3.1. Framework

A risk-neutral Violator has an opportunity to engage in a conduct that harms another risk-neutral party, the Enforcer. The benefit to the Violator is b if it pursues the activity and 0 otherwise, and the harm to the Enforcer is h if the Violator pursues the activity and 0 otherwise.

The Enforcer can threaten to impose any sanction s, where s denotes the monetary equivalent of the disutility of the sanction to the Violator. The cost of sanction to the Enforcer is $c(s)$. For simplicity, assume that the sanction cost function is linear: $c(s) = \alpha + \beta s$, where α is a fixed cost of sanctioning and β is a variable cost multiplier. In some cases, β can be negative, as when the Enforcer collects a monetary fine or damages from the Violator. In other cases, β is positive, representing resources the Enforcer has to invest in inflicting the sanction.

The Enforcer can also offer the Violator a reward r for ceasing the activity. The reward is monetary and thus involves a simple transfer from the Enforcer to the Violator and does not generate additional implementation costs.

If the Violator engages in the harmful activity, its payoff is $b - s + r$ and the Enforcer's payoff is $-h - c(s) - r$.

The parties are rational and have perfect information. The timing of their interaction is as follows: at time 0, the Enforcer announces the sanction-and-reward scheme. At time 1, the Violator chooses whether or not to pursue the harmful activity. If it does, the Enforcer suffers an immediate harm of h. At time 2, the Enforcer can impose a sanction in retaliation. Alternatively, if the Enforcer promised a reward and the Violator complied with the conditions of that reward, the Enforcer must pay the reward. In this setting, the harm occurs immediately at time 1, before and irrespective of any sanction. A sanction can therefore only inflict some cost on the Violator, but it cannot prevent the Violator from engaging in the activity—hence, the sanction is merely retaliatory. However, we also discuss the setting where the sanction can be used to induce the Violator to cease its harmful activity.

3.2. Simple Sanctions and Rewards

Let us consider as a benchmark the effect of simple sanctions. To deter the harmful activity, the Enforcer has to threaten the Violator with a sanction of at least b. When the sanction is merely retaliatory, this threat is not credible. If it imposes the sanction, the Enforcer's payoff is $-h - c(s)$; if it does not impose the sanction, the Enforcer's payoff is $-h$. The payoff is always higher when the sanction is not imposed. Once the harm has occurred, the Enforcer has no incentive to punish the Violator.

Alternatively, if the sanction can cease the harmful activity and thereby reduce the Enforcer's harm to 0, punishment would be rational only if $c(b) \leqslant h$. The Enforcer would have to impose a sanction of $s = b$ to induce the Violator to cease its activity, and would thus have to bear a cost of at least $c(b)$. The Enforcer would choose to pursue a sanction only if its cost were lower than the harm from tolerating the violation. In this case the violation can be eliminated, and the Enforcer's payoff would be $-c(b)$.

Consider now the effect of simple rewards. To induce the Violator to cease the violation, the Enforcer needs to offer a reward of at least b. The Enforcer would choose to do this if $b < h$; that is to say, when it is cheaper to incur the cost of buying off the Violator's compliance than to suffer the harm of Violator's non-compliance.

From the Enforcer's perspective, an incapacitating sanction is superior to a reward whenever the threat to impose such a sanction is credible (whenever $c(b) < h$)—here, the Enforcer would be able to deter the violation at no cost since the threat need not be carried out. In contrast, rewards are superior to sanctions when two conditions hold:

(1) The threat of sanctions is not credible
(2) $b < h$.

Thus, if $b < h < c(b)$, the reward works whereas a sanction does not. Finally, if $h < c(b)$ and $h < b$, neither sanctions nor rewards work—the Enforcer would prefer to bear the harm.

In the remainder of the discussion, we will assume that $c(b) > h$ and that simple sanctions are thus too costly to be credible. We will explore whether reversible rewards could induce compliance at a lower cost, and in a greater set of circumstances, than simple rewards.

3.3. Reversible Rewards

At time 0, the Enforcer sets up a fund and endows it with U. The Enforcer instructs that the fund can be used for either rewarding the Violator for compliance or, failing that, rewarding the Enforcer for punishing the Violator. These instructions cannot be modified or revoked. Specifically, the Enforcer instructs that:

- If the Violator refrains from violation at time 1, the fund's endowment will be transferred in full to the Violator at time 2.

- If the Violator commits the violation at time 1 and the Enforcer punishes him at time 2, the Enforcer's actual cost of punishment will be reimbursed from the fund, up to the full amount available in the fund.

- If the Violator commits the violation and the Enforcer does not punish him, the money in the fund is squandered (e.g., donated to a neutral charity).

Denote by $s^*(U)$ the maximum sanction that could be fully reimbursed from a fund i.e., the highest possible sanction that meets the condition $c(s) \leqslant U$. When $c(s) = \alpha + \beta s$, then

$$s^*(U) = \frac{U - \alpha}{\beta}$$

For example, if $\alpha = 0$ and $\beta = 2$ (namely, $c(s) = 2s$)), then $s^*(U) = 1/2U$. The costlier it is to impose a sanction (i.e., the higher the values for α or β), the lower the maximum sanction that the Fund can credibly support.

What is the minimum necessary fund to induce the Violator to forgo the benefit b and thus refrain from the harmful activity altogether? Denote the minimum fund by $\underline{U}(b)$. The maximum sanction that could be reimbursed from this fund is $s^*(\underline{U}(b))$. The Violator is faced with a choice: either to refrain from the undesired activity and accept the reward of $\underline{U}(b)$, or engage in the activity with a payoff of $b - s^*(\underline{U}(b))$. Given the Violator's choice, the Enforcer chooses the minimum U that induces the Violator to refrain from the activity:

$$U(b) \geqslant b - s^*(U(b))$$

which, after plugging for $s^*(\underline{U})$ and rearranging, yields:

$$\underline{U}(b) = \frac{\alpha + \beta b}{1 + \beta}$$

Several observations can be made:

1. *Cheaper than simple rewards.* The cost of a reversible reward is lower than the cost of a simple reward any time

$$b > \frac{\alpha + \beta b}{1 + \beta}$$

which holds whenever $b > \alpha$. Notice that reversible rewards are superior to simple rewards irrespective of β, the marginal cost of sanctions. The intuition is this: any time $\alpha > b$, the fixed cost of sanction will deplete the reward fund before any pain can be inflicted on the Violator. That is, sanctions are rendered completely ineffective. But if α is lower, sanctions can be used to some (nonzero) degree, and with the ability to impose some sanction, the reward necessary is reduced by the magnitude of this sanction.

2. *Cheaper than simple sanctions.* A fund that is used for the cost of sanction alone has to be funded with $c(b)$. The Enforcer would be better off using a reversible reward scheme any time the cost of sanction exceeds the cost of the reversible reward fund:

$$\alpha + \beta b = \frac{\alpha + \beta b}{1 + \beta}$$

which holds whenever $\beta > 0$. Notice that reversible rewards are superior to a simple sanction irrespective of α, the fixed cost of sanctions. Since α is factored into the Enforcer's cost under both regimes, its magnitude is irrelevant. Notice also that for any positive variable cost of sanction β (that is, any time the cost of the sanction rises with the size of the sanction), a reversible reward is cheaper than a simple sanction, and the higher β, the greater the advantage of the reversible reward regime.

3. *Example.* Assume $c(s) = 100 + s$, and $b = 200$. The minimum effective sanction is 200, which costs 300 to impose. The minimum simple reward is 200. A fund of $\underline{U}$ would generate a credible threat to impose a sanction $s^* = U - 100$. Thus, $\underline{U} = 1/2(100 + b) = 150$. The reversible reward scheme achieves compliance at a cost of 150, less than the cost of simple sanctions or rewards. If $h > 150$, a reversible reward credibly eliminates the harm, whereby a simple sanction is not credible and a simple reward is costlier in comparison.

3.4. Divisible Sanction Costs

By pre-committing a fund, the reversible reward scheme divides the strategic decision into two stages—an initial stage in which the fund is set

and a later stage in which the fund is utilized. We now explore an additional strategic advantage of this two-stage setting: the divisibility of costs. The analysis here responds to the following intuition. If the sanction is costlier than the harm it prevents—in the original example, the sanction of $120 was required to deter a harm of $100—the Enforcer might still deter the Violator by dividing the costs of sanctions across the two periods. In the first period, before the Violator acts, the Enforcer sinks a portion of the sanction costs, leaving the remainder to be spent in the second period, in case the violation occurs. Specifically, the Enforcer may sink upfront only $25, leaving $95 to be spent ex post. Because the remaining portion of $95 of sanction cost is less than the harm of $100, the threat to incur it becomes credible. As a result, the violation is deterred at a substantially lower sunk cost of $25.[5]

We acknowledge the potential of divisible sanctions to generate compliance where non-divisible threats to sanction would not be credible. Yet, we also demonstrate that even under the assumption that the Enforcer can divide its sanctioning costs, reversible rewards are superior to simple sanctions because they require a smaller upfront sunk cost.

3.4.1. *Numerical example.* Return to the example studied in Section 2. We assumed the Violator's benefit to be $80, the harm from the activity $100, and the cost of inflicting a sanction s 1.5s. Simple sanctions are not credible because the Enforcer would need to incur a cost of at least $120 to stop the harm of $100. Cost divisibility could solve this credibility problem. The key would be for the Enforcer to lower its sanctioning costs at time 2, when the decision to inflict the sanction is made. This can be accomplished by dividing its costs into a pre-committed sunk portion and a subsequent avoidable portion. The Enforcer would deposit just enough money in the fund at time 0 (the sunk portion) to render credible its subsequent, time 2 threat to expend the remaining cost of the sanction (the avoidable portion)—thus ensuring that the time 2 threat would not need to be carried out. In the above example, such a scheme would render the simple sanction

5. For a model of the effect of cost divisibility on the threat to enforce, see Bebchuk (1996).

effective. The Enforcer would initially need to deposit just over $20 in the fund. If the Violator subsequently engages in the harmful activity, it would cost the Enforcer less than $100 to inflict a sanction at a total cost of $120. Since the upfront deposit into the fund is sunk and no longer factors into its strategic calculation, it would be rational to spend anything under $100 to terminate the harm of $100. And since the threat to impose a sanction is now credible, the Violator would be deterred. Thus, the Enforcer manages to stop the harm by spending only $20 upfront, never having to actually spend the additional $100.

While the divisibility of simple sanctions can render them cheaper than simple rewards, the costs can be further lowered if, instead, the Enforcer exploits the divisibility feature in setting up a reversible reward fund. Here, the money sunk into the fund at time 0 can be used, not only to fund a subsequent sanction at time 2 but also as a direct reward to the Violator, if the Violator ceases its activity voluntarily at time 1. In this case, we can show that the Enforcer only needs to deposit $8 in the fund—that is, the cost to the Enforcer is reduced from $20 to $8. The reason is that if the Violator is offered $8 as a reward to stop the harmful activity, the Enforcer no longer has to threaten a full sanction of $s = 80$. Instead, a sanction of $s = 72$ would suffice. This is because the wedge between a reward of $8 and a sanction of $72 is, again, $80, equal to the Violator's gain from the activity. Accepting the reward confers to the Violator a payoff of $8. Continuing the harmful activity would also lead to a net payoff of $8 for the Violator: the benefit from activity ($80) minus a sanction ($72). In order to inflict a sanction of $s = 72$, the cost to the Enforcer would be $1.5 \times s$, or $72 \times 1.5 =$ $108. But since $8 would be paid out of the fund, the remaining cost for the Enforcer would only be $100. Hence, the threat to inflict it, and to stop an ongoing harm of $100, would be credible. Thus, setting a fund of just over $8 would make the threat to sanction credible and lead the Violator to cease the activity.

3.4.2. *Formal analysis.* The Enforcer endows an irrevocable fund with U. Consider, first, a scenario in which the fund is used solely to reimburse the Enforcer for the cost of a sanction, but is not offered also as a reward to the Violator. Expecting to be reimbursed up to U, the maximum sanction that the Enforcer can credibly threaten to impose is $s^*(U)$, which is the

solution to:

$$c(s) - U = h.$$

If it inflicts the sanction, the Enforcer stops the harm but incurs a cost of $c(s) - U$. If the Enforcer does not, it incurs a cost the harm, h. Thus, when $c(s) = \alpha + \beta s$, then

$$s^*(U) = \frac{h + U - \alpha}{\beta}$$

If $s^*(U) > b$, the Violator would prefer to stop the Violation, forgo the benefit b, and avoid the sanction $s^*(U)$. Thus, the minimum necessary fund to induce the Violator to forgo the benefit b and refrain from the violation, denoted by $\underline{U}(b)$, must satisfy

$$s^*(U(b)) \geqslant b.$$

Thus,

$$\underline{U}(b) = \alpha + \beta b - h.$$

Notice that the cost of the divisible sanction to the Enforcer is significantly smaller by the amount h than the cost of a simple sanction, $\alpha + \beta b$. Under plausible conditions, it is also cheaper than the cost of a simple reward, b.[6]

The above scenario exploits the divisibility effect in a situation where the Enforcer employs simple sanctions. However, the Enforcer can do even better—i.e., deter the harmful activity at a lower cost—by using Reversible Rewards that combine the divisibility effect with the double-wedge effect. Now, the money in the fund is offered to the Violator in return for ceasing the activity or, alternatively, to the Enforcer for financing the sanction against a non-compliant Violator. The minimum necessary fund to induce the Violator refrain from the activity, $\underline{U}(b)$, must now satisfy

$$U(b) \geqslant b - s^*(U(b)).$$

The Violator's choice is either to refrain from the activity and accept the reward of $U(b)$ or engage in the activity with a payoff of $b - s^*(U(b))$.

6. The cost of the enforcement fund is lower than a simple reward whenever $\alpha + \beta b - h < b$, or $b < (h - \alpha)/(\beta - 1)$. The smaller the fixed cost of sanction and the greater the variable cost, the more likely the enforcement fund to be cheaper than a simple reward.

The Enforcer will thus choose the minimum U that is sufficient to induce the Violator to refrain from the detrimental activity, which yields:

$$\underline{U}(b) = \frac{\alpha + \beta b - h}{1 + \beta}$$

Relative to the simple-sanction fund, the reversible reward fund reduces the size of the minimum necessary fund by a multiplier of $1/(1 + \beta)$. Without the reversible reward, that is, exploiting the divisibility effect alone, the fund needed to be endowed with at least $\alpha + \beta b - h$ for the subsequent threat to be credible. Thus, just like in the basic analysis of reversible rewards (*Remark* 2 above), any time $\beta > 0$—that is, anytime the cost of the sanction increases with the size of the sanction—the reversible reward achieves full deterrence at a lower cost than a simple divisible sanction fund. A reversible reward is also cheaper than a simple reward any time $U(b) < b$, namely, $\alpha - h < b$. Unless the fixed cost of sanction, α, is so high as to overshadow all other costs, the reversible reward scheme makes enforcement more affordable.

4. The Limits of Reversible Rewards

This section identifies two significant limitations of reversible rewards.

4.1. Generic versus Unique Violations

The last example in Section 3 demonstrated that when the cost of sanction is divisible (that is, part of it can be spent upfront and the remainder ex post), reversible rewards are cheaper than simple sanctions, because they require a smaller upfront sunk cost. In that example, the size of the pre-committed fund necessary for a simple sanction was \$20, whereas the size of the pre-committed fund necessary for a simple sanction was only \$8. The reversible reward mechanism dominated the simple sanction.

But now imagine that the violation is "generic": many sequential violators are all engaged in identical consecutive conduct and causing identical harm, and must all be separately deterred. In this scenario, a separate reversible reward of \$8 would be necessary for each of the potential violators, because the money in the fund must in fact be paid out—either as

a reward for compliance or as a reimbursement for punishment. A simple sanction regime, by contrast, would require only a one-time investment of $20, which could be repeatedly used to deter each of the violators, because violators would be deterred without depleting the fund. As long as the potential violators can be arranged along some order—chronological or otherwise—such that the Enforcer can threaten them one by one with a sanction, guaranteeing one's compliance before moving on to the next violator, the single fund of $20 could be repeatedly leveraged.[7]

This is a familiar advantage of sanctions over any type of rewards. When successful in deterring violations, they need not be inflicted. While it might be costlier to set up a pre-committed fund to make the threat credible, once the fund is established, it can be exploited repeatedly at no additional cost. Enforcers who deal with multiple potential violators are therefore likely to prefer one large simple sanction fund—assuming that they do not have liquidity or budget constraints that would prevent them from amassing the funds upfront—that can be used several times over many small reversible reward funds that are paid out and can be used only once, each.

Accordingly, reversible rewards are likely to be more useful in scenarios involving a one-of-a-kind violation. If, say, there is only one potential Violator—such as an individual litigation matter involving one plaintiff and one defendant—the advantage to repeated use of an enforcement fund is irrelevant. Similarly, if the separate sequential violations are aimed at different Enforcers, the single sanctioning fund loses its advantage. This is also true in a setting where there may be multiple Violators but where the Enforcer cares only about the behavior of one primary Violator. For example, in Section 5 we examine an international setting where the Enforcer (the EU or the United States) cares only about violations by the largest Violator (China), which has the ability to inflict significant damage. In this scenario, reversible rewards outperform a simple sanction.

Generic violations also manifest particularly difficult moral hazard problems. In utilizing reversible rewards, the Enforcer must avoid setting a precedent that all good behavior can be made subject to the payment of the reward. Otherwise, multiple individuals would have the incentive to

7. Bar-Gill and Ben-Shahar (2009) discuss the dynamics of such sequential enforcement schemes in the context of resource constrained prosecutors negotiating plea bargaining.

present themselves as potential Violators whose compliance must be bought off with rewards. Reversible rewards are thus suited to strategic settings where the recipient of the reward can be ex ante specified and the use of the reversible reward thereby limited to a unique entity and a unique situation—akin to a single plaintiff in a discrete litigation setting.

4.2. The Limits of Pre-commitment

The pre-commitment element of the fund requires that the money would be truly sunk. As noted above, the problem of pre-commitment has been studied before and does not pose new analytical difficulty when applied to the reversible reward device. The pre-commitment could be accomplished by depositing funds in an irrevocable trust, whereby the trustee is barred from accommodating any conflicting ex post instructions by the fund's initiator. While contract law does not recognize the power of parties to write non-modifiable binary contracts (Jolls, 1997), trust law provides a legal framework to make effective hands-tying commitments (Davis, 2006).

When there are limits to the ability of enforcers to set up legal trusts to fulfill the pre-commitment, other mechanisms can be potentially applied. An Enforcer can contract with an administrator to manage the reversible reward scheme, relying on the reputation of the administrator to prevent ex post modifications. Banks, law firms, arbitrators, and even websites (e.g., StickK.com) specialize in providing such commitment service and sometimes are bound by professional ethics to preserve the original commitment.

Still, we recognize that commitment can be difficult and costly to achieve and that any enforcement device that ultimately depends on pre-commitment—simple sanctions and rewards, as well as reversible rewards—might fail. The advantage of reversible rewards would then fail to materialize in the same way simple sanctions and simple rewards lacking credibility would also fail.

5. Applications

This section illustrates the potential usefulness of reversible rewards in various legal settings, where one party seeks to credibly and cost-effectively change the incentives of another party. Whether it is to perform a contract, refrain from harmful conduct, or settle a suit, the affected party can combine

rewards and sanctions to generate incentives more cheaply than by using sanctions or rewards alone.

Most of our examples focus on private enforcement. In general, reversible rewards are less well suited for public enforcement. Government agencies can credibly deter violations based on their public authority to impose sanctions, coupled with their sunk budgets to enforce law. Some public agencies, however, face budgetary constraints that limit their ability to sanction violations in a way that dilutes the deterrent effect of public enforcement. In some settings, reversible rewards may therefore reduce the costs of public enforcement as well, allowing the agencies to pursue a greater number of violations with the same level of resources, without the downside of diluting the individual violators' incentives to comply. Section 5.5 illustrates one such application.

5.1. Settlement Bargaining

Reversible rewards can be employed by a defendant to improve its strategic position and secure a more favorable settlement.[8] The defendant establishes a fund and offers the money in it to the plaintiff as settlement. If the plaintiff turns down this settlement offer, the defendant uses the money in the fund to pay attorneys to mount a non-compromising defense. To the extent that such defense would make it costlier for the plaintiff to win a judgment, the plaintiff would be better off accepting the settlement offer.

Consider the following illustration. A plaintiff has a claim that, if litigated, would lead to a judgment of $100. If unopposed, the plaintiff would incur no litigation costs. If, instead, the defendant stonewalls the claim, the plaintiff's litigation cost would rise. For simplicity, assume that if the defendant spends any amount C on litigation, the plaintiff would also have to spend an equal amount C to win the $100. In this scenario, the defendant has no incentive to drag the plaintiff to litigation: he prefers to pay $100 outright in settlement rather than incur $100 + C in litigation. Thus,

8. Others have noted how fee arrangements with attorney can affect the strategic structure of settlement bargaining. See, e.g., Bebchuk and Guzman (1996); Croson and Mnookin (1996). Croson and Mnookin examine the effect of pre-committed fee on the *plaintiff's* ability to extract a settlement. Here, instead, we demonstrate the effect of a pre-committed fund on the *defendant's* ability to lower the settlement.

the defendant's threat to litigate and impose costs on the plaintiff is not credible.

The defendant can, instead, utilize a reversible reward in the following way. He would deposit $50 in the fund and offer this sum as final settlement to the plaintiff. If the plaintiff turns down the $50 settlement from the fund and insists on a higher settlement, the money in the fund could be used only to fund litigation cost, up to the full value of $50, which is pre-committed in the fund. Now, the plaintiff would be willing to settle for $50 to avoid litigation, because litigation would yield him a payoff of $50 ($100 judgment minus the litigation costs needs to match the plaintiff's, $50). Because the fund is sunk, the defendant's threat to spend the money to litigate the case is credible. As long as the defendant cannot use the $50 in the fund to pay for a settlement greater than $50 (that is, as long as the maximum settlement paid from the fund is set at $50), the defendant can credibly threaten to litigate by paying an extra $50 rather than settling for the full $100.

In this example, the reversible reward scheme saves the defendant half of the settlement costs by reducing the cost from $100 to $50. In general, the magnitude of the saving depends on the "sanction" that the defendant can impose—i.e., on the proportion by which the plaintiff's costs would rise when the defendant spends C in litigation. If, for example, the plaintiff's costs rise only by 1/2C, then the settlement offer would have to be $67.

Practically, for this technique to work, the defendant has to set up the fund in such a way that would make it impossible to use the money in any way other than stipulated. Specifically, the defendant has to contract with an attorney such that if the settlement offer from the fund is turned down, the attorney must launch a defense with the full sum available in the fund and cannot free the money to pay for higher settlement offers. Otherwise, the plaintiff would be able to undermine this scheme by counter offering a settlement of a little less than $100.

5.2. International Enforcement[9]

Enforcement problems are particularly challenging in the international context. In the near absence of effective supranational enforcement bodies,

9. This section is based on Bradford and Ben-Shahar (2011).

states are not able to rely on an objective third party carrying out enforcement on their behalf. Instead, states are often left to solve their disputes through various diplomatic, political, or—at the most extreme cases—military means. While international law operates in the public sphere, it can effectively be analyzed through the lens of private enforcement model, in which states operate as private enforcers to safeguard their rights with little help from a centralized "police" (Posner and Sykes, 2011). Specifically, states may resort to economic sanctions, and occasionally even to the use of military force—but these tactics are costly and often unsuccessful. Trade sanctions, for example, impose costs on the sanctioning state whose firms and consumers are deprived from the benefits of economic exchange. Other times, states try to enforce international law by offering rewards to violators if they cease their harmful activity. The United States could, for instance, offer direct cash transfer to compensate a polluting country for the cost of reducing pollution and retrofitting its plants. But these rewards, too, are costly and often domestically contentious.[10]

The ongoing effort to negotiate a new global climate change treaty is an illustrative example of a costly enforcement challenge. Recent efforts to enact a new global climate treaty have failed because "enforcers"—states eager to reduce emissions—have been unable to persuade "violators" to join a treaty. The cost of buying off the cooperation of countries like China would simply be too high—China has requested an annual transfer of $300 billion from developed countries to change its emissions practices.[11] The cost of levying effective trade sanctions on economic powers like China is also prohibitive.

In principle, reversible rewards could generate more compliance than the reliance on sanctions or rewards alone. The scheme would work as follows. Enforcers—led by the EU, joined by other states including, possibly, the

10. These enforcement problems are often magnified by collective actions problems. International treaties aimed at solving global cooperation problems are notoriously hard to enforce. International organizations and courts are limited in their ability to levy sanctions on free riders. Individual countries and ad hoc coalitions can at times coordinate sanctions for violations, but for problems of global importance coordination is often elusive.

11. In the Copenhagen Conference in 2009, China requested that developed countries commit one percent of their GDP—amounting to over $300 billion annually—to a fund that would help China and other developing countries to comply with the proposed climate treaty. Levi (2009).

United States—would set up a fund. Instead of endowing it with the full $300 billion, which China is demanding for joining the treaty, Enforcers would deposit only about half the amount in the fund. The fund would reward China for compliance, by financing transformation of its energy infrastructure, transferring environmental technologies, or paying for a host of other tangible inducements. However, if China fails to join a treaty or to fully comply with it, the money in the fund would be used to reimburse Enforcers for the costs of inflicting sanctions against China. If the sanction consisted of a carbon border tax, the fund could compensate adversely affected domestic parties or pay the cost of administering the tariff. Or the fund could grant subsidies for industries that compete with Chinese manufacturers. It could also be used to cover the costs of mitigating the damage from China's possible trade retaliation. Thus, a reversible reward of $150 billion would create an inducement that is roughly equal to a simple reward of $300 billion.

In this context, reversible rewards entail a secondary advantage in that they may mitigate the collective action problem among the various Enforcers who have the incentive to free ride on each other's enforcement efforts. By using a pre-committed fund, each state's participation is measured not by the sanctions it actually levies (on which they have an incentive to cheat and which are hard to monitor), but instead by the amount it contributes to the fund. Unless everybody contributes, no one contributes. Later, if sanctions turn out to be necessary, enforcers have fewer incentives to defect and free ride because their cost of sanctioning is fully reimbursed from the fund. Moreover, coordinating the sanctions through a centralized fund makes it possible to allocate the enforcement burden in the most efficient way. For instance, if the cost of imposing sanctions to one Enforcer is particularly high, this Enforcer does not need to participate in the actual sanctioning and can instead shoulder the burden by paying more to the fund.

We recognize the practical difficulties involved in implementing reversible rewards in a situation involving multiple Violators and multiple Enforcers. The moral hazard problem looms particularly severe because a reward-based mechanism could induce states to portray themselves as Violators and condition their compliance on receiving the reward. Even "genuine" Violators like China might raise their emissions in an effort to ratchet up the reward that the coalition of Enforcers would offer for their

compliance. To mitigate the moral hazard problem, the Enforcers would have to find the way to limit the rewards to genuine Violators. These would be states that stand to be net losers under the climate treaty or states that are economically dependent on outside funding to comply with the treaty. Or simply, Enforcers could offer a reversible reward only to China. Most other countries either emit too little to justify enforcement action or can credibly be deterred by a mere threat of sanctions.

5.3. Private Disputes

Reversible rewards can be used outside the formal legal system to foster private ordering. For instance, they may be used to deter small nuisances that are too costly to enforce. Neighborly grievances—whether they concern a neighbor building an ugly fence or refusing to cut a tree that hangs over the neighboring property—can be difficult to enforce. Relative to the harm they cause, they are often too costly to stop through formal enforcement channels.

Imagine that your neighbor is dumping his garbage on your lawn every week. He is doing it to save a garbage removal cost of $100. You could offer to pay for his garbage removal ($100), but that may seem too costly and unfair. Alternatively, you could hire someone to dump the garbage back at the neighbor's lot, but this too would cost you at least $100. A reversible rewards would enable you to terminate your neighbor's practice at half the cost compared with using sticks or carrots alone. You could set aside $50 and offer it as a reward to the neighbor for ceasing the dumping. If dumping continues, the entire fund would be paid to a garbage service to dump (half of) the garbage back on your neighbor's lawn. As a result, the $50 in the fund works twice: once as a (half) carrot and once as a (half) stick. If the neighbor continues to dump, he will lose twice—the $50 reward, followed by the retaliation measures that $50 can buy. A compliance that would otherwise cost $100 to secure would now cost merely $50.

A similar approach can be used to address contract breach. A contracting party may want to prevent a harmful breach, but cannot do so by merely threatening to sue for damages because the cost of securing a remedy is high (e.g., the cost of litigation), or the remedy would not fully compensate the injured party (e.g., the Enforcer could not obtain consequential damages).

Faced with an impending breach, the Enforcer can instead offer a bonus for adequate performance (the bonus being offered above the already agreed upon price). But if the bonus is structured as a reversible reward, it could be half the size of the otherwise-needed incentive payment The bonus would be offered with the clear message that if it is turned down and the breach occurs, the funds reserved for the bonus would instead be used to finance the cost of litigation.

The reversible reward scheme is also applicable when a party is trying to induce another party to work harder—an employer asking an employee to exert greater effort. In fact, workers are already subject to combined bonus and sanction policies. We could also imagine that employers use reversible rewards in negotiating with labor unions by offering them a small pay rise coupled with the threat of reversing the earmarked funds to pay for their costs of enduring a potential strike by the organized labor.

5.4. Foreclosure

Carrying out foreclosure of mortgaged property or eviction of residential rental property is expensive for mortgage lenders and for landlords. It is time consuming, legally costly, often harmful to the property. Plus, it is simply unpleasant. Indeed, properties coming out of foreclosure sell for a substantial discount, sometimes exceeding 25% of the value of the property (Taub, 2010).

To avoid the costs of evicting the defaulting homeowners through litigation and foreclosure proceedings, creditors can instead offer a reward for those who depart voluntarily and swiftly. But as long as homeowners can impose substantial costs on the creditors by refusing to vacate the property, the reward necessary to buy their compliance might be substantial. Instead, then, creditors could use reversible rewards. Money would be placed in a fund specifically aimed at foreclosing a particular property. It would be offered as reward to a homeowner that voluntary vacates the property in good condition. If the homeowner fails to vacate the premises, the funds would be used to cover the costs of the evictions. In fact, the foreclosure reversible reward fund could be established ex ante and financed by the homeowner. A condition to securing a mortgage could be a contribution by the borrower into a fund that would remain untouched until some fraction of

the mortgage is paid off or until default occurs. If the mortgage is paid off, the money would be returned to the homeowner. If, instead, default occurs, the fund is immediately withdrawn and offered as reward to the homeowner for immediate departure, else used by the creditor to cover foreclosure costs. A reversible reward fund structured this way would convert the enforcement task from a generic one (rewarding all homeowners for vacating their properties voluntarily) into a specific one (each fund being limited to a single homeowner).

5.5. Whistleblower Rewards

In general, reversible rewards are not useful in criminal law because of the problem of multiple, generic violations. It is better to commit funds to sanctions than to pay people for not offending. However, in some specific scenarios, law enforcement is already bolstered by rewards and can be improved if the rewards were reversible. Examples of the use of rewards in criminal investigations involve the whistleblower rewards in connection with securities fraud or cartel investigations. The Dodd-Frank Act, for instance, authorizes the Securities and Exchange Commission (SEC) to use substantial cash rewards to whistleblowers that voluntarily provide the SEC with information that allows it to successfully prosecute securities law violations. The reward is financed by the monetary sanctions the SEC recovers through (civil or criminal) proceedings involving violations of securities laws[12] Antitrust enforcement against cartel activity is similarly bolstered by encouraging self-reporting: those who report their cartel activity and cooperate with the antitrust authorities can obtain immunity and, in some places, are even offered rewards (Office of Fair Trading).

Reversible whistleblower rewards increase deterrence more than simple sanctions, because they harness the insider information to increase detection (in the same way that self-reporting makes detection cheaper) (Kaplow and Shavell, 1994). Members of the cartel would be more likely to defect from the cartel when cooperation with the antitrust authorities would yield them a dual benefit: not only would they escape the penalty, but they would also be entitled to a reward. Linking the funding of the whistleblower reward and

12. Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010, Pub. L. No. 111–203, § 922(a), 124 Stat. 1376, 1841 (2010).

the expenditures of government resources to the investigation of a given suspected cartel or industry would allow the antitrust authorities to offer a smaller reward without diluting the dual effect of its enforcement regime.

6. Concluding Remarks

This paper has demonstrated a novel way in which rewards and sanctions can be combined to reduce private enforcement costs. The idea is to earmark funds that a private party can subsequently use to purchase either a carrot or a stick. By pre-committing the enforcement funds, a reward can be reinforced with a (subsequently costless) threat of sanction. This enforcement scheme doubles the deterrent effect on the Violator.

Reversible rewards can be used to improve compliance in a socially desirable way, for example, by enticing private parties to perform their contracts or countries to pursue efforts to halt climate change. But reversible rewards can also be used in socially harmful ways. For example, a dominant firm seeking to reduce competition can try to intimidate its competitors. It can use rewards (e.g., bribes to competitors to leave a market) or sanctions (e.g. predatory pricing). But since both strategies are costly, a reversible reward could induce the competitor to acquiesce where it otherwise would not and thus allow the dominant firm to capture the market at a smaller cost. Accordingly, the analysis in this paper has no general normative tone. It merely develops a scheme to reduce the costs of private enforcement.

Why have reversible rewards not been used in practice? One reason is their limitation to specific as opposed to generic violations. The idea that sanctions are, indeed, superior as long as the Enforcer can credibly commit to inflicting them is also deeply entrenched in scholarly and public discourse. Most efforts to enhance compliance have therefore been directed at searching for ways to bolster sanctions' credibility. Yet the inability to credibly commit to sanctioning continues to undermine enforcement schemes across numerous areas of private and international law. This limit of simple sanctions has provided the motivation for this paper.

Elements of the reversible rewards scheme are familiar from bounty arrangements. Defendants who post bail are deterred from fleeing in two ways: first, a fleeing defendant loses the bail money; and second, the money that he posted and forfeited can be used to fund bounty hunters, which

increases the likelihood that the defendant will be apprehended. These types of dual incentive schemes are used in many contexts. Internet sites like Facebook offer rewards to hackers who report security flaws in the website, but at the same time pursue enforcement against hackers who exploit such flaws. Governments use rewards to induce informants to report offenses, and punish those that engage in harmful activities and fail to cooperate with the authorities.

But none of these existing schemes combines the pre-commitment and the double effect features in the way Reversible Rewards do. Thus, the idea of reversing the rewards remains unexploited despite its potential to contribute to more credible and less costly enforcement of private rights.

References

Ayres, Ian. 2010. *Carrots and Sticks: Unlock the Power of Incentives to Get Things Done*. New York, NY: Bantam Books.

Bar-Gill, Oren, and Omri Ben-Shahar. 2009. "The Prisoners' (Plea Bargain) Dilemma," 1 *Journal of Legal Analysis* 737.

Bebchuk, Lucian A. 1988. "Suing Solely to Extract a Settlement Offer," 17 *Journal of Legal Studies* 437.

Bebchuk, Lucian A. 1996. "A New Theory Concerning the Credibility and Success of Threats to Sue," 25 *Journal of Legal Studies* 1.

Bebchuk, Lucian A., and Andrew T. Guzman. 1996. "How Would You Like to Pay for That? The Strategic Effects of Fee Arrangements on Settlement Negotiations," 1 *Harvard Negotiation Law Review* 53.

Bradford, Anu, and Omri Ben-Shahar. 2011. "Efficient Enforcement in International Law," 12 *Chicago International Law Journal* 375.

Croson, David, and Robert Mnookin. 1996. "Scaling the Stonewall: Retaining Lawyers to Bolster Credibility," 1 *Harvard Negotiation Law Review* 65.

Dari-Mattiacci, Giuseppe. 2009. "Negative Liability," 38 *Journal of Legal Studies* 21.

Dari-Mattiacci, Giuseppe, and Gerrit De Geest. 2010. "Carrots, Sticks, and the Multiplication Effect," 26 *Journal of Law, Economics, & Organization* 365.

Davis, Kevin E. 2006. "The Demand for Immutable Contracts: Another Look at the Law and Economics of Contract Modification and Renegotiation," 81 *New York University Law Review* 487.

Jolls, Christine. 1997. "Contracts as Bilateral Commitments: A New Perspective on Contract Modification," 26 *Journal of Legal Studies* 203.

Kaplow, Louis, and Steven Shavell. 1994. "Optimal Law Enforcement with Self-Reporting of Behavior," 102 *Journal of Political Economy* 583.

Levi, Michael. 2009. "Copenhagen's Inconvenient Truth," *Financial Times*, September/October.

Levmore, Saul. 1986. "Waiting for Rescue: An Essay on the Evolution and Incentive Structure of the Law of Affirmative Obligations, 72 *Virginia Law Review* 879.

Levmore, Saul. 2000. "Carrots and Torts," in Eric Posner ed., *Chicago Lectures in Law and Economics* 203, 221. New York, NY: Foundation Press.

Office of Fair Trading. 2012. "Rewards for Information About Cartels." http://www.oft.gov.uk/OFTwork/competition-act-and-cartels/cartels/rewards (accessed October 10, 2012).

Posner, Eric A., and Alan O. Sykes. 2011. "Efficient Breach of International Law: Optimal Remedies, 'Legalized Noncompliance,' and Related Issues," 110 *Michigan Law Review* 243.

Taub, Daniel. 2010. "Foreclosed Home in the U.S. Selling at 28% Discount." Bloomberg.com. January 26, 2010, http://www.bloomberg.com/apps/news?pid=newsarchive&sid=acYOhFiTDKsc (accessed October 10, 2012).

Wittman, Donald A. 1984. "Liability for Harm or Restitution of Benefit?," 13 *Journal of Legal Studies* 57.