2011

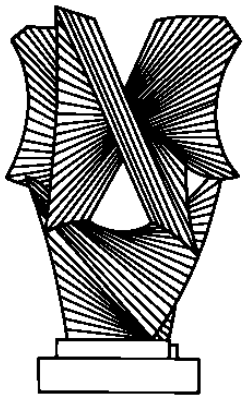
Reversible Rewards

Omri Ben-Shahar
omri@uchicago.edu

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

CHICAGO

JOHN M. OLIN LAW & ECONOMICS WORKING PAPER NO. 557
(2D SERIES)



Reversible Rewards

Omri Ben-Shahar and Anu Bradford

THE LAW SCHOOL
THE UNIVERSITY OF CHICAGO

Revised November 2011

Electronic copy available at: http://ssrn.com/abstract=1864043

REVERSIBLE REWARDS

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

Abstract

This article offers a new mechanism of law enforcement, combining sanctions and rewards into a scheme of "reversible rewards." A law enforcer 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 enforcers to stop violations that would otherwise be too costly to deter. It argues that reversible rewards could be used to bolster enforcement in selective areas of private and public law, and could also be applied strategically in litigation.

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

Electronic copy available at: http://ssrn.com/abstract=1864043

INTRODUCTION

There are two general ways to induce people into action. One is to reward them for the desired behavior; the other is to punish them for the undesired behavior. The typical normative inquiry in the law enforcement literature focuses on the carrot versus stick selection and asks when one device is more effective than the other.[1] A basic insight advanced in this literature is that sanctions are superior to rewards when they are credible: a credible threat of sanctions does not need to be carried out and thus costs nothing. An effective reward, by 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 to advance the normative inquiry on efficient enforcement. 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. We show that reversible rewards can induce compliance in situations where simple sanctions or simple rewards fail.

---

[1] Donald A. Wittman, *Liability for Harm or Restitution of Benefit?*, 13 J. Legal Stud. 57 (1984); Saul Levmore, Saul, *Waiting for Rescue: An Essay on the Evolution and Incentive Structure of the Law of Affirmative Obligations*, 72 Virg. L. Rev. 879 (1986); Giuseppe Dari-Mattiacci, *Negative Liability*, 38 J. Legal Stud. 21 (2009); Giuseppe Dari-Mattiacci and Gerrit De Geest. *Carrots, Sticks, and the Multiplication Effect*, 26 J. L., Econ. & Org. 365 (2010).

Electronic copy available at: http://ssrn.com/abstract=1864043

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: to pay someone to punish 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 it is too costly. Sanctions could be too expensive for the Enforcer relative to the harm they save. Similarly, rewards needed to buy off the Violator's compliance might be too high. In these scenarios, where ordinary enforcement falls short, 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 enforcement schemes suggest.

To illustrate intuitively the double effect of reversible rewards, imagine a party 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 original amount of money into it, and offering candidate A the choice between taking this sum or 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 the rival gains 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, 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 law enforcement: an Enforcer 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. Reversible rewards could also be used in litigation: a defendant could offer a plaintiff a smaller settlement, coupled with the threat that declining it would lead to trial and attrition. Reversible rewards may similarly be used to advance regulation: the government could offer rewards for regulatory compliance, backed up by sanctions for violations, and interlink the two through a pre-committed fund. Governments could also resort to reversible rewards to enhance compliance with international law, where ordinary enforcement methods are limited, threats of sanctions are rarely credible and rewards are often too costly to fund.

We use a simple framework to demonstrate the potential advantage of reversible rewards over simple sanctions or simple rewards. In that framework we show when and why reversible rewards are cheaper instruments to generate compliance. However, we also recognize that the practical application of this mechanism is limited. Simple sanctions are always superior to reversible rewards when a threat of sanctions is credible and thus need not be carried out. We also note that when numerous different violations can potentially occur, a simple sanction scheme normally dominates reversible rewards, as long as the

mere threat of sanctions deter and need not be imposed in every case. 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.

The concept of reversible rewards contributes to the rewards-versus-sanctions inquiry. It is also closely related to the literature on credible enforcement. This literature has wrestled with the question how legal rights might 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 overcome the credibility problem.[2] 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 enforcement in areas where the reach of ordinary sanctions and rewards is limited.

I. AN ILLUSTRATION OF THE ARGUMENT

Party A engages in an activity that is harmful to party B, and 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, and the challenge is to devise an enforcement scheme that stops the

[2] Lucian A. Bebchuk, *Suing Solely to Extract a Settlement Offer*, 17 J. Legal Stud. 437 (1988) (uncertainty about costs); Lucian A. Bebchuk, *A New Theory Concerning the Credibility and Success of Threats to Sue*, 25 J. Legal Stud. 1 (1996) (divisibility of costs); Oren Bar-Gill and Omri Ben-Shahar, *The Prisoners' (Plea Bargain) Dilemma*, 1 J. Legal. Anal. 737 (2009) (sequential enforcement).

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 1.5*s*. For example, to inflict a loss of $100 on the Violator the Enforcer would have to bear a cost of $150, and so on. Alternatively, the Enforcer can offer a reward *r*, and let us assume that the cost of such reward is *r*. Namely, sanctions cost the Enforcer more than the pain they inflict on the Violator, whereas rewards consist of simple transfers of cash.

A. 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. If the Enforcer derives no utility from retaliation, it would have no incentive to inflict a sanction if its cost is greater than 0. Alternatively, it can be assumed that the sanction has an incapacitating effect, forcing the Violator to cease the harmful activity. Then, it would be rational for the Enforcer to impose it only 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.[3]

B. 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. Assuming, for the moment, that the Enforcer has greater bargaining power, it can offer a reward of slightly more than $80 in return for the Violator ceasing the violation. In reality, of course, various factors might make a reward bargain costly to achieve, or might enable the Violator to extract a higher payoff. What this example shows, then, is *not* that a reward would necessarily succeed. All it shows is that, under perfect conditions, a successful reward can cease the violation at the minimum cost of $80 to the Enforcer.

The question we are interested in is whether the Enforcer can do better. Can it induce compliance without having to spend this much money in a reward?

C. Reversible Rewards

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

[3] In much of the analysis below, we will assume that sanctions are merely retaliatory. This assumption will help us sharpen the insight that even when sanctions are least credible, the reversible reward scheme can use them to create stronger deterrence.

three simple elements:

(1) The Enforcer deposits money in an irrevocable fund, which can 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 and the Enforcer decides to punish the Violator in retaliation, 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. Since the cost of inflicting the sanction would be reimbursed to the Enforcer from the fund, we say that the reward is reversible *i.e.,* it can be diverted into a sanction after being rejected by the Violator. Note that if the Enforcer fails to punish an ongoing violation, the money would be squandered and may not be recovered by the Enforcer. That the Enforcer can only recover this money as a reimbursement for a sanction 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 minimum effective non-reversible reward. Under some assumptions, it is enough to deposit $48 in the fund to stop the violation. 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 maximum sanction that the Enforcer would have an incentive to inflict is $s = 32$. This sanction would cost the Enforcer $1.5s$ (namely $1.5 \times 32 = 48$), exactly the amount available in the fund. Thus, the threat to inflict a sanction of

$32 would be credible. A lower sanction would also be credible, but the Enforcer gains nothing by lowering the sanction (the marginal cost to him of unit of sanction is $0.) A higher sanction, exceeding $32, would not be fully reimbursed and thus—when the sanction is merely retaliatory—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: a violation, which would entail a net payoff of $48 (that is, a gain of $80 from continuing violation minus a sure sanction of $32); or ceasing the violation and collecting the reward, which would yield an immediate identical payoff of $48. Thus, endowing the fund with a little more than $48 (say, $50) would be enough to make the Violator strictly prefer compliance over violation. A reversible reward of at least $48 can lead to full compliance.

D. 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.

1. *The Double Effect*

A reversible reward uses the same money twice. In the Introduction, 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 witnessing the same resource being redirected to his or her opponent.[4]

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 by doubling the wedge and thus doubling the effect of the Enforcer's fund.

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 to enhance 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 has studied the potential simultaneous use of sanctions and rewards.[5] Levmore identified situations in which laws or contracts reward good behavior and at the same time punish bad behavior. For example, some jurisdictions provide rewards for rescue and penalties for failure to rescue. Or, stores incentivize sales staff by offering commissions for generating high sales and penalties for

[4]Obviously, in equilibrium the money can be used at most once, but because it is factored into the off-equilibrium moves—because parties act in the "shadow" of what this money can do in different scenarios—it has the double effect.

[5] Saul Levmore, *Waiting for Rescue: An Essay o the Evolution of Incentive Structure of the Law of Affirmative Obligations*, 72 Va. L. Rev 879, 891-906 (1986); Saul Levmore, *Carrots and Torts*, in CHICAGO LECTURES IN LAW AND ECONOMICS 203, 221 (Eric Posner, ed., 2000).

generating low sales. Or, construction contracts sometimes include "risk versus reward" terms, with penalties for delayed completion and rewards for ahead-of-schedule completion. The reversible reward mechanism differs from these examples by utilizing the double wedge effect in a specific manner: Not only are rewards and sanctions offered simultaneously, but their funding is linked.

We use the term "double" effect loosely. More precisely, the effect of reversible rewards could be more or less than double the effect of a simple reward, depending on the precise sanctions' cost function. In the example above, we used a cost multiplier of 1.5*s*, which diminished the sanction's effect on the "wedge." Thus, the reward fund of $80 was reduced to somewhat less than half, $48. A multiplier of 2*s*, to take another example, would further mitigate the double effect, increasing the size of the reversible reward to at least $54.[6] If, instead, the sanction involved excess efficiency (a cost multiplier of less than 1, say 0.5*s*), the money used to finance the sanction would have a larger incentive effect than the same money used for a reward. In this case the reversible reward more than doubles the incentive effect and reward fund would have to be less than half the simple reward.[7] 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.[8] The

[6] Such a reversible reward fund could finance a sanction of $27. The Violator would prefer to take the reversible reward of $54 than commit a violation and net $80 - $27 = $53.

[7] When the cost of sanction is 0.5*s*, the minimum necessary fund would be just over $27. Such a fund could finance a sanction of $54. The Violator would prefer to take the reversible reward of $27 than commit a violation and net $80 - $54 = $26. It should be pointed out that when sanctions are cheap, it would more likely be the case that simple sanction (that has incapacitating power) could be used to fully deter the behavior, and reversible rewards would not be needed.

[8] 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 1.25*r*, 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

more effective sanctions are, the smaller the necessary reversible reward. At the same time, when sanctions are more effective, the threat of simple sanctions is likely enough to create deterrence, as we will show below.

2. *Credible Commitment*

Reversible rewards can be used with or without a pre-committed fund. We mentioned the risk versus reward clauses in contracts and bonus versus fine schemes in employment. The particular reversible reward scheme we develop 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 his threat to sanction. This is particularly helpful in situations where the Enforcer would have to inflict the sanction *after* having already suffered the harm, and thus would have no incentive to do so. With the effective cost of sanctioning reduced to zero at the time the sanctions have to be implemented, the Enforcer's threat to inflict even purely retaliatory sanctions becomes credible through this mechanism. Indeed, as we will demonstrate in the next section, the size of the fund that needs to be pre-committed depends on the type of sanctions involved. Pure retaliatory sanctions require a larger pre-commitment than sanctions that also have an incapacitating effect.

The problem of pre-commitment to enforcement has been studied before, and this paper offers no new theory on how to resolve it. As discussed below, contracts, irrevocable trusts, agency, reputation, sunk costs, and various combinations of these mechanisms have been identified as commitment devices. Reversible rewards are not a new commitment

---

and a reward of $36. The Violator would prefer to take the reversible reward of $36 than commit a violation and net $80 - $45 = $35.

device, but rather a method to take advantage of pre-commitment strategies.[9] Thus, the crux of our argument is that, to the extent that pre-commitment is possible, reversible rewards reduce the amount of the necessary pre-committed fund. We will comment below the use of pre-commitments in a simple sanction regime.

II. FORMAL ANALYSIS

A. 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 he 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

[9] Another mechanism that couples a pre-commitment with a double effect is Ian Ayres' StickK concept. See www.stickk.com. See *also* Ian Ayres, CARROTS AND STICKS: UNLOCK THE POWER OF INCENTIVES TO GET THINGS DONE (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.

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.

B. 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 $c(b) \leq h$. 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 $b < h$, namely, 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 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 cheaper cost, and in a greater set of circumstances, than simple rewards.

C. 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. (We will comment below on the legal foundations for this assumption). 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).

The fund makes the Enforcer's threat to inflict sanctions credible in a number of circumstances where a threat to inflict simple sanctions would lack this credibility. 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 $c(s) \leq U$. 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 $s^*(U) = \frac{1}{2}U$. The costlier it is to impose a sanction (*i.e.,* the higher the values for α or β), the lower is the maximum sanction that the Fund can credibly support.

The question we are interested in is the following: 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 Violator is faced with a choice: either to refrain from the 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:

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

which 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 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 $b > \alpha$, the money in the fund is not completely eaten up by the fixed cost and some of it can be used to generate a non-zero sanction. In contrast, when $\alpha > b$, the fixed cost of sanction will deplete the reward fund before any pain can be inflicted on the Violator. With the ability to impose some sanction, the reward necessary is reduced by the magnitude of this sanction.

2. *Cheaper than simple sanctions.* When credible, sanctions can terminate the violation at no cost. Imagine a scenario in which the sanction, while credible, does not deter the violation (for example, because the Violator does not believe it to be credible). The Enforcer would then inflict the sanction and terminate the violation, but suffer a cost of $c(b)$ (assuming that $c(b) < h$). In this scenario, 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 is, the greater the advantage of the reversible reward regime. If $\beta = 0$, there is no advantage to a reversible reward because any necessary increase in the sanction can be imposed at no extra cost. It is only when increasing the sanction is costly that a reward can have a cost-saving effect. This also suggests that reversible rewards could work as an effective enforcement scheme in situations in which simple sanctions fail.

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 the minimum necessary fund is $U = ½ (100+b)$, which equals 150. It offers a reward of 150, backed by a sanction of 50.The reversible reward scheme achieves compliance at a cost of 150, which is 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.

D. 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 divisibility effect.[10]

1. Numerical Example

Return to the example studied in Section I. We assumed the Violator's benefit to be $80, the harm from the activity $100, and the cost of inflicting a sanction *s* 1.5*s.* We noted that a merely retaliatory sanction would never be credible because it would not eliminate the harm of $100 and would simply amount to another expense. Incapacitating sanctions would also lack credibility as the Enforcer would still need to incur a cost of at least $120 to deter 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 his 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 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 his 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 becomes credible at this stage—the Violator now knows that the

[10] For a model of the effect of cost divisibility on the threat to enforce, see Bebchuk, supra note &&.

Enforcer can pay the full $120 to impose a sanction that costs the Violator a disutility of $80—the Violator would be deterred. Thus, the Enforcer manages to stop the Violator's activity by spending only $20 upfront, and never having to actually spend the additional $100. As long as the money in the fund is sunk, the threat to punish becomes credible. Further, the money deposited into the fund remains there – and can continue to work as sanction-credibility enhancement to deter further violations. It does not need to be paid out as a reward.

While the divisibility of simple sanctions can render them cheaper than simple rewards, the enforcement costs can be further lowered if the Enforcer exploits the divisibility feature in setting up a Reversible Reward fund. Here, the money deposited in 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 his activity voluntarily at time 1. In this case, we can show, the Enforcer only needs to deposit $8 in the fund—that is, the cost to the Enforcer is reduced from $20 to $8. Here is why: 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 the Violator a payoff of $8 which is no less than the net payoff of $8 the Violator obtains from continuing the harmful activity (the benefit from activity ($80) – 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 the fund, the remaining cost for the Enforcer would only be $100, and 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.

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 he inflicts the sanction, the Enforcer stops the harm but incurs a cost of $c(s) - U$. If he doesn't, he 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

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

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.[11]

The above scenario exploits the divisibility effect in a situation where the Enforcer

[11] 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 is the enforcement fund to be cheaper than a simple reward.

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

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

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 $b - s^*(U(b))$. 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, and 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.

A caveat is in order. While Reversible Reward fund is smaller than the simple sanction fund, it does have to be paid to a complying Violator, whereas the simple sanction fund remains unused. Enforcers who deal with multiples potential violators might prefer one large simple sanction fund—assuming 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 only be used once, each.

III. THE LIMITS OF REVERSIBLE REWARDS

This section identifies two significant limitations of Reversible Rewards.

A. Generic v. Unique Violations

The last example in Section II 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 can be repeatedly used to deter each of the violators. All 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.[12]

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.

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 possibility to repeatedly 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 only cares about the behavior of one primary Violator. For example, in Section IV we examine an international enforcement application, where the Enforcer only cares about violations by the largest Violator (China), which has the ability to inflict greatest 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

[12] Bar-Gill & Ben-Shahar, supra note 2, discuss the dynamics of such sequential enforcement schemes in the context of resource constrained prosecutors negotiating plea bargaining.

individuals would have the incentive to 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.

B. 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,[13] trust law provides a legal framework to make effective hands-tying commitments.[14]

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 a middleman to administer the reversible reward scheme, relying on this intermediary's reputation to prevent ex-post modifications. Banks, law and accounting firms, arbitrators, even websites (*e.g.,* StickK.com), specialize in providing such commitment service, and sometimes are bound by professional ethics to preserve the original commitment.

---

[13] Christine Jolls, *Contracts as Bilateral Commitments: A New Perspective on Contract Modification*, 26 Journal of Legal Studies 203 (1997).
[14] Kevin E. Davis, *The Demand for Immutable Contracts: Another Look at the Law and Economics of Contract Modification and Renegotiation*, 81 *New York University Law Review* 487 (2006).

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.

IV. 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.

For example, 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, or the remedy would not fully compensate the injured party. Faced with a threat to breach a contract, this party can deposit some money in an irrevocable fund and offer it to its counterparty as a reward for adequate performance (namely, as a bonus above the already agreed upon price). If the counterparty turns down the bonus and breaches the contract, the deposited fund could be used instead to finance the cost of securing a remedy. By using such a reversible bonus, a contracting party can secure full performance of the contractual obligations at a lower cost.

The same is true if a party is trying to prevent a harmful action by another—nuisance, trespass, defamation, or pollution. When the threat to sue for redress is not credible and does not deter, a reversible reward could be a feasible way to secure the right.

The 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. As we mentioned, construction contracts sometimes include a bonus for early completion as well as a fine for late completion. We could also imagine employers use reversible rewards to force a settlement on labor unions by offering them a smaller 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.

In the remainder of this Section we survey some applications in more detail. In all cases, a reversible reward goes beyond the simple co-utilization of rewards and sanctions; it pre-commits a fund to finance sanctions when rewards are turned down, thus making the threat of sanctions more credible.

A. Settlement Bargaining

Reversible rewards can be employed by a defendant to improve its strategic position and secure a more favorable settlement.[15] 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.

[15] Others have noted how fee arrangements with attorney can affect the strategic structure of settlement bargaining. See, e.g., Lucian A. Bebchuk, and Andrew T. Guzman, *How Would You Like to Pay for That? The Strategic Effects of Fee Arrangements on Settlement Negotiations*, 1 Harv. Neg'n L. Rev. 53 (1996); David Croson and Robert Mnookin, *Scaling the Stonewall: Retaining Lawyers to Bolster Credibility*, 1 Harv. Neg'n L. Rev. 65 (1996). Croson and Mnooking 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.

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. Assume that the more the defendant spends on litigation, the costlier it would be for the plaintiff to secure the $100 judgment. For simplicity, assume that if defendant spends any amount C on litigation, the plaintiff would also have to spend and 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 than incur $100+C in litigation. Thus, 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 that 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 need 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 costs rise only by ½C, then the settlement offer would have to be $67.

Practically, for this technique to work, the defendant has to set up the fund in a way that would make it impossible to use the money in any other way 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.

B. International Enforcement

Enforcement problems are particularly challenging in the international context. In the absence of supranational enforcement bodies, states are not able to rely on an objective third party carrying out enforcement on their behalf. States sometimes resort to economic sanctions, and occasionally the use of military force—but these 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 US 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.

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.

The ongoing effort to negotiate a new global climate change treaty is an illustrative example of a complex multilateral enforcement challenge. While all states would benefit from collective efforts to limit their emissions, they also have the incentive to free ride on other states' efforts to protect the climate. 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.[16] The cost of levying effective trade sanctions on economic powers like China is also prohibitive.

In theory, 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 United States—would set up a fund. Instead of endowing it with the full $300 billion that 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 China's 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 reimburse Enforcers for the costs of inflicting sanctions against China. If the sanction consists of a carbon border tax, the fund could compensate adversely

[16] 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. Michael Levi, *Copenhagen's Inconvenient Truth*, FIN. TIMES, Sept./Oct., 2009.

affected domestic parties. 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 if the reversible reward is, say, $150 billion, it would create an inducement that is roughly equal to a simple reward of $300 billion. The choice for a Violator between violation and compliance has a payoff effect of $300 billion.

The primary advantage of the reversible reward is its ability to enhance the credibility and reduce the cost of enforcement. A secondary advantage is in mitigating 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, to the extent that sanctions are costly to administer, coordinating the sanctions through a centralized fund makes it possible for participating states to allocate the enforcement burden in the most efficient way. For instance, if the cost to one Enforcer of imposing sanctions is particularly high, this Enforcer does not need to participate in the actual sanctioning and can instead shoulder the burden by paying more in setting up the reversible reward fund. Finally, as the total cost of enforcement is reduced through reversible rewards, it is likely to be easier to harness a larger coalition of Enforcers to join the enforcement effort in the first place. This could entice more reluctant Enforcers,

including the United States, to join the enforcement effort.

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 when Enforcers attempt to solve a global collective action problem akin to climate change where every country has the incentive to free ride on other states' efforts to curtail emissions. Any reward-based mechanism could attract numerous "frivolous" Violators all conditioning their compliance on receiving the reward. Alternatively, "genuine" Violators like China might raise their GHG emission levels in an effort to ratchet up the magnitude of 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. To start, Enforcers could offer a reversible reward only to China. Most other countries either emit too little GHGs to justify enforcement action or can credibly be deterred by a mere threat of sanctions.

C. Private Disputes

The settlement bargaining example involved an individual litigant employing reversible rewards to obtain a more favorable settlement. But reversible rewards can also 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 $100 cost of garbage removal. 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. However, reversible rewards would enable you to terminate your neighbor's practice at half the cost compared to using sticks or carrots alone. You could set up a reward fund and deposit $50 to this fund. You could then offer the entire fund ($50) as a reward to the neighbor for ceasing the dumping entirely. If dumping continues, the entire fund would be paid to a garbage service to dump some 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. This "double loss" would likely offset the $100 saving in garbage removal service, leading the neighbor to comply at a mere cost of $50.

D. 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, during which the property depreciates at a greater pace and, at worst, may even get destroyed by recalcitrant private users. In these settings the law prohibits self-help, and so eviction requires non-trivial legal costs and private enforcement measures, and can be simply unpleasant. Indeed, properties coming out of foreclosure sell for a substantial discount,

sometimes exceeding 25% of the value of the property.[17] The weakness of enforcement is costly to the creditors.

To avoid the costs of sanctioning 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, 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. Because this money is sunk and can only be accessed by the foreclosing agents, homeowners are more likely to face immediate forceful eviction and would thus be induced to accept the reward instead. As we showed above, it would require a smaller reward to successfully vacate the reluctant homeowner when that reward is reversible.

Further, the 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. Again, a reversible reward fund structured this way would convert the enforcement task

[17] http://www.bloomberg.com/apps/news?pid=newsarchive&sid=acYOhFiTDKsc

from a generic one (rewarding all homeowners for vacating their properties voluntarily) to a specific one (each fund being limited to a single homeowner).

E. 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.[18] 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.[19]

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.[20]) 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

[18] Dodd-Frank Act at § 922(a).
[19] http://www.oft.gov.uk/OFTwork/competition-act-and-cartels/cartels/rewards
[20] Louis Kaplow and Steven Shavell, *Optimal Law Enforcement with Self-Reporting of Behavior*, 102 J. Pol. Econ. 583 (1994).

reward. Linking the funding of the whistleblower reward and 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.

CONCLUDING REMARKS

This paper has demonstrated a novel way in which rewards and sanctions can be combined to reduce enforcement costs. The idea is to initially set aside earmarked funds that can subsequently be used to purchase either a carrot or a stick. By pre-committing the fund, a reward can be reinforced with a (now costless) threat of sanction. This scheme doubles the deterrent effect on the target.

Reversible rewards can be used to improve enforcement in a socially desirable way, for example by enticing corporations to adopt better safety standards, 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. price war). 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.

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 public law. The limits of simple sanctions has provided the motivation for our inquiry.

Since simple sanctions and rewards are used all the time—and sometimes interchangeably—it is puzzling why an enforcement scheme that has the potential to outperform them in certain circumstances is not utilized. Elements of this scheme are familiar from other arrangements. For example, attorneys are sometimes paid a retainer fee—a pre-committed remuneration irrespective of the amount of work invested—which enhances the strategic position of their client vis-à-vis its counterparty. Another example relates to bounty arrangements. Defendants who posts 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. Bounties are used in many other 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 around the world 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 schemes combine 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 law.

Readers with comments should address them to:

Professor Anu Bradford
University of Chicago Law School
1111 East 60th Street
Chicago, IL 60637
abradford@uchicago.edu

Chicago Working Papers in Law and Economics
(Second Series)

For a listing of papers 1–500 please go to Working Papers at http://www.law.uchicago.edu/Lawecon/index.html

501. Saul Levmore, Interest Groups and the Problem with Incrementalism (November 2009)
502. Tom Ginsburg, The Arbitrator as Agent: Why Deferential Review Is Not Always Pro-Arbitration (December 2009)
503. Nuno Garoupa and Tom Ginsburg, Reputation, Information and the Organization of the Judiciary (December 2009)
504. Eric A. Posner and Alan O. Sykes, Economic Foundations of the Law of the Sea (December 2009)
505. Jacob E. Gersen and Anne Joseph O'Connell, Hiding in Plain Sight? Timing and Transparency in the Administrative State (December 2009)
506. Richard A. Epstein, Impermissible Ratemaking in Health-Insurance Reform: Why the Reid Bill is Unconstitutional (December 2009)
507. Tom Ginsburg and Eric A. Posner, Subconstitutionalism (January 2010)
508. Stephen J. Choi, Mitu Gulati, and Eric A. Posner, What Do Federal District Judges Want? An Analysis of Publications, Citations, and Reversals (January 2010)
509. Joseph Isenbergh, The Future of Taxation (January 2010)
510. Lee Epstein, William M. Landes, and Richard A. Posner, Why (and When) Judges Dissent: A Theoretical and Empirical Analysis (January 2010)
511. Tom Ginsburg, James Melton, and Zachary Elkiins, The Endurance of National Constitutions (February 2010)
512. Omri Ben-Shahar and Anu Bradford, The Economics of Climate Enforcement (February 2010)
513. Neta-li E. Gottlieb, Free to Air? Legal Protection for TV Program Formats (February 2010)
514. Omri Ben-Shahar and Eric A. Posner, The Right to Withdraw in Contract Law (March 2010)
515. Richard A. Epstein, Inside the Coasean Firm: Competence as a Random Variable (March 2010)
516. Omri Ben-Shahar and Carl E. Schneider, The Failure of Mandated Disclosure (March 2010)
517. Kenneth W. Dam, The Subprime Crisis and Financial Regulation: International and Comparative Perspectives (March 2010)
518. Lee Anne Fennell, Unbundling Risk (April 2010)
519. Stephen J. Choi, Mitu Gulati, and Eric A. Posner, Judicial Ability and Securities Class Actions (April 2010)
520. Jonathan S. Masur and Jonathan Remy Nash, The Institutional Dynamics of Transition Relief (April 2010)
521. M. Todd Henderson, Implicit Compensation, May 2010
522. Lee Anne Fennell, Possession Puzzles, June 2010
523. Randal C. Picker, Organizing Competition and Cooperation after *American Needle*, June 2010
524. Richard A. Epstein, What Is So Special about Intangible Property? The Case for intelligent Carryovers, August 2010
525. Jonathan S. Masur and Eric A. Posner, Climate Regulation and the Limits of Cost-Benefit Analysis, August 2010
526. Richard A. Epstein, Carbon Dioxide: Our Newest Pollutant, August 2010
527. Richard A. Epstein and F. Scott Kieff, Questioning the Frequency and Wisdom of Compulsory Licensing for Pharmaceutical Patents, August 2010
528. Richard A. Epstein, One Bridge Too Far: Why the Employee Free Choice Act Has, and Should, Fail, August 2010
529. Jonathan Masur, Patent Inflation, August 2010
530. Bernard E. Harcourt and Tracey L. Meares, Randomization and the Fourth Amendment, August 2010
531. Ariel Porat and Avraham Tabbach, Risk of Death, August 2010
532. Randal C. Picker, The Razors-and-Blades Myth(s), September 2010
533. Lior J. Strahilevitz, Pseudonymous Litigation, September 2010
534. Omri Ben Shahar, Damanged for Unlicensed Use, September 2010
535. Bermard E. Harcourt, Risk As a Proxy for Race, September 2010
536. Christopher R. Berry and Jacob E. Gersen, Voters, Non-Voters, and the Implications of Election Timing for Public Policy, September 2010
537. Eric A. Posner, Human Rights, the Laws of War, and Reciprocity, September 2010

538. Lee Anne Fennell, Willpower Taxes, October 2010
539. Christopher R. Berry and Jacob E. Gersen, Agency Design and Distributive Politics, October 2010
540. Eric A. Posner, The Constitution of the Roman Republic: A Political Economy Perspective, November 2010
541. Stephen J. Choi, Mitu Gulati, and Eric A. Posner, Pricing Terms in Sovereign Debt Contracts: A Greek Case Study with Implications for the European Crisis Resolution Mechanism, Novemer 2010
542. Bernard E. Harcourt, Reducint Mass Incarceration: Lessons from the Deinstitutionalization of Mental Hospitals in the 1960s, January 2011
543. Jacob E. Gersen, Designing Agencies, January 201l
544. Bernard E. Harcourt, Making Willing Bodies: Manufacturing Consent among Prisoners and Soldiers, Creating Human Subjects, Patriots, and Everyday Citizens—The University of Chicago Malaria Experiments on Prisoners at Stateville Penitentiary, February 2011
545. Tom Ginsburg and Thomas J. Miles, Empiricism and the Rising Incidence of Coauthorship in Law, February 2011
546. Eric A. Posner and Alan O. Sykes, Efficient Breach of International Law: Optimal Remedies, "Legalized Noncompliance," and Related Issues, March 2011
547 Ariel Porat, Misalignments in Tort Law, March 2011
548. Tom Ginsburg, An Economic Interpretation of the Pastunwalli, March 2011
549. Eduardo Moises Penalver and Lior Strahilevitz, Judicial Takings or Due Process, April 2011
550. Stephen J. Choi, Gurang Mitu Gulati, and Eric A. Posner, The Law and Policy of Judicial Retirement, April 2011
551. Douglas G. Baird, Car Trouble, May 2011
552. Omri Ben-Shahar, Fixing Unfair Contracts, May 2011
553. Saul Levmore and Ariel Porat, Bargaining with Double Jeopardy, May 2011
554. Adam B. Cox and Richard T. Holden, Reconsidering Racial and Partisan Gerrymandering, May 2011
555. David S. Evans, The Antitrust Economics of Free, May 2011
556. Lee Anne Fennell, Property and Precaution, June 2011
557. Omri Ben-Shahar and Anu Bradford, Reversible Rewards, June 2011