Lie Aversion and Self-Reporting in Optimal Law Enforcement

Robert Innes *

March 2016

Abstract

Requirements that individuals or companies self-report violations are common in regulation and law enforcement. This paper studies how violators' aversion to lying affects the design and merit of enforcement regimes that require self-reporting. Because false reports compel violators to bear costs of lies, we find that self-reporting enables greater deterrence of violations at a lower cost of monitoring, even when self-reporting enjoys no economic advantage in the absence of lie aversion. Corollaries to this result are that (1) the presence of lie aversion enhances social welfare and (2) enforcement regimes that elicit more noxious lies when false reports are made - for example, a compulsory vs. voluntary self-reporting feature - are advantageous.

Keywords: Self-Reporting, Law Enforcement, Lie Aversion

JEL Classifications: K42, K32, D23

*Economics Dept., University of California, Merced, CA 95434, email: rinnes@ucmerced.edu.

1 Introduction

A common feature of regulation, law enforcement and tax collection is to require agents to self-report an outcome to the government. For example, regulations protecting food safety, product safety and the environment require companies to report violations of legal standards and face sanctions for a failure to do so. Economic and legal scholars have identified a variety of societal benefits from these types of self-reporting schemes. The literature starts with Kaplow and Shavell (1994) and Malik (1993), who show that self-reporting can economize on enforcement by enabling regulators to forgo inspections of reporting violators. Subsequent work shows that self-reporting can be advantageous in promoting remediation of harm (Livernois and McKenna, 1999; Innes, 1999a, 1999b), preventing costly efforts to avoid apprehension (Innes, 2001a), encouraging self-auditing (Friesen, 2006; Innes, 2001b), enabling private enforcement with cost-effective citizen suits (Langpap, 2008), better tailoring fines to agents with different probabilities of apprehension (Innes, 2000), and eliciting reports in criminal teams (e.g., Motta and Polo, 2003; Buccirossi and Spagnolo, 2006; Aubert et al., 2006; Fees and Walzl, 2004). [1]

In this literature, an optimal design of self-reporting enforcement accounts for agents' incentives to truthfully report or not report a violation to the government. An agent only reports correctly if the expected cost of doing so is no greater than the expected cost of falsely reporting "no violation" and facing probabilistic inspection and fines from government officials. However, this weighing of expected costs does not account for a feature of individual preferences that recent work has shown to be pervasive in societies around the world, namely, an individual aversion to lies. Building on initial work of Gneezy (2005), a large literature now documents the presence of lie aversion and a variety of forces that drive the extent of lie aversion (see the survey by Rosenbaum et al., 2014).[2] Recent papers show that there is a

[1] Kaplow and Shavell (1994) show, in addition, that self-reporting can save costs of prison sanctions and improve risk sharing when agents are risk averse. Friesen (2006) challenges merits of self-reporting inducements in environmental law enforcement when harm is small and self-auditing costs are high. There is also a vibrant empirical and experimental literature on self-reporting and self-enforcement; for some recent examples (with apologies for omissions) see Bigoni et al. (2012), Friesen and Gangadharan (2013), Toffel and Short (2011), Pfaff and Sanchirico (2004), Short and Toffel (2008), Stafford (2005, 2007), Guerrero and Innes (2013).

[2] For example, recent studies show that lying aversion is affected by the consequences for both sides of the interaction (Gneezy, 2005; Gibson et al., 2013), social cues on how often others lie (Innes and Mitra, 2013),

particularly strong aversion to lies in the field vs. the lab (Abeler et al., 2014) and that the extent of lie aversion varies across individuals (Gibson et al., 2013).

In this paper, I study how an aversion to lies affects both the societal benefits from self-reporting schemes and the optimal design of these programs. Self-reporting is shown to enhance social welfare when agents have heterogeneous lie averse preferences even though (i) self-reporting enjoys no advantage when there is no lie aversion, (ii) lie aversion costs are typically borne by some violators/liars in the optimal regime, and (iii) the aversion costs of lies count in measuring social welfare. A corollary to this result is that the presence of lie aversion, vs. no lie aversion, enhances welfare by lowering costs of enforcement; that is, lie aversion is an advantageous trait for society, even though it creates costs of lies. A related conclusion is that an *increase* in lie aversion generally enhances social welfare. For example, if self-reporting is compulsory rather than voluntary, then agents must explicitly lie in order to avoid the report of harm, rather simply withhold a report. The explicit lie, vs. the smaller "withholding" lie, produces a greater lie aversion cost (Friesen and Gangadharan, 2013) and thereby enables greater enforcement economies.[3] The presence of lie aversion also affects the optimal choice of enforcement effort and the extent to which harm is deterred. On both counts, I find that lie aversion is salutary, saving enforcement resources even while producing greater deterrence on average.

To illustrate the forces that lie aversion brings to bear, consider a simple illustration. Suppose that a violation causes harm of 100, that the government discovers a violation with 50 percent probability, and that a maximal fine of 150 can be levied when a violation is uncovered. In addition, let us suppose that potential violators are of three types, "high lie

gender (Dreber and Johanneson, 2008; Friesen and Gangadharan, 2012), the extent of the lie (Lundquist et al., 2009; Fischbacher and Heusi, 2013), team incentives (Conrads et al., 2013), and cooperation in prior play (Ellingsen et al., 2009). Lie aversion has been found in experiments conducted in the U.S., Australia, India, Israel, and several European countries.

[3]Friesen and Gangadharan (FG, 2013) point to two potentially competing effects of compulsory vs. voluntary reporting on lie aversion. The first is to increase lie aversion by requiring a stronger lie (as argued here). The second is based on work showing that economic incentives can sometimes crowd out intrinsic motives for pro-social behavior (see, for example, Gneezy and Rustichini, 2000). FG suggest that a compulsory (vs. voluntary) regime might also deplete intrinsic motives for truthfulness. In their experiments, FG find that compulsory vs. voluntary self-reporting enforcement regimes lead to more truthful behavior; perhaps because both compulsory and voluntary regimes "price" the truth, the net effect of a compulsory approach is to increase their subjects' lie aversion.

aversion" with a monetary equivalent cost of a lie equal to 30, "middle lie aversion" with monetary equivalent cost of lies equal to 11, and zero ("low") lie aversion. Without self-reporting, all agents face an expected sanction of 75 (150 X 0.50) and are under-deterred because the average fine is less than true harm (100). Now suppose that a compulsory self-reporting regime is introduced with a self-reporting sanction equal to 85. The "high" and "middle" aversion types then truthfully self-report because the alternative (falsely reporting "no violation") leads to an equivalent cost of 75 (the average sanction) plus the respective lie aversion penalties of 30 and 11, for total costs of 105 and 86, both higher than the self-reporting alternative (85). The zero/low aversion types do not truthfully report and face the same cost as in the absence of a self-reporting scheme, 75. For the high and middle types, deterrence is improved, because the agent's cost of a violation (85 with self-reporting vs. 75 without) is closer to true harm. Moreover, no lie aversion costs are actually borne and welfare is, as a result, strictly improved.

Suppose instead that the self-reporting fine is raised to equal true harm, 100. The "high" types still self-report truthfully (and bear no lie aversion costs), but the "middle" types now face a lower total cost with a false report (86) than with a truthful report (100). Deterrence is improved for the "high" types who now face true harm. However, for the "middle" types, deterrence is marginally improved relative to the lower self-reporting fine (from 85 to 86) but at the cost of the resulting lie, 11. If lie aversion costs are considered illicit (so they don't count in social welfare), then the latter cost is immaterial and welfare is strictly improved due to the gain in deterrence for the "high" types. However, if aversion costs "count," then the question is whether the deterrence benefit of the higher self-reporting fine exceeds the resulting cost of lies to the "middle" types. In a more general model (studied below), this tradeoff produces an interior choice of self-reporting sanction, between the average sanction for a "false reporter" and true harm.

Beyond optimality of self-reporting, one key implication of this logic is that, with heterogeneous lie aversion, an optimal enforcement regime typically elicits truthful self-reporting by some and false reporting (of no violation) by others. A few other studies also find motive for this outcome. With heterogeneous probabilities of apprehension, those with high apprehen-

sion risk self-report and those with low risk do not (Innes, 2000; Fees and Walzl, 2006). When the regulator cannot commit to ex-post enforcement, optimal regimes must produce incentives for enforcement by leaving some "guilty" agents who have not self-reported (Gerlach, 2013). Heterogeneous lie aversion provides a distinct motive for heterogeneous responses to self-reporting incentives, which in turn may serve to relax the constraints on enforcement created by the lack of an ex-ante commitment capability (Gerlach, 2013).

While I develop the implications of lie aversion for self-reporting enforcement in the simplest possible setting, the forces at play also have implications for more complicated environments. For example, an interesting paper by Burlando and Motta (2016) shows that a regime of self-reporting - that legalizes and taxes, vs. enforces and fines - combats corruption because taxpayers/self-reporters are no longer in the bribery market. An individual aversion to misreports, as modeled in the present paper, gives rise to an added advantage of the self-reporting regime of Burlando and Motta (2016) by increasing incentives for self-reporting and withdrawing from corrupt exchanges.

This paper's results on the salutary effects of lie aversion relate generally to Kaplow and Shavell's (2007) description of optimal moral sentiments.[4] Kaplow and Shavell (2007) show how guilt and virtue can be inculcated to promote virtuous acts, and deter harmful acts, thereby economizing on ex-post regulatory and subsidy mechanisms that otherwise encourage these choices. In their treatment, guilt and virtue affect the evaluation of harm-creating or -avoiding actions - in our case, the decision to create harm or to take care to prevent harm. While the spirit of Kaplow and Shavell (2007) on beneficial regulatory consequences of morality surely applies here, the present paper is distinquished by its focus on a prevalent moral sentiment (lie aversion) that concerns communication, rather than harm creation.

Section 2 below describes the model and Section 3 shows that self-reporting is optimal in the presence of lie aversion. Sections 4 and 5 characterize optimal enforcement policy in our baseline model (when costs of lies count) and alternative models (when costs of lies are illicit). Section 6 considers welfare benefits of lie aversion and Section 7 concludes.

[4]See also Shavell (2002), Baron (2010) (who studies implications of altruism for self-regulation) and Benabou and Tirole (2006) (who study how a menu of well-documented social preferences affect pro-social behavior under alternate incentive regimes).

2 The Model

A risk neutral agent engages in an activity that can cause external harm h with an endogenous probability p. The event that harm is caused will be called an "accident" or a "violation." The probability of harm depends upon the agent's choice of care c, $p = p(c)$, where higher care leads to a lower risk of harm, $p' < 0$, at a diminishing rate, $p'' > 0$. [5] Care c is measured in monetary-equivalent cost, and is unobservable to the regulator.

If an accident produces an expected sanction (fee) of F to the agent (and no fee is charged when an accident does not occur), the agent chooses care to minimize his/her costs,

$$c^*(F) = \text{argmin } [p(c)F + c] \tag{1}$$

The First Best. If accidents/harm could be costlessly observed and freely sanctioned, then the agent could be charged a fee equal to harm, $F = h$, when an accident occurs. A first-best would then be achieved, $c^{**} = c^*(h) = \text{argmin } p(c)h + c$.

Sanctions and Enforcement Without Self-Reporting. Although the agent observes an accident when it occurs, the regulator can only detect an accident by engaging in costly monitoring. The regulator chooses the probability with which an accident is detected, $q \in [0, 1]$, at cost $m(q)$, where $m' > 0$ and $m'' \geq 0$.

Agents can be sanctioned at any level $s \in [0, \overline{s}]$. Negative sanctions (subsidies) are precluded, perhaps due to budgetary constraints or to avoid entry into accident-causing enterprises. Liability limits or other statutory constraints place an upper bound on sanctions, $\overline{s} \geq h > 0$. $\overline{s}$ may be thought of as available assets and, for simplicity, is assumed not to directly drive inefficient sanctions ($\overline{s} \geq h$).

By the logic of Becker (1968), an optimal (no self reporting) enforcement regime sanctions maximally when an accident occurs ($s = \overline{s}$) and minimally when an accident does not occur ($s = 0$) in order to provide any given incentive for care at minimal monitoring cost $m(q)$.

[5] To ensure an interior optimum, we make the standard assumptions, $|p'(0)| \approx \infty$ and $|p'(\overline{c})| \approx 0$ for $\overline{c} > 0$. In addition, we assume $p''' \geq 0$ (sufficient but not necessary for $c^{*\prime\prime} < 0$).

Optimal monitoring in turn solves the societal cost minimization problem,

$$\min_{q} [p(c)h + c + m(q)] \text{ s.t. } c = c^*(q\overline{s}) \tag{2}$$

$$\rightarrow q_{NSR}\text{: } [p'(c^*())h + 1]c^{*\prime}(q\overline{s})\overline{s} + m'(q) = 0 \tag{3}$$

By our assumptions, condition (3) defines a unique interior monitoring optimum, $q_{NSR} \in (0, 1)$, that underdeters accidents relative to a first-best:[6]

$$p'(c^*())h + 1 < 0 \Leftrightarrow q\overline{s} < h. \tag{4}$$

In choosing q, the regulator trades off benefits of greater expected sanctions in increasing care and lowering social costs ($p'h+1 < 0$) against associated higher expenditures on enforcement ($m' > 0$).

Lie Aversion and Self-Reporting. The model departs from prior work on self-reporting by considering an agent's possible aversion to lies - an intrinsic aversion to falsely reporting that an accident has not occurred when in fact it has. The agent has a monetary-equivalent disutility of lying equal to a. The parameter a is private information and is drawn from the publicly known density / distribution $g(a)/G(a)$ that has positive support on $[0, \overline{a}]$, $\overline{a} > 0$. Consistent with experimental evidence (Gneezy, 2005; Gibson et al., 2013), agents have heterogeneous aversion to lies, some with no aversion at all ($a = 0$), some with a great deal of aversion ($a = \overline{a}$), and others in between. [7]

Under a self-reporting enforcement regime, the agent can self-report an accident to the regulator and thereby face a pre-specified sanction s_R. A self-reporting regime can be either voluntary or compulsory. Under a voluntary regime, the agent is not obligated to report anything. Under a compulsory regime, the agent must report whether an accident has or has not occurred. Under a voluntary regime, an agent need not directly lie when he/she does not self-report an accident. Under a compulsory regime, an agent must directly lie

[6]The standard result (4) follows from equation (3), $c^{*\prime} > 0$, $\overline{s} > 0$, and $m'(q) > 0$.

[7]The analysis is easily modified to allow for fixed probabilities of no lie aversion ($a = 0$) and complete aversion to lies (agents who never lie), together with those on the continuum, $a \in (0, \overline{a})$. For simplicity, we restrict attention to the continuum.

in order to avoid the report of an accident that has occurred. While "white lies" - failing to report an accident - may lead to some lie aversion, the aversion will be less than for direct lies (see, for example, Friesen and Gangadharan, 2013).[8] Later in the paper, we will consider whether a self-reporting regime that requires stronger lies (compulsory) or weaker lies (voluntary) is more advantageous. For the moment - in order to focus on the broader question of whether and how lie aversion affects the optimal enforcement regime - we consider the case of compulsory self-reporting.

Basics of Self-Reporting. Under a self-reporting regime, the agent will truthfully report an accident when

$$q\overline{s} + a \geq s_R \rightarrow \text{truthfully self report accident} \tag{5}$$

The expected costs of falsely reporting "no accident" are the expected sanction, $q\overline{s}$, plus the incurred aversion to the associated lie, a. When these costs are greater than the self-reporting sanction s_R, the agent will choose the truthful self-reporting strategy, and vice versa:

$$q\overline{s} + a < s_R \rightarrow \text{falsely report no accident} \tag{6}$$

Equations (5)-(6) implies the gross fee for an accident,

$$F_{SR}(a; q, s_R) = min\left(q\overline{s} + a, s_R\right) \tag{7}$$

Without loss, we can restrict attention to self-reporting sanctions in the following interval:

$$s_R \in [q\overline{s}, q\overline{s} + \overline{a}] \tag{8}$$

An optimal enforcement regime will set s_R at or above $q\overline{s}$ because, if not, q can be lowered

[8] A related literature shows that stronger lies (that involve stronger statements) lead to an increased lie aversion on average (Lundquist et al., 2009; Gawn and Innes, 2015). A "white lie" as described here involves a lesser (indeed no) statement of a falsehood. (Note that this meaning for the term "white lie" differs from nomenclature in some related literature; for example, in Gneezy (2005) and Erat and Gneezy (2012) "white lies" are direct lies that are virtuous in the sense that they benefit their recipient.)

and enforcement costs thereby saved without altering agent sanctions or care or accident risk. An s_R sanction at or above the upper bound, $q\overline{s} + \overline{a}$, elicits false reports by all agents, regardless of their aversion costs a, and is therefore equivalent to an s_R sanction equal to the upper bound. Equation (7) implies that there is a critical $a^* \in [0, \overline{a}]$ that partitions agents between those who truthfully report accidents ($a \geq a^*$) and falsely report ($a < a^*$):

$$q\overline{s} + a^* = s_R \tag{9}$$

Enforcement Costs Under Self Reporting. Self-reporting can affect enforcement costs as described in Kaplow and Shavell (1994) and Malik (1993) (KSM). If monitoring is specific to agent activities - as when government inspectors periodically visit regulated facilities - then self-reported accidents/violations need not be inspected. A given probability of monitoring non-reporters (q) can then be achieved with fewer inspections. For conceptual clarity in identifying how lying aversion affects benefits of self-reporting enforcement regimes, we void the KSM effect in our baseline model:

Assumption A1. Enforcement costs are $m(q)$ under all regulatory regimes.

Assumption A1 reflects generalized enforcement, when violations are detected by monitoring of the general environment (Shavell, 1991). For example, when police officers monitor traffic or crime-prone neighborhoods, the probability of discovery is driven by the rotation of officers and not outcomes of self-reporting. Alternately, pollution events may be detected by testing of local air and water vs. facility inspections.

Later in the paper, we will consider alternate models that incorporate the KSM effect:

Assumption A2 (KSM). Enforcement costs are (i) linear in q, with unit cost m (the cost of an inspection) and (ii) reduced by self-reporting of violations that need not be monitored.

Under Assumption A2, total (expected) enforcement costs equal

$$qm[1 - [1 - G(a^*)]p(c^*(s_R))] = \text{KSM enforcement costs under self-reporting,} \tag{10}$$

i.e., unit costs of inspections m, times the probability that non-self-reporters are inspected

q, times the fraction of non-reporters in the population of potential reporters (one minus the fraction of agents who truthfully self-report given an accident, that is, $1 - G(a^*)$ times the probability that a truth-teller has an accident, $p(c^*(s_R))$.

Welfare Under Self-Reporting. There are two alternate ways in which lie aversion costs might be treated in a welfare calculus for enforcement design. In our baseline model, we assume:

Assumption B1. Lie aversion costs, a, count in social welfare (when borne by an agent).

Assumption B1 is a natural baseline for two reasons. First, on a normative level, if a government regulatory policy produces lies that are costly to the agents making them, these costs are relevant to the choice of policy design. Second, if costs of lie aversion are ignored, then eliciting lies provides a "free" sanction to violations: an agent with a violation bears the "lie aversion sanction" a that is costlessly imposed by the government. In our baseline model, we are interested in whether and how lie aversion can be exploited when this advantage is not present.

A competing perspective is that lie aversion costs are akin to benefits of crime (to criminals) and should not be included in a social welfare calculus (see, for example, Lewin and Turnbull, 1990; Stigler, 1970):

Assumption B2. Lie aversion costs, a, are illicit and excluded from social welfare.

In what follows we consider model variations suggested by the competing assumptions:

Baseline Model 1 (M1): Assumptions A1 (generalized enforcement) and B1 (welfare-relevant lie aversion costs).

Model 2 (M2): Assumptions A1 (generalized enforcement) and B2 (illicit lie aversion costs).

Model 3 (M3): Assumptions A2 (KSM enforcement) and B2 (illicit lie aversion costs).

In our baseline model M1, welfare costs include (i) harm from accidents, (ii) agent lie aversion costs when untruthful self-reports are made, (iii) agent costs of care c, and (iv) regulatory enforcement costs $m(q)$. Models M2 and M3 exclude the lie aversion costs (ii),

and model M3 incorporates the revised (KSM) enforcement costs (equation (10) in place of $m(q)$ in (iv)). Because illicit lie aversion costs (Assumption B2) impart large enforcement benefits of lies (in model M2), we consider a revised KSM model M3 (with Assumption A2) that adds enforcement benefits of truths.

3 Logic of Optimal Self-Reporting Under Lie Aversion

A self-reporting (SR) enforcement regime produces identical incentives for care as a "no-self-reporting" (NSR) regime if s_R is set equal to $q\overline{s}$ (at $q = q_{NSR}$ from equation (3)). All agents then self-report and face the same expected sanction as under the NSR regime. Under Assumption A1, the two regimes also produce identical enforcement costs and an identical calculus of enforcement effort (choice of q). Moreover, absent any lie aversion costs, a higher s_R sanction ($s_R > q\overline{s}$) has no effect because no agent will self-report.

Observation 1. If there are no lie aversion costs ($a = 0$ for all agents) and Assumption A1 holds, then SR enforcement has no welfare advantage over NSR enforcement.

With lie aversion costs, does an optimal s_R regime depart from the NSR-mimicing regime, $s_R = q\overline{s}$? For our baseline model M1 in which the lie aversion costs count, consider raising s_R marginally above q (to $s_R = q\overline{s} + \epsilon$). For all remaining self-reporters ($a \geq \epsilon$), deterrence rises with the sanction. Recalling that the initial NSR regime underdeters (equation (4)), this deterrence effect is strictly welfare improving. The cost of the increase in s_R is that it leads low lie aversion agents to lie and bear the associated aversion costs, $a \in [0, \epsilon)$. By construction, however, the latter costs are negligible (as the rise in s_R is marginal, with ϵ arbitrarily small). The increase in s_R thus improves social welfare, meaning that an optimal SR regime sets $s_R > q\overline{s}$ and does better than the NSR-replicating SR regime: [9]

Observation 2. If there are lie aversion costs, then optimal SR enforcement is welfare superior to optimal NSR enforcement.

Self-reporting is advantageous because higher self-reporting sanctions, over and above average sanctions imposed on lying non-reporters, can elicit truthful reports from agents

[9] This logic establishes the welfare superiority of self-reporting in models M1 and M2. For the adapted KS model M3, SR enforcement is optimal even without lie aversion costs; there, self-reporting with $s_R = q_{NSR}\overline{s}$ reduces enforcement costs (vs. no self-reporting) without sacrificing deterrence.

who are sufficiently lie averse. These reports involve no lie aversion costs because they are truthful, but the higher (SR) sanction improves welfare by promoting accident prevention. Although the higher sanction also produces false reports (of no accident), it does so only for those who have low lie aversion costs. An optimal SR regime balances these benefits and costs.

When lie aversion costs are illicit (Assumption B2), benefits of a self-reporting regime are even greater because costs of inducing lies are not relevant to the welfare calculus and all that matters is the deterrence benefit of self-reporting. For this case, Figure 1 depicts the improvement in accident sanctions produced by moving from the optimal NSR regime (with the average sanction $q\overline{s}$) to an SR regime with s_R set equal to true harm h (above $q\overline{s}$ by equation (4)). Because the lie aversion costs borne by the low-a agents do not count (by Assumption B2), the improved incentives are "free" to the regulator and strictly welfare improving.

4 Optimal Self-Reporting in the Baseline Model

Let individual a's aversion to a lie be αa, where the parameter $\alpha(> 0)$ will be useful later and for now is set equal to one, $\alpha = 1$. In model M1, welfare costs under self-reporting enforcement can then be written: [10]

$$W^*(\alpha) = \min_{q,s_R} W(q, s_R, \alpha) = \int_0^{a^*} [p(c^*(q\overline{s} + \alpha a))(h + \alpha a) + c^*(q\overline{s} + \alpha a)]g(a)da$$

$$+ (1 - G(a^*))(p(c^*(s_R))h + c^*(s_R)) + m(q) \tag{11}$$

s.t. a^*: $q\overline{s} + \alpha a^* = s_R$, $s_R \in [q\overline{s}, q\overline{s} + \alpha\overline{a}]$, $q \in [0, 1]$.

The first (integral) term on the right measures welfare costs of accidents (harm h), lies (a), and accident prevention care c for false reporters, those with low lie aversion costs who falsely report "no accident." The second term measures costs of accidents and care for truthful self-reporters who bear the accident sanction s_R. The final term captures enforcement costs $m(q)$.

[10] An additional constraint to problem (11) is that the self-reporting sanction not exceed the bound $\overline{s}$, $s_R \leq \overline{s}$. In what follows, we will see that this constraint does not bind in an optimum.

Differentiating (11) with respect to the SR sanction s_R, we have:

$$\frac{\partial W}{\partial s_R} = \frac{\partial a^*}{\partial s_R} p(c^*(s_R))\alpha a^* g(a^*) + (1 - G(a^*))[p'(c^*(s_R))h + 1]c^{*\prime}(s_R) \tag{12}$$

where $\frac{\partial a^*}{\partial s_R} = \frac{1}{\alpha} = 1 > 0$. The regulator simultaneously sets q optimally, $\frac{\partial W}{\partial q} = 0$. The question is where s_R is set in the interval $[q\overline{s}, q\overline{s} + \alpha\overline{a}]$, the lower end ($s_R = q\overline{s}$) representing the NSR-replicating regime. The bounds for s_R correspond exactly with a "critical agent" of $a^* = 0$ (for s_R at the lower bound, $q\overline{s}$) and $a^* = \overline{a}$ (for s_R at the upper bound). Evaluating (12) at the bounds, we have:

$$\frac{\partial W}{\partial s_R}|_{a^*=0} = [p'(c^*(q\overline{s})h + 1]c^{*\prime}(q\overline{s}) < 0 \tag{13}$$

where the inequality follows from (4), and

$$\frac{\partial W}{\partial s_R}|_{a^*=\overline{a}} = (1/\alpha)p(c^*(s_R))\overline{a}g(\overline{a}) > 0 \tag{14}$$

By (13) and (14), the solution to (11) is interior,

$$q\overline{s} < s_R^* < q\overline{s} + \alpha\overline{a}, \tag{15}$$

with s_R^* solving $\frac{\partial W}{\partial s_R} = 0$. The optimum trades off benefits of a higher s_R sanction in increasing deterrence for self-reporters (the second term on the right of equation (12)) against costs of more lie aversion costs at the margin (the first right-hand term in (12)). An immediate implication of this tradeoff is under-deterrence of self-reporters at the optimum, [11]

$$p'(c^*(s_R^*))h + 1 < 0 \leftrightarrow s_R^* < h. \tag{16}$$

Because "false reporters" face lower average sanctions than truthful self-reporters, $q\overline{s} + \alpha a < s_R^*$, they too are under-deterred. In summary:

Proposition 1A. In an optimal enforcement regime for model M1:

[11] Equation (16) also implies that the optimal self-reporting sanction is lower than $\overline{s}$ (by our initial premise that $\overline{s} \geq h$).

(i) self-reporting is optimal ($s_R^* > q\overline{s}$),

(ii) untruthful reporting of "no accident" occurs for some agents ($a \in [0, a^*)$, $a^* > 0$),

(iii) truthful self-reporting of accidents occurs for agents with high lie aversion costs ($a \in [a^*, \overline{a}]$, $a^* < \overline{a}$), and

(iv) all agents are under-deterred relative to a first-best, $F_{SR} < h$ for all $a \in [0, \overline{a}]$, but truth-tellers are less under-deterred (and exercise more care) than liars, $q\overline{s} + \alpha a < s_R < h \leq h + \alpha a$ for $a \in [0, a^*)$.

Somewhat more subtle (and analytically challenging) is a comparison of the optimal self-reporting (SR) enforcement regime to the optimal regime without a self-reporting feature (NSR). The following is proven in the Appendix:

Proposition 1B. Relative to the optimal NSR regime, the optimal SR regime in Model M1 (i) increases deterrence of the truthful self-reporters, (ii) improves deterrence on average, and (iii) reduces enforcement effort, $q_{SR} < q_{NSR}$.[12]

Self-reporting enables a "free" extra accident deterrent to truthful self-reporters who face higher sanctions and willingly pay them in order to avoid their high lie aversion costs. Ceteris paribus (in particular, fixing accident monitoring effort), lying reporters also face a greater accident deterrent than under enforcement without self-reporting due to costs of lies that they bear in the event of a violation. Deterrence is thereby raised under self-reporting, which in turn reduces the government's marginal incentives to achieve deterrence with costly enforcement effort. Overall, self-reporting therefore improves deterrence while saving enforcement resources.

5 Self-Reporting When Costs of Lies Are Illicit

In model M2, welfare costs differ from M1 only because the lie aversion costs, a, are not added to harm h as a social cost (inside the integral of equation (11)):

[12]Improving deterrence on average is defined in the Appendix as reducing the average deterrent benefit (reduction in welfare cost) that results from a marginal (and uniform) increase in sanction.

$$W^*(\alpha) = \min_{q,s_R} W(q, s_R, \alpha) = \int_0^{a^*} [p(c^*(q\overline{s} + \alpha a))h + c^*(q\overline{s} + \alpha a)]g(a)da$$

$$+(1 - G(a^*))(p(c^*(s_R))h + c^*(s_R)) + m(q) \tag{17}$$

In this case, the prior cost of raising the self-reporting sanction s_R, due to more lying at the margin, no longer "counts." The marginal effect of raising s_R on welfare costs thus reduces to the second term in equation (12), the deterrence effect:

$$\frac{\partial W}{\partial s_R} = (1 - G(a^*))[p'(c*(s_R))h + 1]c^{*\prime}(s_R) \tag{18}$$

At the lower bound for the s_R sanction, $q\overline{s}$, welfare-enhancing benefits of raising the sanction are the same as before (equation (13)). Deterrence is improved at no cost (relative to the under-deterrence of the NSR-replicating regime).

However, the optimal setting for s_R differs from M1 because now only the deterrence effect matters. There are two possible cases: (1) $q\overline{s} + \alpha\overline{a} \leq h$ when $s_R = h$,[13] or alternately, (2) $q\overline{s} + \alpha\overline{a} > h$. In the first case, an optimum stipulates a maximal self-reporting sanction, $s_R = q\overline{s} + \alpha\overline{a} \leq h$, and all agents falsely report "no accident." A self-reporting regime is socially advantageous because the lie aversion costs provide "free" sanctions to accidents.

The second case is more interesting and perhaps more plausible, arising so long as maximal lie aversion is sufficiently high and/or marginal costs of detection ($m'(q)$) are sufficiently modest. In this case, an optimum stipulates a self-reporting sanction equal to harm, $s_R = h$, setting the derivative $\partial W/\partial s_R$ in (18) equal to zero. Unlike case (1), there is an interior partition of agents, $a^* \in (0\,\overline{a})$, where a^* equates the "false reporting" cost, $q\overline{s} + \alpha a^*$, with the SR sanction, $s_R = h$. More lie averse agents ($a \geq a^*$) self-report and choose first-best care. Less lie averse agents ($a < a^*$) falsely report "no accident" and are under-deterred, $c^*(q\overline{s} + \alpha a) < c^*(h)$. In essence, the prescription depicted in Figure 1 is implemented.

Proposition 2. In an optimal enforcement regime for model M2:

(i) a self-reporting feature is optimal ($s_R^* > q\overline{s}$);

[13] To be more precise, the first case inequality is evaluated at the corresponding optimal q with no truthful reporting, with q solving: $\int_0^{\overline{a}} [p'(c^*(q\overline{s} + \alpha a))h + 1]c^{*\prime}()g(a)da + m'(q) = 0$.

(ii) there are two possible cases, one in which the self-reporting sanction s_R is set maximally (case (1), $s_R = q\overline{s} + \alpha\overline{a} \leq h$) and another in which s_R is set equal to harm (case (2), $s_R = h < q\overline{s} + \alpha\overline{a}$);

(iii) in case (2), there is a set of truthful self-reporters of accidents and a set of false reporters; the former choose first-best care ($c = c^*(h)$); the latter are under-deterred relative to a first-best ($c < c^*(h)$); and the optimal self-reporting sanction is higher than in model M1 (where lie aversion costs "count"); [14]

(iv) in case (1), no agents truthfully self-report accidents and all bear an aversion cost of a lie when an accident occurs; and

(v) relative to the optimal NSR regime, the optimal SR regime in Model M2 improves deterrence on average and reduces enforcement effort, $q_{SR} < q_{NSR}$.

The KSM Model M3. Model M2 embeds benefits of lies that are made when an accident occurs: they implicitly punish agents for accidents and thereby provide a "free" accident deterrent. However, unlike M1, this setting imparts no social benefit to truth, because lie aversion costs are considered illicit and excluded from the welfare calculus. There are other potential benefits of truths beyond avoiding lie aversion costs. The enforcement structure of Kaplow and Shavell (KS, 1994) and Malik (1993) (KSM) provides a particularly relevant example. Model M3 embeds this structure. Here, truthful self-reports of accidents enable regulators to more intensively inspect agents who report "no accident" by diverting monitoring investments away from the self-reporters. In an optimum, this benefit of truth is traded off against deterrence benefits of lies. Under plausible conditions, this trade-off returns the optimal regime to an interior solution as in the baseline model.

In M3, self-reporting enforcement is welfare-improving with or without lie aversion costs due to the logic of KSM. By setting $s_R = q\overline{s}$, self-reports are elicited from all agents without sacrificing any deterrence; this strategy reduces enforcement costs because only agents who do not have accidents need to be inspected. The question becomes: Should s_R be set to elicit truth from all agents ($a^* = 0$), no agents ($a^* = \overline{a}$), or some agents ($a \geq a^*$, $a^* \in (0\,\overline{a})$)

[14] In case (2), $s_R^* = h$ in M2 vs. $s_R^* < h$ in M1 (by Propositions 1 and 2).

but not others ($a < a^*$)?

In M3, welfare can be written,

$$W^*(\alpha) = \min_{q,s_R} W(q, s_R, \alpha) = \int_0^{a^*} [p(c^*(q\overline{s} + \alpha a))h + c^*(q\overline{s} + \alpha a)]g(a)da$$

$$+(1 - G(a^*))(p(c^*(s_R))h + c^*(s_R)) + qm[1 - [1 - G(a^*)]p(c^*(s_R))] \tag{19}$$

where a^*: $q\overline{s}+\alpha a^* = s_R$. In (19), we assume that there is some untruthful reporting ($a^* > 0$). The first and second terms capture costs of accidents and care by the untruthful and truthful agents, respectively; the third term measures the KSM enforcement costs. Differentiating with respect to s_R:

$$\frac{\partial W}{\partial s_R} = (1 - G(a^*))[p'(c^*)h + 1]c^{*\prime}(s_R)$$

$$-mq(1 - G(a^*))p'(c^*)c^{*\prime}(s_R) + mqg(a^*)p(c^*)\frac{1}{\alpha} \tag{20}$$

at $c^* = c^*(s_R)$. The first two terms in (20) reflect net deterrence benefits of raising s_R and thereby eliciting higher care from the truthful reporters ($a \geq a^*$), net of an enforcement cost penalty to reducing accident risk for truth-tellers (which reduces the frequency with which they are exempted from inspection). The third term gives the enforcement cost penalty of shifting the marginal a^* agent from truthful to untruthful reporting. When there is no truthful reporting (with $a^* = \overline{a}$), the first two terms vanish and we are left with the third effect:

$$\frac{\partial W}{\partial s_R}|_{a^*=\overline{a}} = mqg(\overline{a})p(c^*(s_R)) > 0 \tag{21}$$

By equation (21), an optimal SR regime sets $s_R < q\overline{s} + \alpha\overline{a}$ and elicits truthful reports from some agents ($a \in [a^* \, \overline{a}]$, $a^* < \overline{a}$).

Evaluating the other bound for s_R, $s_R = q\overline{s}$ (where $a^* = 0$), is more complicated. Elevating s_R above this bound improves deterrence for all truthful reports, that is, all agents at this setting. However, the enforcement cost penalty of shifting the marginal agent ($a^* = 0$) from truthful to untruthful reporting (the third term in (20)) does not vanish. If there is a

sufficiently low density of the marginal agents with no lie aversion ($a = 0$), the latter effect is dominated by the deterrence benefits of a higher s_R sanction:

Condition C1. $(q\overline{s})g(0) < (1-p)/p$ at $p = p(c^*)$ and $c^* = c^*(q\overline{s})$), where

$$q = \text{argmin}\, W^{**}(q) = p(c^*)h + c^* + mq(1 - p(c^*)) \tag{22}$$

Proposition 3. In an optimal enforcement regime for model M3, (i) self-reporting is optimal, (ii) truthful self-reporting occurs for some agents ($a \in [a^*, \overline{a}]$, $a^* < \overline{a}$), (iii) untruthful self-reporting occurs for some agents when condition C1 holds ($a \in [0, a^*)$, $a^* > 0$), and (iv) all agents are under-deterred relative to a first-best, but truth-tellers are less under-deterred (and exercise more care) than liars.

Proof. (ii) By (21). (iii) Evaluating (20) at $a^* = 0$ where $s_R = q\overline{s}$ and substituting from the first order condition for the associated optimal q, $q^{**} = \text{argmin}\ W^{**}(q)$ (from (22)):

$$\frac{\partial W}{\partial s_R}|_{a^*=0} = -(m/\overline{s})[1 - p - p\, q\overline{s}\, g(0)] < 0 \ \text{ under (C1)} \tag{23}$$

By (23), $s_R^* > q\overline{s}$. (iv) If $s_R^* \in (q\overline{s}, q\overline{s} + \alpha\overline{a})$ (e.g., if (C1) holds), (iv) follows from equation (20) ($p'(c)h + 1 < 0$ at $c = c^*(s_R^*)$) and $q\overline{s} + \alpha a < s_R$ for $a < a^*$. If an optimum sets $s_R = q\overline{s}$, then all agents are truth-tellers and

$q = \text{argmin}\, W^{**}(q) : (p'(c^*)h+1)c^{*\prime}\overline{s} + m(1-p(c^*)) - mqp'(c^*)c^{*\prime}\overline{s} = 0 \leftrightarrow p'(c^*)h+1 < 0$ at $c^* = c^*(s_R^*)$.

6 Social Benefits of Lie Aversion and Optimality of Compulsory (vs. Voluntary) Self-Reporting

In models M1-M3, there is an enforcement regime that (a) is optimal when there are no lie aversion costs ($a = 0$ for all agents), (b) is available as a possible enforcement regime when there are lie aversion costs, (c) involves no lies, and (d) produces the same welfare costs with or without the presence of lie aversion. This is the SR regime that elicits truthful reports from all agents, $s_R = q\overline{s}$, with q chosen optimally. If this regime is not optimal under lie

aversion then (by revealed preference) the optimal regime under lie aversion must produce lower welfare costs than the optimal regime under no lie aversion. Therefore, by Propositions 1-3, we have:

Proposition 4. In models M1 and M2, the presence of lie aversion enhances social welfare. In model M3, the presence of lie aversion (a) does not diminish social welfare and (b) enhances social welfare if condition C1 holds.

While the presence of lie aversion (vs. none) can enhance social welfare, does an increase in lie aversion do the same? This question is potentially relevant when moving from a voluntary to a compulsory self-reporting regime. Now, if there is no aversion to the "white lie" of not reporting a violation, then a voluntary self-reporting enforcement regime involves no lie aversion ($\alpha = 0$) and, by Proposition 4, is welfare dominated by a compulsory self-reporting regime under which false reports are associated with positive lie aversion costs ($\alpha = 1$).

The comparison is somewhat more tricky when "white lies" (under voluntary reporting) involve positive lie aversion costs, but lower costs than do "bald lies" (under compulsory reporting). Let us suppose that $\alpha = \alpha_v$ (< 1) under voluntary SR and $\alpha = \alpha_c = 1$ under compulsory SR. Does the increase in lie aversion, from α_v to $\alpha_c = 1 > \alpha_v$, enhance welfare?

The "illicit lying cost" (Assumption B2) models M2 and M3 have the simplest answer to this question, so we begin with them. For model M2, we can differentiate the welfare costs in equation (17):

$$\frac{\partial W}{\partial \alpha} = \int_0^{a^*} [p'(c^*(q\overline{s} + \alpha a))h + 1]c^{*\prime}(q\overline{s} + \alpha a)ag(a)da < 0 \tag{24}$$

Increasing α, by raising lie aversion costs, increases deterrence of the false reporters ($a < a^*$). Because these agents are underdeterred in the optimum (Proposition 2), this effect enhances social welfare (lowering welfare costs, as indicated in (24)).

Similarly for model M3 (with $a^* > 0$), we can differentiate $W^*(\alpha)$ in equation (19):

$$\frac{\partial W}{\partial \alpha} = \int_0^{a^*} [p'(c^*(q\overline{s} + \alpha a))h + 1]c^{*\prime}(q\overline{s} + \alpha a)ag(a)da + q\, m\dot{g}(a^*)\, \frac{\partial a^*}{\partial \alpha}\, p(c^*(s_R)) < 0 \quad (25)$$

The inequality in (25) follows from Proposition 3 ($p'h + 1 < 0$) and $\frac{\partial a^*}{\partial \alpha} = -\frac{a^*}{\alpha} < 0$. Here there is an additional enforcement cost advantage of greater lying aversion: The marginal untruthful reporter shifts to truthful self-reporting, which reduces the number of agents subject to inspections. Of course, this advantage arises only when there are some untruthful reporters in the optimum ($a^* > 0$).

Proposition 5A. An increased level of lying aversion (higher α) enhances social welfare in (i) model M2 and (ii) model M3 when $a^* > 0$.

In model M1, there is an added cost to increased lying aversion because associated cost increases to the lying agents, $a \in [0\, a^*)$, "count." However, under an arguably plausible restriction, the enhanced deterrence benefits of the heightened lie aversion more than offset their added cost. To develop this point, consider starting with the optimal SR regime when $\alpha = \alpha_v$ (under voluntary self-reporting). Figure 2 depicts the change in accident sanctions and lie aversion costs under this regime when α rises to $\alpha_c = 1$. With $s_R < h$ (by Proposition 1), the increase in lie aversion reduces the extent of under-deterrence for all agents that lie under the original voluntary ($\alpha = \alpha_v$) regime, $a \in [0, a^*(\alpha))$. Some agents, $a \in [a^*(1), a^*(\alpha_v)]$, lie under the voluntary regime, but no longer lie under the compulsory regime (where $\alpha = 1$). Other agents, $a \in [0, a^*(1))$, still lie but face a higher accident penalty, $q\overline{s} + a > q\overline{s} + \alpha_v a$. The former agents no longer bear costs of lies, while the latter bear higher costs of lies. If the savings in lie aversion costs for those no longer lying exceeds the cost increase to those who still lie, then overall the increase in lie aversion reduces welfare costs.

Writing down the change in lie aversion costs when moving from $\alpha = \alpha_v < 1$ to $\alpha = 1$, we have

$$\Delta(\alpha) = (1 - \alpha) \int_0^{a^*(1)} ag(a)da - \alpha \int_{a^*(1)}^{a^*(\alpha)} ag(a)da = -\int_\alpha^1 \frac{\partial \Delta(x)}{\partial x} dx \quad (26)$$

where $a^*(\alpha) : q\bar{s} + \alpha a = s_R$ and the second equality follows from $\Delta(1) = 0$. The first term in (26) gives the increased cost of lies to continuing liars and the second gives the cost savings to those who no longer lie. Differentiating Δ:

$$\frac{\partial \Delta(\alpha)}{\partial \alpha} = -\int_0^{a^*(\alpha)} ag(a)da - \alpha(\frac{\partial a^*}{\partial \alpha})a^*(\alpha)g(a^*(\alpha)) = \int_0^{a^*(\alpha)} [2g(a^*(\alpha)) - g(a)]ada \quad (27)$$

Now suppose the following condition holds:

Condition C2. $g(a) \leq 2g(a^*(\alpha_0))$ for $\alpha_0 \in [\alpha_v, 1]$ and $a \in [0, a^*(\alpha_0))$.

C2 requires that the density of low-lie-averse agents ($a < a^*(\alpha)$) does not exhibit large declines; "mid-range" lie averse agents ($a = a^*(\alpha)$) are not excessively scarce relative to agents with lower levels of lie aversion ($a < a^*(\alpha)$). By (26) and (27), C2 is sufficient for $\Delta(\alpha) \leq 0$, that is, for an increase in lying aversion to (weakly) lower total direct costs of lies to liars.

Proposition 5B. An increased level of lying aversion (higher α) enhances social welfare in model M1 if condition C2 holds.

Corollary. Suppose a shift from voluntary self-reporting to compulsory self-reporting raises lie aversion proportionally, increasing α from $\alpha_v \in (0, 1)$ to $\alpha_c = 1$. This shift enhances social welfare in model M2; in model M3 if $\alpha^* > 0$; and in model M1 if condition C2 holds.

7 Conclusion

Self reporting of behavior is a common requirement in law enforcement and regulation. Polluters are often required to report pollution exceedances; food producers are required to report instances of contamination; product producers are required to report product flaws that can endanger users' safety. Enforcement regimes can elicit self-reports of violations by confronting agents with lower costs when they self-report their offenses rather than face probabilistic discovery and prosecution. This paper considers the merits and design of self-reporting enforcement regimes in a standard accident regulation model with one key twist: agents exhibit a heterogeneous aversion to lies. Experimental evidence documents that many individuals are averse to lies and that the extent of aversion differs from one person to the

next (e.g., Gneezy, 2005; Gibson et al., 2013). With lie aversion, agents have an added incentive to truthfully report violations, rather than falsely report "no violation," because the truth avoids the cost of a lie. Truthful reports can thereby be elicited even when resulting fines are higher than they would be under conventional enforcement - so long as the excess doesn't exceed the lie aversion margin. The higher fines enable increased deterrence of violations without any additional enforcement effort. As a result, a self-reporting feature is optimal even when, without lie aversion, it would enjoy no advantage. This conclusion, in turn, implies that lie aversion is a good thing for society; it enables savings in enforcement costs. That is, moral sentiments can be advantageous not only when they inculcate guilt and/or virtue that promote harm prevention (Kaplow and Shavell, 2007), but also when they advance truthful communication.

Lie aversion is important to other features of optimal enforcement, including the adoption of a compulsory (vs. voluntary) self-reporting requirement. One noteworthy symptom of optimal enforcement under lie aversion is that self-reporting sanctions are higher than average sanctions experienced with false reports, and particularly high if the social planner is not concerned about the aversion costs of lies borne by liars. Prior work on self-reporting prescribes fines for truthful reporting agents that are either equal to or lower than average "non-reporting" sanctions. For example, costs of auditing programs or efforts to avoid apprehension (as in Malik, 1990) can motivate self-reporting fines lower than average non-reporting sanctions (Innes, 2001a; Innes 2001b; Pfaff and Sanchirico, 2000). Lower sanctions are in fact observed in some contexts, for example states with sweeping immunity protections for violations uncovered and reported as a result of individual companies' self-auditing programs (see Guerrero and Innes, 2013, and Short and Toffel, 2008, for discussion). However, in other contexts, self-reporting sanctions are seemingly large relative to those that extant theoretical models would prescribe. For example, the EPA's environmental self-auditing policies provide only limited exemptions from sanctions for self-reported infractions (EPA, 1995, 2000). At least when costs of lies "count," such policies are unlikely to find support from the calculus of this paper's model. This said, significant lie aversion justifies concern about excessively liberal "leniency" programs that deplete harm prevention incentives of the most lie averse violators.

Results here are presumably limited to environments in which lie aversion is important in relation to stakes in reporting decisions. For individuals and small businesses engaged in small-scale interactions with government officials (as in the case of tax compliance and in many of the potentially corrupt exchanges of interest in Burlando and Motta (2016), for example), this condition is likely to be satisfied. However, even with larger companies in which company compliance officers are making reporting decisions, reputational concerns of both individuals and the companies themselves are likely to produce significant lie aversion in some cases. Although these forces are beyond the scope of modeling in the present paper, Baron's (2009) work on competition in corporate social responsibility suggest a potential role for lie aversion in the corporate marketplace. [15]

While this paper studies how lie aversion can affect optimal law enforcement, the forces at play are potentially relevant to other settings, including reporting regimes in employer-employee, buyer-supplier, seller-consumer and borrower-lender relationships. [16] Perhaps an over-arching message is that lie aversion elevates the importance of compulsory communication to the design of efficient contracts and arrangements between individuals, companies, and the government.

[15] See also Baron (2010), Feddersen and Gilligan (2001), and Besley and Ghatak (2007) for a few examples in a growing literature on corporate social responsibility.

[16] See related work on costs of lies in cheap talk and signaling games (Chen et al., 2008; Kartik, 2009; Deneckere and Severinov, 2007) and communication and coordination interactions (Demichelis and Weibull, 2008; Ellingsen and Ostling, 2010).

8 References

Abeler, J., A. Becker and A. Falk. 2014. "Representative Evidence on Lying Costs." *Journal of Public Economics* 113: 96-104.

Aubert, C., P. Rey and W. Kovacic. 2006. "The Impact of Leniency and Whistle-Blowing Programs on Cartels." *International Journal of Industrial Organization* 24: 1241-66.

Baron, D. 2009. "A Positive Theory of Moral Management, Social Pressure, and Corporate Social Performance." *Journal of Economics and Management Strategy* 18: 7-43.

Baron, D. 2010. "Morally Motivated Self-Regulation." *American Economic Review* 100: 1299-1329.

Becker, G. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76: 169-217.

Benabou, R. and J. Tirole. 2006. "Incentives and Pro-Social Behavior." *American Economic Review* 96: 1652-78.

Besley, T. and M. Ghatak. 2007. "Retailing Public Goods: The Economics of Corporate Social Responsibility." *Journal of Public Economics* 91: 1645-53.

Bigoni, M., S. Fridolfsson, C. Le Coq and G. Spagnola. 2012. "Fines, Leniency, and Rewards in Antitrust." *RAND Journal of Economics* 43: 368-90.

Buccirossi, P. and G. Spagnola. 2006. "Leniency Policies and Illegal Transactions." *Journal of Public Economics* 90: 1281-97.

Burlando, A. and Al. Motta. 2016. "Legalize, Tax, and Deter: Optimal Enforcement Policies for Corruptible Officials." *Journal of Development Economics*, in press.

Chen, Y., N. Kartik and J. Sobel. 2008. "Selecting Cheap-Talk Equilibria." *Econometrica* 76: 117-36.

Conrads, J., B. Irlenbusch, R. Rilke and G. Walkowitz. 2013. "Lying and Team Incentives." *Journal of Economic Psychology* 34: 17.

Demichelis, S. and J. Weibull. 2008. "Language, Meaning and Games: A Model of Communication, Coordination and Evolution." *American Economic Review* 98 (4): 12921311.

Deneckere, R. and S. Severinov. 2007. "Optimal Screening with Costly Misrepresentation." Working Paper, University of Wisconsin.

Dreber, A. and M. Johannesson. 2008. "Gender Differences in Deception." *Economics Letters* 99: 197199.

Ellingsen, T., M. Johannesson, J. Lilja and H. Zetterqvist. 2009. "Trust and Truth." *Economic Journal* 119: 252-276.

Ellingsen, T. and R. Ostling. 2010. "When Does Communication Improve Coordination¿' *American Economic Review* 100: 16951724.

Erat, S. and U. Gneezy. 2012. "White Lies." *Management Science* 58: 723-33.

Feddersen, T. and T. Gilligan. 2001. "Saints and Markets: Activists and the Supply of Credence Goods." *Journal of Economics and Management Strategy* 10: 149-71.

Fees, E. and M. Walzl. 2004. "Self-Reporting in Optimal Law Enforcement When There Are Criminal Teams." *Economica* 71: 333-48.

Fees, E. and M. Walzl. 2006. "Heterogeneity and Optimal Self-Reporting." *Journal of Institutional and Theoretical Economics* 162: 277-90.

Fischbacher, U. and F. Fllmi-Heusi. 2013. "Lies in Disguise An Experimental Study on Cheating." *Journal of the European Economic Association* 11: 525-547.

Friesen, L. 2006. "The Social Welfare Implications of Industry Self-Auditing." *Journal of Environmental Economics and Management* 51: 280-294.

Friesen, L. and L. Gangadharan. 2012. "Individual Level Evidence of Dishonesty and the Gender Effect." *Economics Letters* 117: 624-626.

Friesen, L. and L. Gangadharan. 2013. "Designing Self-Reporting Regimes to Encourage Truth-Tellling: An Experimental Study." *Journal of Economic Behavior and Organization* 94: 90-102.

Gawn, G. and R. Innes. 2015. "Language and Lies: Does the Strength of a Communication Affect the Intrinsic Aversion to Dishonesty? " Working Paper, U.C. Merced.

Gerlach, H. 2013. "Self-Reporting, Investigation and Evidentiary Standards." *Journal of Law and Economics* 56: 1061-90.

Gibson, R., C. Tanner and A. Wagner. 2013. "Preferences for Truthfulness: Heterogeneity Among Within Individuals." *American Economic Review* 103: 532-48.

Gneezy, U. 2005. "Deception: The Role of Consequences." *American Economic Review* 95: 38494.

Gnnezy, U. and A. Rustichini. 2000. "A Fine is a Price." *Journal of Legal Studies* 29: 1-17.

Guerrero, S. and R. Innes. 2013. "Self-Policing Statutes: Do They Reduce Pollution and Save Regulatory Costs? " *Journal of Law, Economics and Organization* .

Innes, R. 1999a. "Remediation and Self-Reporting in Optimal Law Enforcement." *Journal of Public Economics* 72: 379-393.

Innes, R. 1999b. "Self-Policing and Optimal Law Enforcement when Violator Remediation is Valuable." *Journal of Political Economy* 107: 1305-1325.

Innes, R. 2000. "Self-Reporting in Optimal Law Enforcement When Violators Have Heterogeneous Probabilities of Apprehension." *Journal of Legal Studies* 29: 287-300.

Innes, R. 2001a. "Violator Avoidance Activities and Self-Reporting in Optimal Law Enforcement." *Journal of Law, Economics and Organization* 17: 239-56.

Innes, R. 2001b. "Self-Enforcement of Environmental Law," in A. Heyes, ed., *The Law and Economics of the Environment*. Cheltenham: Elgar.

Innes, R. and A. Mitra. 2013. "Is Dishonesty Contagious? " *Economic Inquiry* 51: 722-34.

Kaplow L. and S. Shavell. 1994. "Optimal Law Enforcement with Self-Reporting of Behavior." *Journal of Political Economy* 102: 583-606.

Kaplow, L. and S. Shavell. 2007. "Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System." *Journal of Political Economy* 115: 494-514.

Kartik, N. 2009. "Strategic Communication With Lying Costs." *Review of Economic Studies* 76: 1359-95.

Langpap, C. 2008. "Self-Reporting and Private Enforcement in Environmental Regulation." *Environmental and Resource Economics* 40: 489-506.

Lewin, J. and W. Trumbull. 1990. "The Social Value of Criime? " *International Review of Law and Economics* 10: 271-84.

Livernois, J. and C. J. McKenna. 1999. "Truth or Consequences: Enforcing Pollution Standards with Self-Reporting." *Journal of Public Economics* 71: 415-440.

Lundquist, T., T. Ellingsen, E. Gribbe and M. Johannesson. 2009. "The Aversion to Lying." *Journal of Economic Behavior and Organization* 70: 81-92.

Malik, A. 1990. "Avoidance, Screening and Optimum Enforcement." *RAND Journal of Economics* 21: 341-53.

Malik, A. S. 1993. "Self-Reporting and the Design of Policies for Regulating Stochastic Pollution." *Journal of Environmental Economics and Management* 24: 241-257.

Motta, M. and M. Polo. 2003. "Leniency Programs and Cartel Prosecution." *International Journal of Industrial Organization* 21: 347-79.

Pfaff, A. and C. Sanchirico. 2000. "Environmental Self-Auditing: Setting the Proper Incentives for Discovery and Correction of Environmental Harm." *Journal of Law, Economics and Organization* 16: 189-208.

Pfaff, A. and C. Sanchirico. 2004. "Big Field, Small Potatoes: An Empirical Assessment of EPAs Self-Audit Policy." *Journal of Policy Analysis and Management* 23: 415-432.

Rosenbaum, S., S. Billinger and N. Stieglitz. 2014. "Lets be honest: A review of experimental evidence of honesty and truth-telling." *Journal of Economic Psychology* 45: 181-96.

Shavell, S. 1991. "Specific vs. General Enforcement of Law." *Journal of Political Economy* 99: 1088-1108.

Shavell, S. 2002. "Law versus Morality as Regulators of Conduct." *American Law and Economics Review* 4: 227-57.

Short, J. and M. Toffel. 2008. "Coerced Confessions: Self-Policing in the Shadow of the Regulator." *Journal of Law, Economics and Organization* 24: 45-71.

Stafford, S. 2005. "Does Self-Policing Help the Environment? EPAs Audit Policy and Hazardous Waste Compliance." *Vermont Journal of Environmental Law* 6: 2.

Stafford, S. 2007. "Should You Turn Yourself In? The Consequences of Self-Policing." *Journal of Policy Analysis and Management* 26: 305-26.

Stigler, G. 1970. "The Optimum Enforcement of Laws." *Journal of Political Economy* 78: 526-56.

Toffel, M. and J. Short. 2011. "Coming Clean and Cleaning Up: Does Voluntary Self-Reporting Indicate Effective Self-Policing? " *Journal of Law and Economics* 54: 609-49.

U.S. Environmental Protection Agency (EPA). 1995, 2000. "Incentives for Self-Policing: Discovery, Disclosure, Correction, and Prevention of Violations." 60 Fed. Reg. 66705 / 65 Fed. Reg. 6576-3.

9 Appendix: Proof of Propositions 1B and 2(v)

Define the following measures of under-deterrence:

$$U_1(s) =\mid p'(c^*(s))h + 1 \mid c^{*\prime}(s)\overline{s} \tag{28}$$

$$U_2(s,a) =\mid p'(c^*(s+a))(h+a) + 1 \mid c^{*\prime}(s+a)\overline{s} \tag{29}$$

Note that, with under-deterrence ($s \leq h$), these measures are monotone decreasing in the sanction s. That is, under-deterrence falls when the sanction is raised:

$$\frac{\partial U_i}{\partial s} = -p''(h+\delta a)(c^{*\prime})^2\overline{s} - (p'(h+\delta a)+1)c^{*\prime\prime}\overline{s} < 0 \,\text{for}\, i \in [1,2] \tag{30}$$

where $\delta = 0$ for $i = 1$, $\delta = 1$ for $i = 2$, and the inequality follows from $p'' > 0$, $c^{*\prime} > 0$, $s \leq h$ (implying $p'(h+\delta a)+1 \leq 0$), and with $p' < 0$, $p'' > 0$, and $p''' \geq 0$,

$$c^{*\prime\prime} = (c^{*\prime}/s)[(p'p'''/p''^2)2] < 0 \tag{31}$$

We can define corresponding indices of under-deterrence for agents under NSR and SR enforcement regimes: regimes:

$$U_{NSR} = U_1(q_{NSR}\overline{s}) \;=\; \text{under-deterrence under NSR enforcement} \tag{32}$$

$$U_{SRT} = U_1(s_R) \;=\; \text{under-deterrence for truthful self-reporters under SR enforcement} \tag{33}$$

$$U_{SRL}^{M1} = (1/G(a^*)) \textstyle\int_0^{a^*} U_2(q_{SR}\overline{s}, a)g(a)da =$$

$$\text{average under-deterrence for liars under SR enforcement (M1)} \tag{34}$$

$$U_{SRL}^{M2} = (1/G(a^*)) \textstyle\int_0^{a^*} U_1(q_{SR}\overline{s} + a)g(a)da =$$

$$\text{average under-deterrence for liars under SR enforcement (M2)} \tag{35}$$

The optimality condition for q_{NSR} (eq. (3)) can be written:

$$q_{NSR} : -U_{NSR} + m'(q) = 0 \tag{36}$$

Combining optimality conditions for (q_{SR}, s_R) for M1 (eq. (11)) and M2 (eq. (17)):

$$q_{SR} : -U^*_{SR} + m'(q) = 0 \ , \text{ where } U^*_{SR} = G(a^*)U_{SRL} + (1 - G(a^*))U_{SRT} \tag{37}$$

The following results (3 and 4) establish Proposition 1B and Proposition 2(v), where U^*_{SR} measures "average under-deterrence":

Result 1. $U^*_{SR} - U_{NSR} = m'(q_{SR}) - m'(q_{NSR})$.

Proof: Subtract (37) from (36).

Result 2. $U_{SRL} > U_{SRT}$.

Proof: For $0 < a < a^*$ (with $a^* > 0$ by Propositions 1A and 2),

$$U_2(q_{SR}\overline{s}, a) > U_1(q_{SR}\overline{s} + a) > U_1(s_R) \tag{38}$$

where the inequalities follow from the definitions of U_2 and U_1 (with $p' < 0$), under-deterrence in the SR optimum, $q_{SR}\overline{s} + a < s_R \leq h < h + a$ for $a \in (0, a^*)$, and (30). (38) implies Result 2 by the definitions of U_{SRL} and U_{SRT} in (33) to (35).

Result 3. $s_R > q_{NSR}\overline{s}$ ($\leftrightarrow \ U_{SRT} < U_{NSR}$).

Proof: Suppose not, $s_R \leq q_{NSR}\overline{s}$ (and therefore, by their definitions and (30), $U_{SRT} \geq U_{NSR}$). Using Result 2, we then have

$$U_{SRL} > U_{SRT} \geq U_{NSR} \rightarrow U^*_{SR} > U_{NSR} \tag{39}$$

With $a^* > 0$ (by Propositions 1A and 2) and $s_R = q_{SR}\overline{s} + a^* \leq q_{NSR}\overline{s}$, we also have $q_{SR} < q_{NSR}$. However, $U^*_{SR} > U_{NSR}$ and $q_{SR} < q_{NSR}$ contradict Result 1 (with $m'' \geq 0$).

Result 4. (a) $q_{SR} < q_{NSR}$, and (b) $U^*_{SR} < U_{NSR}$.

Proof: (b) follows from Results 1 and 4(a). To prove (a), suppose the contrary, $q_{SR} \geq q_{NSR}$. There are the two cases, M1 and M2. For M1, note that for $s = q_{SR}\overline{s}$ and $a \in [0, a^*]$,

$$\frac{\partial U_2(s,a)}{\partial a} = -[p''(h+a)c^{*\prime} + p']c^{*\prime}\overline{s} - [p'(h+a)+1]c^{*\prime\prime}\overline{s} < 0 \tag{40}$$

where the equality follows from under-deterrence (by Proposition 1A) and the inequality follows from $c^{*\prime\prime} < 0$ (by (31)), $p'(h+a)+1 < 0$ (under-deterrence), and with $p' < 0$ and $h > s$ by Proposition 1A,

$$[p''(h+a)c^{*\prime} + p'] = [p'/(s+a)](s-h) > 0, \tag{41}$$

where $c^{*\prime} = -p'/[p''(s+a)] > 0$ has been substituted. (40) implies that (for $a \in (0, a^*]$)

$$U_2(s,a) < U_2(s,0) = U_1(s) \tag{42}$$

and, therefore,

$$U_{SRL}^{M1} < U_1(q_{SR}\overline{s}) \leq U_1(q_{NSR}\overline{s}) = U_{NSR} \tag{43}$$

where the second inequality is due to the initial premise, $q_{SR} \geq q_{NSR}$ and (30). (43) and Result 3 imply that $U_{SR}^* < U_{NSR}$, contradicting Result 1 (with $m'' \geq 0$ and $q_{SR} \geq q_{NSR}$).

For M2, $q_{SR} \geq q_{NSR}$ implies that

$$q_{SR}\overline{s} + a > q_{NSR}\overline{s} \text{ for } a \in (0, a^*] \tag{44}$$

Together with (30) (and the definition in (35)), (44) implies

$$U_{SRL}^{M2} < U_{NSR} \rightarrow U_{SR}^* < U_{NSR} \rightarrow q_{SR} < q_{NSR} \tag{45}$$

where the first implication is due to the definition of U_{SR}^* in (37) and Result 3, and the second is due to Result 1. (45) contradicts the premise, $q_{SR} \geq q_{NSR}$.

Figure 1. Improving Incentives with Self-Reporting

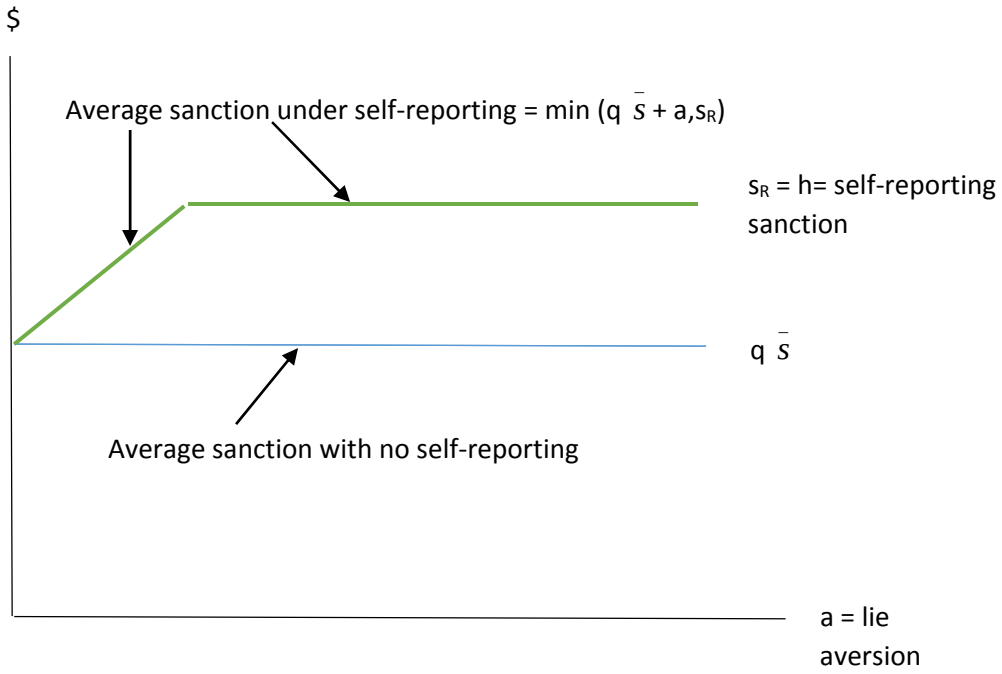

Figure 2. Change in Costs When Lie Aversion Increases (from α<1 to α=1)