

# Whistleblowers and the Regulation of Environmental Risk

Anthony G. Heyes

Royal Holloway, UNIVERSITY OF LONDON

Short title: WHISTLEBLOWERS

---

<sup>1</sup>Present address Department of Economics, Royal Holloway College, UNIVERSITY OF LONDON, England TW20 OEX, e-mail a.heyes@rhul.ac.uk. I am grateful to David Martimort, Bernard Sinclair-Desagne, Charles Mason, Mark Raymond, Peth Tuppe and seminar participants in Toulouse and London for helpful comments. Errors are mine.

**Abstract.** ‘Whistleblowing’ is a common feature of our regulatory landscape, yet there is no formal economic model of it. We propose such a model. Sociological and psychological research in the area points to three alternative theories as to why individuals might disclose, even when such action is not in their (apparent) self-interest. Individuals who disclose are either conscience cleansers, welfarist, or punishment-motivated. The policy problem is to decide how frequently to pursue disclosures made by whistleblowers, and how substantially to fine firms whose plans for wrong-doing are detected in this way. Not surprisingly, optimal policy depends upon the motives attributed to whistleblowers, but is not in general characterised by maximal penalties nor routine pursuit of complaints, even when pursuit is costless. **Keywords:** **Regulation - enforcement and compliance - behavioral law and economics.**

## 1. Introduction

‘Whistleblowing’ - whereby an individual working within an organization helps an external agency to identify and prevent wrong-doing by that organization - is a common feature of the regulatory landscape. In the US whistleblowing clauses are included in pieces of legislation as diverse as the Occupational Health and Safety Act (OSHA) of 1970, the Clean Air Act Amendments (CAAA) of 1977 and the Financial Institutions Reform Act of 1989.

In light of this it is surprising that (to the best of our knowledge) there exists no formal economic model of whistleblowing.<sup>1</sup>

Glazer and Glazer (1989) define the whistleblower as one who (1) acts to prevent harm to others, not him or herself, (2) while possessing evidence that would convince a reasonable person. Though other definitions exist this is a frequently-cited one, and we adopt it here.

The biggest hurdle in setting-up an economic model of compliance based on disclosure by whistleblowers, is the need to explain why disclosure is likely to happen at all. Whistleblowing may well be a privately costly action but the benefits accrue (by definition) to others, so the behavior is not easily explained under conventional assumptions about rational, self-interested behavior. The model presented here explores alternative behavioral motivations and is in the spirit of the recent behavioral law and

---

<sup>1</sup>Outside economics there is an extensive literature. Twenty books about whistleblowing are available from Amazon.com, and more than 100 articles have been published in the sociology, psychology, business and public administration literatures. This research will inform the assumptions here. There is also, of course, a long history of popular press coverage of individual whistle-blowing cases. There is, of course, an established literature on self-reporting of offences, and reporting by victims (examples are Mookerjee and P’ng (1992) and Livernois and McKenna (1999) respectively)).

economics literature associated with scholars such as Cass Sunstein, Richard Zeckhafer, Thomas Ulen and others (see, for examples, the work contained in Sunstein (2000)).

### 1.1. What Makes a Whistleblower Whistle?

So what do scholars in other disciplines have to say about the motivations that might underpin the behavior of whistleblowers?

Empirical efforts to identify a whistleblower ‘type’ - and more generally efforts to predict what sorts of people become ‘rescuers’ - have been comparatively unfruitful. Correlations with observables such as social class, religion and educational attainment have proved insignificant or small (see, for examples, Bauman (1989), Oliner and Oliner (1998) and citations there-in). “Rescuers come from all corners and sectors of social structure, thereby calling the bluff of those believing there to be largely social determinants of moral behavior” (Bauman (1989:5)). This is recollective of the findings of Milgram (1974) in his famous study *Obedience to Authority* that there were no convincing social psychological correlations with disobedience, even with the various measures of moral development used by social psychologists.

Alford (2001) develops a theory of whistle-blowing that builds on concepts of individuality. He presents interview-based case study evidence of motivations consistent with deontological justifications of obligation and duty. Individuals, he concludes, are caught between loyalties to their employers, society and ego. He proposes a novel ethical category which he calls ‘narcissism moralized’. In summary, “(n)arcissism becomes moral when the self’s commitment to the highest ideals is based in the internalized image of an ideal self, so bound to its ideals that there is in the end no difference between the ideal self and ideals of the self ... (w)histleblowers disclose because they dread living with a corrupted self more than they dread the other outcomes” (Alford (2001: 90)).

Alford’s description of motivation does not, however, tell us the behavior that those

highest ideals prescribe. In a legal context it does not tell us when - if ever - violation of a law is defensible? Is it, for example, defensible when the cost of compliance is sufficiently large?

The answers to such questions rely on individual moralities. Hunt (1997) proposes that a 'justifiable disclosure' must at minimum (a) serve some purpose in correcting or preventing harm and (b) do more harm than good. The definition of whistleblowing we adopt implies that (a) is necessarily satisfied. (b) implies some weighing up of social costs and benefits and implies, importantly, that the disclosure decision is forward-looking. Hunt notes that in practice the whistleblower may find it difficult or impossible to ascertain whether (b) is satisfied. Conceptually, though, he notes that:

“...all the well-rehearsed arguments for and against utilitarian calculation could be invoked at this point. The moral codes of some will lead them to take the view that it reasonable to make a disclosure simply and only because ‘it is the right thing to do’ even if harmful consequences are known to be more likely than beneficial ones. A whistleblower in this position might feel, for example, that they are answerable to God who will judge them *only for following moral principles of honesty and fortitude, not for the consequences of the rightful act*” (Hunt (1997:2), italics added).

He also, we note in passing, regards the rectification criterion as contestable. “The whistleblower, as we have seen, may not be concerned so much with the consequences of the disclosure as with simply making the truth known, because it is the truth.” Some may disclose to redeem themselves for complicity or collusion in past or planned organizational wrong-doing, though “the evidence remains that most whistleblowers appeal to rectification to justify their act” (Hunt (1997: 3)).

Alford’s notion of narcissism moralized is, as he observes, closely related to sense of shame. “Though not all whistleblowers use the language of shame or guilt, many talk

about feeling dirty or corrupted by the acts of others with whom they are associated, and an inability to live with that corruption” (p 74).

In terms of unobservable characteristics of an individual psyche which might be expected to determine the propensity to disclose, psychologists offer the concepts of ‘imagination for consequences’ and ‘doubling’.

Different people may attach the same weight to the well-being of others, but differ in their ability to anticipate or comprehend the impact of action (or inaction) upon others. Arendt (1978) promotes imaginative capability and the ability to take account of others as the ground of all ethics.

Another dimension along which individuals are likely to differ is in their ability to ‘double’ (Lifton (1986)). Doubling takes place when a part of the self comes to act autonomously. Ethical qualms that our whole self might have can be ignored or suspended at work because a ‘work self’ temporarily prevails. Different individuals differ in the extent to which they are able to insulate their work self from their whole being, and to tolerate behavior from the former which would be unconscionable to the latter. As one manager (not a whistleblower) put it: “What is right in the corporation is not what is right in a man’s house or in his church. What is right in the corporation is what the guy above you wants from you. That’s what morality is in the corporation” (Jackall (1988:6)). “Doubling is a sophisticated emotional and cognitive act, one that whistleblowers have difficulty performing. In this sense they might be regarded as dysfunctional actors in modern society” (Alford (2001: 73)).

An alternative and popular notion is of whistleblowers as what Alford calls ‘hysterical malcontents’, motivated by a desire to punish the organization of which they are part. A third assumption that we might make, then, is that whistleblowers are driven by a desire to punish the firm that employs them.

A certain proportion of staff might be “disgruntled employees”, unhappy for reasons unconnected with the firm’s planned non-compliance with the regulation, but

opportunistic in blowing the whistle when so doing creates sufficiently substantial discomfort (*i.e.* cost) for their employer.

Alternatively, the individual's desire to punish the firm may relate to the non-compliance decision itself. There is growing experimental evidence that at least a subset of people are willing to act to punish others who have behaved anti-socially or illegally, even if that punishment is not in their apparent self-interest. Fehr and Schmidt (1999) and Fehr and Falk (2002) outline parts of this evidence, and draw out some of the implications for economic incentives. Such punishment might also be perceived as making an altruistic contribution to the public good, and so yield the whistleblower with a feeling of 'warm glow' (Andreoni (1990)).

This type of motivation would be consistent with the proposal by Sunstein (2003) that people behave according to an 'outrage heuristic'. As he notes in a Section entitled 'Pointless Punishment':

Peoples intuitions about punishment seem quite disconnected from the consequences of punishment. Suppose, for example, that a corporation has engaged in serious wrongdoing. People are likely to want to punish the corporation as if it were a person. They are unlikely to inquire into the future consequences of punishment. Punishment judgements are rooted in a simple heuristic, to the effect that penalties should be a proportional response to the outrageousness of the act (Sunstein (2003:5)).

Significantly outrageousness of an act may be unconnected to the forward-looking costs and benefits of regulatory intervention. He presents experimental evidence that subjects (including judges) recommend similar levels of punitive damages in different hypothetical contexts in which though the past violation was common, the next period implications of penalty levels differed. This again suggests that individuals do not look forward at the consequences of action (in our case disclosure), but rather backward at

the characteristics of the act itself.<sup>2</sup>

In developing a tractable economic treatment of compliance incentives and enforcement informed by disclosure, some concrete motivation for disclosure must be embedded in the model. Whilst we have presented a cursory summary of parts of the relevant literature, a wider reading suggests three preponderant alternative views about why employees blow the whistle on errant employers: (a) conscience-cleansing (an unwillingness or inability to be complicit with wrongdoing), (b) social motivation (based on a calculation of social costs and benefits) and (c) the desire to punish the wrongdoer. Because of the difficulty of arriving at a consensus view we conduct our analysis under three different assumptions, proposing simple decision rules which operationalize each of these.

## 2. A Regulatory Setting

To explore the relationship between whistleblower motives, agency behavior and compliance requires we establish a generic compliance/enforcement setting in which whistleblowing can occur.

The definition of whistleblowing we have adopted requires, recall, that the individual (1) acts to prevent harm to others, not him or herself (2) while possessing evidence of intent that would convince a reasonable person (Glazer and Glazer (1989:4)).

To capture (1) in a single-period setting requires there be a lag between a decision being taken to violate a law, and the moment at which the violation occurs and harm

---

<sup>2</sup>Sunstein, Schkade and Kahneman (2000) invoke the same heuristic to ‘explain’ why people *do not appear to want* optimal deterrence. In their experiments they found that varying the probability of detection had no effect on punitive awards. Even when peoples attention was explicitly directed to the differing probabilities of detection in alternative scenarios, people were indifferent to it.

results.<sup>3</sup> It is easy to think of stories consistent with this. An employee may be aware that a decision has been made to dispose of some unit of waste illegally rather than using a licensed contractor, or that during construction of a plant faulty materials are being used to cut cost, or critical safety features are being omitted. Before actual harm results the employee has the opportunity to report to the regulatory agency that violation is planned. Alternatively, in a repeated setting we might think of future violation being anticipated on the basis of past record. It must also be the case, again plausible, that the regulator has time to respond to the report and prevent the plans for violation being put into practice (as in Heyes (1996)).

Consistent with (2), and the examples provided, we will assume that the whistleblower has inside information regarding a firm's intent to violate, information which an external agent could not observe directly but which can be evidenced.<sup>4</sup>

Consider the following stylized representation. There is some regulation, compliance with which prevents a firm imposing external damage  $d$ . Compliance is costly. Firm

---

<sup>3</sup>Retrospective reporting of harm due to violation may occur in many settings, and a commitment that those reports will be pursued and penalised might be expected to strengthen *ex ante* compliance incentives. Such reporting does not, however, constitute whistleblowing.

<sup>4</sup>These are the incriminating e-mails and unshredded internal memos that are the stuff of high-profile whistleblowing cases. The evidence is assumed secure enough to form the basis for the regulatory agency to impose a penalty for intent to violate. Alternatively one can think of the whistleblower not holding evidence herself, but being able to provide the inspector with a tip-off on where in the firm to find it. It is important to note that only in cases where the employee holds *hard* information (in the formal sense) can we regard disclosure as whistleblowing according to Glazer and Glazer's definition. I am grateful to David Martimort for useful discussion on this point.

$i$ 's cost of compliance is  $c_i$  and is drawn from a distribution on the interval  $[0, \infty)$  with density  $g(c)$ .

At time  $t = 1$  the firm (which we can equally think of as a single principal) makes a plan either to 'comply' or to 'violate'. That decision may be known to a number of employees, but to abstract from the possibility of multiple potential whistleblowers we assume that the content of the decision is known to a single agent. The agent holds evidence of the decision and this is externally verifiable. He also knows  $c_i$  though information on costs will - as is common - be assumed unverifiable.

At  $t = 2$  the agent has the opportunity to report planned violation to the regulator.

At  $t = 3$  the regulator decides whether to act upon any report received. As information from whistleblowers is qualitatively uniform (indicating only a firm's intent to violate) the agency acts upon the report with probability  $\pi$ , and ignores it with probability  $(1 - \pi)$ . Initially we will regard  $\pi$  as exogenous. Later we will consider how, at  $t = 0$ , the agency will go about determining how responsive its institutions and practices should be to the activities of whistleblowers.

An enforcement action entails two things - (1) the firm is coerced into compliance and, (2), is subject to a fine  $f$  for intent to violate.

For simplicity we will proceed *as if* there are no pecuniary costs or benefits to the whistleblower. There are numerous legal protections for whistleblowers in the US and Europe,<sup>5</sup> and the aim of associated compensation programs is to ensure that

---

<sup>5</sup>Examples of American Acts that prohibit discharging or discriminating against an employee who reports an actual or anticipated violation of the Act include the Clean Air Act Amendments (1977), Energy Reorganisation Act (1974) (in the context of nuclear safety codes), the Financial Institutions Reform, Recovery and Enforcement Act (1989) and the Defense Contractor Act (1986). There is an overarching Whistleblower Protection Act (1989). There are also protections in common law.

individuals are compensated for any private losses that they might incur. The most restrictive assumption we might make is that such a program based on protections and/or compensation is in place and functioning perfectly. That would not generally be realistic, however. There is plenty of evidence that legal protection is less than full, and that whistleblowing imposes a significant cost upon the individuals concerned (for examples see Miethe (1999)). The key costs usually cited are things such as premature dismissal, foregone promotion, less favorable work or office assignments, and other diminutions of career prospects. In terms of our model, though, the important thing is that there do not appear to be significant social costs. The examples cited can be regarded as largely redistributive in nature - the job, promotion, bigger office, salary increase or whatever, goes to another employee - and as such private but not social costs. They can therefore be regarded as captured in the  $\mu_i, \theta_i, \delta_i$  ‘thresholds’ introduced in (1.1) through (1.3) below.

In the most general case we would expect the decision to disclose would depend upon  $c_i, \pi$  and  $f$ . We posit a general ‘motivation function’ that describes the probability that an individual will disclose when faced with a particular set of these variables and parameters,

$$\rho_{\Delta}(c_i, \pi, f).$$

We work the model under three alternative assumptions about what motivates disclosure, and  $\Delta \in \{\alpha, \beta, \gamma\}$  denotes the assumption in play.

### 2.1. Scenario $\alpha$ : Whistleblowing as Conscience Cleansing

One assumption we might make about a whistleblower is that his or her decision regarding whether or not to disclose is based solely upon the characteristics of the proposed offence. In particular the moral defensibility of the employers decision (consistent with Hunt (1997) already quoted).

The prospective violator, recall, takes a decision to impose external environmental

damage of value  $d$  in order to save private costs  $c_i$ . An individual can take a view on the morality of that decision - and decide whether or not his conscience allows him to ‘live’ with that decision - without reference to the enforcement environment. The impact of non-compliance is fixed at  $d$ , so we can regard  $c_i$  as a measure of the defensibility of the firm’s decision to, then assume that an individual has some threshold of defensibility beyond which his conscience compels him to speak out.

Choosing to not comply if  $c$  is very small (even zero) would be a particularly outrageous/inexcusable/indefensible decision and particularly likely to prompt disclosure. As  $c$  gets larger it is more defensible that the firm would opt to violate, though individuals may differ in their ‘forgiveness’ (consider attitudes if  $c$  were to approach infinity).

We operationalize this by making *Assumption  $\alpha$  : Individual  $i$  reports planned violation if and only if*

$$c_i < \mu_i. \tag{1}$$

It is plausible to suppose that  $\mu_i$  will vary across people (there may also be variability in the  $\mu$  that a person applies in decision-making through time) in a way that can only be observed by the individual. The probability that planned violation by firm  $i$  will be disclosed to the regulator can then be described as some function of the defensibility of the act,  $\rho_\alpha(c_i)$ . The enforcement environment,  $\pi$  and  $f$ , do not impact the disclosure decision and so we omit them to save on notation. Note that  $c_i < c_j \leftrightarrow \rho_\alpha(c_i) \geq \rho_\alpha(c_j)$ .

In Figure 1 we represent  $\rho_\alpha(c_i)$  as an ogive - consistent with  $\mu_i$  being distributed according to some continuous single-peaked distribution - but nothing rests on this. We can pick out two special cases. First, that in which  $\rho_\alpha(d) = \frac{1}{2}$ , meaning that *on average* employees regard  $c_i > d$  as a reasonable defence for law-breaking. Second,  $\rho_\alpha(c_i) = 1$  for all  $c_i$  would be the case law-breaking was never regarded as defensible and employees

always felt morally compelled to disclose. These would correspond to the ‘pathologically honest’ individuals identified by Heyes (2001).

Significantly, the probability of disclosure does not depend upon the characteristics of the enforcement regime.

Under assumption  $\alpha$  firm  $i$  will comply voluntarily if

$$c_i \leq \rho_\alpha(c_i) \cdot \pi \cdot (c_i + f) \quad (2)$$

The left-hand side is the cost of compliance, the right-hand side is the expected cost of non-compliance.

Compliance incentives - the net expected benefits from compliance - are monotonically decreasing in  $c_i$ . That is  $(\rho_\alpha(c_i)\pi(c_i + f) - c_i)$  is decreasing in  $c_i$ . An increase in  $c_i$  both makes compliance more costly and reduces the chance that non-compliance will be reported and therefore prosecuted. For values of  $c_i$  close to 0 net benefits must be positive. Therefore, the firm will comply if and only if its cost is less than some critical value  $\widehat{c}(\pi, f|\alpha)$  implicitly defined by

$$0 = \rho_\alpha(\widehat{c}) \cdot \pi \cdot (\widehat{c} + f) - \widehat{c}. \quad (3)$$

It is straight-forward to confirm that  $\partial\widehat{c}/\partial\pi$  and  $\partial\widehat{c}/\partial f$  are both positive. In summary:

**Remark 1** *Under Assumption  $\alpha$  the net private benefits from compliance are monotonically decreasing in the cost of compliance. For a given enforcement environment  $\{\pi, f\}$  the realised probability of compliance is non-increasing in  $c_i$  and equals 1 if  $c_i \leq \widehat{c}$ , equals  $\rho_\alpha(\widehat{c}) \cdot \pi$  otherwise.*

In this case, then, everything works in the expected direction.

The policy questions are two-fold: (a) How responsive should the regulatory agency be to reports from whistleblowers? And (b) how should firms caught through such disclosures be punished?

Assume that the regulator acts to minimize social loss, defined as the sum of expected compliance costs and external damage. The regulator's problem is to choose  $0 \leq \pi \leq 1$  and  $f \leq F$  to minimize:

$$SL(\pi, f|\alpha) = \int_0^{\hat{c}} c_i g(c) + \int_{\hat{c}}^{\infty} [\rho_{\alpha}(c_i)\pi c_i + (1 - \rho_{\alpha}(c_i)\pi)d] g(c) dc \quad (4)$$

Firms with compliance costs below  $\hat{c}$  comply voluntarily. Firms with costs above  $\hat{c}$  will comply only if coerced. This occurs with probability  $\rho_{\alpha}(c_i)\pi$ . Otherwise the firm will be left to execute its plans for violation, imposing external damage  $d$ .

We can first observe that the welfare-motivated regulatory agency will never choose to implement a regime inducing  $\hat{c} \geq d$ . Consider any regime  $\{\pi', f'\}$  such that  $\hat{c}(\pi', f'|\alpha) \geq d$ . Differentiating 4 gives:

$$\frac{\partial SL(\pi', f'|\alpha)}{\partial \pi} = \frac{\partial \hat{c}}{\partial \pi} (1 - \rho_{\alpha}(\hat{c})\pi)(\hat{c} - d) + \int_{\hat{c}}^{\infty} \rho_{\alpha}(c_i)(c_i - d)g(c)dc. \quad (5)$$

Both terms on the right-hand side are positive implying  $SL(\pi', f'|\alpha) > SL(\pi' - \epsilon, f'|\alpha)$ , such that  $\{\pi', f'\}$  cannot be optimal. Optimal policy  $\{\pi^*, f^*\}$  must, then, be such that  $\hat{c}(\pi^*, f^*|\alpha) < d$ .

Given this, consider any pair  $\{\pi', f'\}$  such that  $\hat{c}(\pi', f'|\alpha) < d$ . Then,

$$\frac{\partial SL(\pi', f'|\alpha)}{\partial f} = \frac{\partial \hat{c}}{\partial f} (1 - \rho_{\alpha}(\hat{c})\pi)(\hat{c} - d) < 0 \quad (6)$$

Therefore  $SL(\pi', f'|\alpha) > SL(\pi', F|\alpha)$  for any  $f' < F$ . The optimal penalty must be maximal,  $f^* = F$ .

An interior solution to the Agency's choice of  $\pi$  will then be characterized by

$$\frac{\partial SL(\pi^*, F|\alpha)}{\partial \pi} = \frac{\partial \hat{c}}{\partial \pi} (1 - \rho_{\alpha}(\hat{c})\pi^*)(\hat{c} - d) + \int_{\hat{c}}^{\infty} \rho_{\alpha}(c_i)(c_i - d)g(c)dc = 0 \quad (7)$$

where

$$\hat{c} = \hat{c}(\pi^*, F|\alpha)$$

The first composite term on the right-hand side of 7 is negative. It captures the fact that an increase in inspection probability extends the interval of firms that voluntarily comply, which at the  $\hat{c} < d$  margin increases welfare. The second captures the social loss associated with the increased frequency with which compliance will be coerced at firms with compliance costs above  $\hat{c}$ , which is positive in the vicinity of an interior solution.

If the expression in 7 is negative when evaluated at  $\pi = 1$  then the corner solution  $\pi^* = 1$  results. For current purposes we will assume that this upper corner solution is not binding, which amounts to an assumption that the instruments of enforcement are sufficiently potent. In particular, *if* both enforcement variables were set at their highest levels (that is  $\pi = 1$ ,  $f = F$ ) then the agency would be able to achieve a level of  $\hat{c}$  greater than  $d$ .

**Proposition 1** *If whistleblowing is motivated by conscience cleansing (that is, under Assumption  $\alpha$ ), and the regulator's enforcement instruments are potent enough that  $\hat{c}(1, F|\alpha) > d$  then optimal policy will be characterised by (a) a maximal penalty and (b) an inspection intensity set less than maximally. This applies even though inspection is costless.*

A well-known result in the economics of enforcement, attributed to Becker (1968), that to achieve a particular level of compliance the agency should set fines maximally and adjust inspection intensity correspondingly. His observation is based on inspections being costly, fines not. Here (where inspections are costless) a similar result holds but for quite different reasons. In raising *voluntary* compliance to any particular level there is a strict preference based not on consideration of enforcement costs, but on the pattern of *actual* compliance that results.

The agency can raise  $\hat{c}$  by raising either  $\pi$  or  $f$ . So doing through increases in  $f$  will always be beneficial provided  $\hat{c}$  remains below  $d$ . The increase in  $f$  has an affect only at the margin, and the additional voluntary compliance induced at that margin is

welfare-improving. Raising  $\hat{c}$  through increases in  $\pi$  is different. Again, at the margin the additional voluntary compliance is beneficial provided  $\hat{c}$  remains below  $d$ . But there is a welfare cost of so doing, which is the increased frequency with which compliance is coerced at firms where compliance is welfare-reducing.

With  $f$  fixed at  $F$ , in raising  $\pi$  Equation 7 says that the agency trades-off the net benefits from increased voluntary compliance at the margin with the net disbenefits from increased coerced compliance above.

Given that the pursuit of reports is costless, the choice of  $\pi$  and  $f$  depends only upon the efficiency of the pattern of compliance that is realized. First-best would be described by compliance with probability 1 for  $c_i \leq d$ , zero otherwise. Against this benchmark we can note that in equilibrium firms can be divided into three classes. Those with compliance costs  $c_i \leq \hat{c}(\pi^*, F|\alpha)$  will comply pre-emptively, and that compliance will be socially desirable. The mid-range interval of firms  $\hat{c}(\pi^*, F|\alpha) < c_i \leq d$  comply only when coerced (that is, with probability  $\rho_\alpha(c_i)\pi$ ), and that compliance is socially desirable. The interval of high-cost firms  $c_i > d$  also comply only when coerced, but that compliance is socially undesirable. In summary,

**Remark 2** *Under Assumption  $\alpha$  optimal policy implements an equilibrium in which an interval of low cost firms comply and this is efficient. An interval of high cost firms comply too often, an interval of intermediate cost firms comply not often enough.*

The comparison between realized compliance probabilities given optimal policy (solid line) and the first-best pattern (broken line) is illustrated in Figure 2.

## 2.2. Scenario $\beta$ : Whistleblowing as a Social Act

The key characteristic of Assumption  $\alpha$  was that the prospective whistleblower's decision to report was dependent only upon the characteristics of the offence itself. The conscientiability or justification for a proposed act of non-compliance is increasing in the

cost of compliance, and the ‘conscience cleanser’ was assumed to have some threshold of conscientiability beyond which his internal ethic would require him to speak out.

Importantly the whistleblower did not take account of the characteristics of the enforcement regime in deciding whether or not to disclose.

The second behavioral assumption that we propose conceives of a utilitarian or social welfare basis for disclosure. In particular, we propose is that the whistleblower discloses if the expected *social* benefits from so doing are sufficiently large. *Assumption  $\beta$* : *Individual  $i$  discloses planned violation if and only if*

$$\pi \cdot (d - c_i) > \theta_i. \quad (8)$$

Such an individual will not report in cases where the social returns to complaint are comparatively small, but will when the returns are large.<sup>6</sup> Again, note that  $\theta_i$  might capture the private costs associated with becoming a whistleblower, under the assumption that those costs do not constitute a social loss (see penultimate paragraph in sub-section (1)).  $\theta_i$  is non-negative for all  $i$ . If  $\theta_i$  varies across individuals, or if there is some variability in the  $\theta$  that a representative individual applies in decision-making, then the probability that planned violation by firm  $i$  will be disclosed is described as some function of the expected social return to complaint (the left-hand side of 8),  $\rho_\beta(c_i, \pi)$ .

Observe that in deciding whether to report the individual takes account of the likelihood that report would lead to intervention (since that effects the welfare return) but does not take account of the penalty for non-compliance (which is a simple transfer).

We can note that  $c_i \geq d \implies \rho_\beta = 0$ ,  $\partial \rho_\beta(c_i, \pi) / \partial c_i \leq 0$  and  $\partial \rho_\beta(c_i, \pi) / \partial \pi \geq 0$ . The latter says that other things equal an increase in the rate at which reports are

---

<sup>6</sup>An alternative specification, similar in spirit, would have been to have individuals attaching differential weights to  $c$  and  $d$ , the relative weights offering a measure of the social conscience versus corporate loyalty of the individual.

pursued increases the probability that a given decision to violate will be reported. We might plausibly think of  $\rho_\beta(c_i, \pi)$  as an ogive in  $c_i$  - with  $\theta_i$  distributed according to some single-peaked distribution - but again nothing rests on this. Such an example is presented in Figure 3.

Firm  $i$  will comply voluntarily if

$$c_i \leq \rho_\beta(c_i, \pi) \cdot \pi \cdot (c_i + f) \quad (9)$$

As in (1.1) it is straight-forward to establish that the net expected benefits from compliance are everywhere (weakly) decreasing in  $c_i$ . Firm  $i$  complies voluntarily if  $c_i$  is less than some critical value  $\hat{c}(\pi, f|\beta) < d$ .  $\partial\hat{c}/\partial\pi$  and  $\partial\hat{c}/f$  are again positive. Everything again works in the ‘right’ direction, and we duplicate Remark 1.

**Remark 3** *Under Assumption  $\beta$  the net private benefits from compliance are monotonically decreasing in the cost of compliance. For a given enforcement environment  $\{\pi, f\}$  the realised probability of compliance is non-increasing in  $c_i$  and equals 1 if  $c_i \leq \hat{c}$ , equals  $\rho_\beta(\hat{c}, \pi)\pi$  otherwise.*

As in (1.1) the agency sets policy to minimize expected social loss subject to the behavioral Assumption  $\beta$ , and the self-selection behavior of firms. That is chooses  $0 \leq \pi \leq 1$  and  $f \leq F$  to minimize

$$SL(\pi, f|\beta) = \int_0^{\hat{c}} c_i g(c) dc + \int_{\hat{c}}^{\infty} [\rho_\beta(c_i, \pi)\pi c_i + (1 - \rho_\beta(c_i, \pi)\pi)d] g(c) dc \quad (10)$$

The terms are analogous to those in 4.

Since (a)  $c_i \geq d \implies \rho_\beta = 0$ , a complaint from a whistleblower will only arise when the cost of compliance is less than the damage avoided, (b)  $\rho_\beta$  is everywhere non-decreasing in  $\pi_i$  and invariant to  $f$  and, (c),  $\partial\hat{c}/\partial\pi$  and  $\partial\hat{c}/f$  are positive, it is apparent that the agency can do no better than set  $\pi = 1$  and  $f = F$ .

**Proposition 2** *If whistleblowing is motivated by social welfare (that is, under Assumption  $\beta$ ), then optimal policy will be characterised by (a) a maximal penalty and (b) a maximal inspection intensity.*

The logic is straight-forward. Whistleblowing only occurs when compliance is welfare-improving ((a) above), so  $\hat{c}$  can never exceed  $d$ . Since increasing  $\pi$  both encourages whistleblowing and increases  $\hat{c}$  ((b) and (c) above) the agency will wish to raise it as far as possible. Since increasing  $f$  increases  $\hat{c}$  ((c) above) the agency will wish to raise it, too, to its maximal level.

Naturally this result might be altered if inspections were costly.

In terms of the resulting pattern of compliance,

**Remark 4** *Under Assumption  $\beta$  optimal policy implements an equilibrium in which an interval of low cost firms comply and this is efficient. An interval of high cost firms never complies, and this is efficient. If  $\hat{c}(1, F|\beta) < d$  an interval of intermediate cost firms comply not often enough.*

The case in which  $\hat{c}(1, F|\beta) < d$  is illustrated in Figure 4. The only departure from first best in this example is due to the interval of firms between  $\hat{c}(1, F|\beta)$  and  $d$  who comply only if compelled.

### 2.3. Scenario $\gamma$ : Whistleblowing as Punishment, and the Disgruntled Employee

A third assumption that we might make about what motivates whistleblowers is that they are driven by a desire to punish the firm that employs them.

A certain proportion of staff might be “disgruntled employees”, unhappy for reasons unconnected with the firms planned non-compliance with the regulation, but opportunistic in blowing the whistle when so doing creates sufficiently substantial discomfort (*i.e.* cost) for their employer. This would correspond with what Alford

(2001:18) refers to as the popular conception of the whistleblower as ‘corporate malcontent’.

Alternatively, the individual’s desire to punish the firm may relate to the non-compliance decision itself. In a variety of experimental scenarios there is evidence that subjects are willing to incur private cost to punish others who have behaved in uncooperative ways, or in ways out of line with what they perceive to be socially appropriate.

Introspection may lead many to identify the punishment of wrongdoers as a motivation in itself, regardless of whether in that punishment generates welfare-improving changes in future incentives. Experimental evidence summarized by Fehr and Schmidt (1999) and others suggests that at least some fraction of people are willing to punish wrong-doers, even if such punishment is not in their own (apparent) self-interest. This is consistent with much of the evidence discussed in Sunstein (2003).

We operationalize this notion of punishment-motivated disclosure by making *Assumption  $\gamma$* : *An employee discloses planned violation if and only if the expected cost impact upon the firm is sufficiently large:*

$$\pi \cdot (c_i + f) > \delta_i \tag{11}$$

Again, the threshold  $\delta_i$  can be interpreted as varying across individuals, or as the motivation of a representative individual being subject to variation.

An implication of such an assumption is that the probability that planned violation will be disclosed is *increasing* in the cost of compliance. It is also now increasing in the two enforcement parameters. Prospective whistleblowers want to punish, and so increased expected penalty increases the likelihood of report.

We can specify this probability  $\rho_\gamma(c_i, \pi, f)$  where  $\partial\rho_\gamma/\partial c_i \geq 0$ ,  $\partial\rho_\gamma/\partial\pi \geq 0$  and  $\partial\rho_\gamma/\partial f \geq 0$ . An example is provided in Figure 5. Embedded here as a special case (which we will use below) is the degenerate one in which  $\delta_i = \delta$  for all  $i$ . In that case

$\rho_\gamma(c_i, \pi, f) = 1$  if  $c_i > \left[\frac{\delta}{\pi} - f\right]$ , equals 0 otherwise. This is superimposed as the bold line in the figure.

In contrast to earlier cases,  $f$  now matters not just for the firm's compliance incentives directly, but also indirectly through its impact on the reporting behavior of whistleblowers.

For a given enforcement environment  $\{\pi, f\}$  firm  $i$ 's net benefits to compliance are:

$$\eta \cdot \rho_\gamma(c_i, \pi, F) \cdot \pi \cdot (c_i + F) - c_i \quad (12)$$

Under assumptions  $\alpha$  and  $\beta$  net benefits were monotonically decreasing in  $c$  - an increase in  $c$  made pre-emptive compliance more costly, and also made non-compliance less likely to be reported. Here, however:

**Remark 5** *If whistleblowing is punishment motivated (that is, under Assumption  $\gamma$ ) the net private benefits from compliance may increase or decrease with cost of compliance.*

Differentiating 14 with respect to  $c$  gives:

$$\pi \cdot \left[ \frac{\partial \rho_\gamma}{\partial c_i} \cdot (c_i + f) + \rho_\gamma \right] - 1 \gtrless 0 \quad (13)$$

An increase in  $c$  makes pre-emptive compliance more costly (the second term), but now *increases* the likelihood that any given decisions of non-compliance will be reported (the positive first term). There is no general way to sign the whole.

Further, in contrast to the earlier cases:

**Proposition 3** *Realised non-compliance probabilities may be non-monotonic in cost of compliance.*

For simplicity we prove this by means of the simple degenerate example described by  $\delta_i = \delta$  for all  $i$ . Assume that parameters are such that

$$\frac{\pi}{(1 - \pi)} f > \frac{\delta}{\pi} - f.$$

In that case firms comply voluntarily if and only if

$$\frac{\pi}{(1-\pi)} \cdot f > c_i > \frac{\delta}{\pi} - f$$

For  $c_i$  outside this mid-range the firm will only comply if coerced, which will happen with probability  $\pi$  for  $c_i > \frac{\pi}{(1-\pi)}f$ , with probability zero for  $c_i < \frac{\delta}{\pi} - f$  (see Figure 4). In this degenerate case the firms above the mid-range do not comply pre-emptively because so doing is too expensive and it is more attractive to wait to see if they will be coerced. Firms below the range do not comply pre-emptively because their low costs places them below the point at which their non-compliance will be disclosed. In a non-degenerate version the low propensity to comply at the lower end would be driven by the comparatively low likelihood of non-compliance being reported.

If we restrict attention to settings in which the net benefits from compliance are monotonically decreasing in  $c_i$  (the additional up-front cost effect outweighs the increased probability of report effect) then firm  $i$  will pre-emptively comply if  $c_i$  is less than  $\hat{c}(\pi, f|\gamma)$ , implicitly defined by

$$\hat{c} = \rho_\gamma(\hat{c}, \pi, f) \cdot \pi \cdot (\hat{c} + f) \quad (14)$$

An argument similar to that used earlier can be used to establish that optimal policy will involve implementing some  $\hat{c} < d$ . We can no longer assert, however, that for any combination of  $\pi'$  and  $f'$  (such as the optimal one) generating  $\hat{c} < d$  that social loss will be decreasing in  $f$ . To see this note that

$$\frac{\partial SL(\pi', f'|\gamma)}{\partial f} = \frac{\partial \hat{c}}{\partial f} \cdot (1 - \pi \rho_\gamma(\hat{c}, \pi, f')) (\hat{c} - d) + \int_{\hat{c}}^{\infty} \pi \frac{\partial \rho_\gamma(\hat{c}, \pi, f')}{\partial F} \quad (15)$$

where  $\hat{c} = \hat{c}(\pi', f'|\gamma) < d$ . The first term here (under our retained assumption) is negative. The second term is ambiguous in sign and will be positive in the vicinity of an interior solution to the Agency's problem. Without further restriction on  $\rho$  there is no basis for signing the expression overall.

**Proposition 4** *If whistleblowing is punishment motivated, for any given probability that disclosure will be pursued the optimal penalty for planned violation may be less than maximal. This applies even if  $\hat{c} < d$ .*

Under assumptions  $\alpha$  and  $\beta$ , having established that equilibrium would involve  $\hat{c} < d$  it was simple to understand why the instrument  $f$  should be raised to its maximum level. Increases in  $f$  increased voluntary compliance at the  $\hat{c}$  (welfare-enhancing) without any cost. Under  $\gamma$ , however, increases in  $f$  now impact compliance away from the margin, increasing the frequency with which planned violations at high-cost firms are reported and compliance coerced. This effect may lead the regulator to wish to refrain from raising  $f$  to its maximal level.

### 3. Mixed Motivations and General Insights

Manipulating the basic model under three alternative assumptions about why employees become whistleblowers meant we could avoid having to make a definitive choice in a context where the informing literature lacks consensus.

In any real world population there will be a variety of motives for why individuals report planned wrong-doing in some contexts but not others. This implies that the  $\rho$  function is likely to be much more difficult to pin down. With mixed motivations there is, for example, no particular reason to think that it will be monotonic (a population made up of some  $\alpha$ -types and some  $\beta$ -types could easily generate a  $u$ -shaped  $\rho$  function, for example).

Allowing for mixed motives analytically would imply loss of tractability. Though the model is stylized and particular, the analysis points to two general lessons, both of which counsel that in designing an enforcement regime informed by the reports of whistleblowers, care needs to be taken to be clear about whistleblower motives.

First, the value of the information that whistleblowers bring to the enforcement

agency - and what the agency will wish to do with that information - depends upon the motives assumed to whistleblowers. If the motive is either conscience cleansing or welfarist ( $\alpha$  or  $\beta$ ) then whistleblowers will be more likely to report a planned act of violation at a firm with low compliance costs. These are the cases in which the agency would find it beneficial to coerce compliance. If, on the other hand, the motivation is punishment ( $\gamma$ ) then other things equal a case is more likely to be disclosed at a firm where compliance costs are high, precisely those cases where coerced compliance is of least (or even negative) social value.

Second, in adjusting the enforcement instruments -  $\pi$  and  $f$  - attention has to be paid to the change induced in the flow of disclosures. Again, the quantitative and qualitative response will depend upon whistleblower motives. Under conscience cleansing ( $\alpha$ ) the propensity for individuals to disclose isn't sensitive to the enforcement regime. Under the welfarist motivation ( $\beta$ ) the agency can induce employees to report more frequently by increasing the frequency with which reports are pursued. When the motive is punishment (and popular wisdom tends to portray disgruntled employees as the source of most disclosures) the flow of complaints can be increased by raising either  $\pi$  or  $f$ .

Naturally, alternative considerations may impact disclosure propensities in more subtle ways. In a repeated setting a firm's past performance, for example, may matter, though in ways that may not be straight-forward to predict. An individual may be able to 'forgive' or turn a blind eye to wrongdoing on a single occasion, but not when it becomes a pattern. Alternatively, individuals working in organizations with a long record of good behavior may come to regard the firm as implicitly contracted to continue working in that way. In that case a non-compliance decision may be perceived by an employee as a betrayal, and 'betrayal aversion' (Koehler and Gershoff (2002)) or the

related ‘betrayal heuristic’ (Sunstein (2003)) may make disclosure more likely.<sup>7</sup>

The model here has been stylized, and we have quite explicitly focussed on the role of alternative motivational assumptions. There are a number of ways in which model here could be developed. Two that are priorities in future work are, (a) to explore the role played by whistleblower rewards or ‘bounties’ and (b), to investigate how whistleblower-informed inspections might be combined with a more conventional enforcement program based on random inspections.

#### 4. Bibliography

Alford, C. Fred (2001). *Whistleblowers*, Cornell University Press: Ithaca NY.

Andreoni, James (1990). “Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving?” *Economic Journal* 100: 464-77.

Bauman, Zygmunt (1989). *Modernity and Ambivalence*, Cornell University Press: Ithaca NY.

Clark, Charles (1997). “Whistleblowers”. *CQ Researcher* 7: 1057-80.

Elliston, Frederick E. (1985). *Whistleblowing Research: Methodological and Moral Issues*, Praeger: New York.

Fehr, Ernst and A. Falk, (2002). “Psychological Foundations of Incentives”,

---

<sup>7</sup>“A betrayal of trust is likely to produce a great deal of outrage, and it should not be surprising that people will favor greater punishments for betrayals than for otherwise identical crimes.” (Sunstein (2003: 6)). Koehler and Gershoff (2002) provide experimental evidence of this. The critical issue in the conjecture that we make here - that a past record of good performance will make an employee more likely to want to report a given incident of wrongdoing - is the extent to which employees of socially responsible organisations feel that the organisation has implicitly committed to continued ‘good behavior’.

*European Economic Review* 46(4-5): 687-724.

Fehr, Ernst and Klaus M. Schmidt (1999). "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics*, 114(3): 817-68.

Glazer, Myron P. and Penina Glazer (1989). *The Whistleblowers*. Basic Books: New York.

Heyes, Anthony (1996). "Cutting Environmental Penalties to Increase Compliance", *Journal of Public Economics* 60(2): 251-65.

Heyes, Anthony (2001). "Honesty in a Regulatory Context - Good Thing or Bad?", *European Economic Review*.45: 215-32.

Hunt, Geoffrey (1997). "Whistleblowing" in *The Encyclopedia of Applied Ethics*, Academic Press: New York.

Jos, Phillip and Mark Tompkins (1989). "In Praise of Difficult People: A Portrait of the Committed Whistleblower". *Public Administration Review* 49: 552-61.

Jubb, Peter (1999). "Whistleblowing: A Restrictive Definition and Interpretation". *Journal of Business Ethics* 21: 77-94.

Koehler, Jonathan J. and Andrew D. Gershoff (2002). "Betrayal Aversion: When Agents of Protection Become Agents of Harm", *Organizational and Human Processes*.

Lifton, Robert J. (1986). *The Nazi Doctors: Medical Killing and the Psychology of Genocide*, Harvard University Press: Cambridge MA.

Livernois, John and Chris J. McKenna (1999). "Truth or Consequences: Enforcing Pollution Standards" *Journal of Public Economics* 71(3): 415-40.

Miceli, Marcia and Janet Near (1992). *Blowing the Whistle: The Organizational and Legal Implications for Companies and Employees*, Lexington Books: New York.

Miethe, Terrance (1999). *Whistleblowing at Work: Tough Choices in Exposing Fraud, Waste and Abuse on the Job*, Westview Press: Boulder, Colorado.

Mookerjee, Dilip and Ivan P. L. P'ng (1992). "Monitoring vis-a-vis Investigation in the Enforcement of Law", *American Economic Review* 82(3):556-63.

Milgram, Stanley. (1974). *Obedience to Authority*, Harper & Row: New York.

Sunstein, Cass R. (2000). *Behavioral Law & Economics*, Cambridge Series on Judgement and Decision-making, Cambridge University Press.

Sunstein, Cass R. (2003). “Moral Heuristics”, John M. Olin Law & Economics Working Paper #180, University of Chicago Law School.

Sunstein, Cass R., David Schkade and Daniel Kahneman (2000). “Do People Want Optimal Deterrence?”, *Journal of Legal Studies* 237.