An Economic Model of Whistleblowing

Anthony Heyes*  Sandeep Kapur
Royal Holloway College  Birkbeck College
University of London  University of London

February 2007

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. We adopt the methods of behavioral law and economics, in the spirit of recent work by Thomas Ulen, Cass Sunstein and others. The three competing theories are operationalized as alternative behavioral heuristics. 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, behavioral law and economics

JEL Codes: K42

A final version of this paper appeared in
Journal of Law, Economics & Organization, 2009
volume 25 (1), pp 157-182

*Corresponding author: Department of Economics, Royal Holloway College, University of London, Egham, Surrey TW20 0EX. email a.heyes@rhul.ac.uk. We are grateful for constructive comments to Thomas Ulen, Cees Withagen, Daniel G. Arce, John List, Bernard Sinclair-Desagne, Charles Mason, Mark Raymond, seminar participants at the LSE, Toulouse, Nottingham and the 2005 Industrial Organization Society Meetings (Atlanta); two anonymous referees and Editor Pablo Spiller. Errors are ours.
1 Introduction

‘Whistleblowing’ is a common phenomenon in many regulatory settings.\footnote{In the US explicit 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 there is no established economic model of it.\footnote{Apesteguia \textit{et al} (2004) consider the potential for a cartelist to inform a regulator about the wrongdoing of fellow cartel members in exchange for a reward. This is quite different to an agent (employee) blowing the whistle on a principal (employer) which is what we consider here. 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 body of research informs the assumptions that we make in setting-up the model that follows.}

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. This is a frequently-cited definition, and we adopt it here.

The biggest hurdle in modeling whistleblowing follows directly from part (1) of the definition. The benefits from the activity accrue, by definition, to other than the actor. As such the behavior is not easily explained under conventional assumptions about rational, self-interested behavior.

We proceed in the following way. We begin by summarizing our reading of the sociological and psychological research on why employees blow the whistle on law-breaking employers. We identify three broad schools of thought. In Section 2 we propose a regulatory setting. Though it has ‘generic’ features the story told here relates to the decision by an employer to comply with an environmental rule, such as correctly disposing of toxic wastes. The prospective whistleblower who inhabits the model is initially ascribed a general ‘motivation function’, within which is nested each of the three competing behavioral theories. The model is used to characterize optimal enforcement policy, and to highlight the substantial way in which policy prescriptions vary according to the analyst’s retained assumption regarding motivation. Section 3 relaxes the assumption that whistleblowers are fully knowledgeable: we assess the extent to which our conclusions regarding the enforcement policy are robust to the possibility of errors.

1.1 Whistleblower Motivation: Three Schools

What does research in other disciplines say about the motivations that might underpin the behavior of whistleblowers?

The first, and preponderant school of thought is that whistleblowing constitutes \textbf{conscience cleansing}. The hypothesis is that individuals have moral codes that condition their behavior. In our context, the moral code of an agent (employee) may preclude him from being complicit with, ‘going along with’ or ‘keeping quiet about’ immoral activities
proposed by his principal (employer).

Being complicit with an immoral activity corrupts the self and “... (w)histleblowers disclose because they dread living with a corrupted self more than they dread the other outcomes” (Alford (2001: 90)). The agent will not comply unquestioningly with an instruction issued by the principal but will, rather, subject that instruction to a moral screen. Authors ascribing to and providing evidence of this view include Jos and Tompkins (1989) and citations therein.

Determination of the ‘morality’ or otherwise of a particular activity may depend on cultural, religious and other forces that we do not need to detail here. Importantly, however, the decision by an employee to disclose any planned illegal activity of an employer depends only upon the characteristics of the planned activity itself. If an activity is sufficiently ‘bad’ - where the dimension along which ‘badness’ is defined, and the sufficiency threshold that a particular agent will apply, are as yet unspecific - then the agent is morally unable from being complicit with and feels compelled to speak out to cleanse his conscience. Significantly for policy the whistleblowing decision is independent of the characteristics of the enforcement regime. This detachment will be captured by the ‘Assumption α’ that we introduce below. Importantly the regulator in this case will not be able to change the pattern of whistleblowers coming forward by changing the enforcement regime.

Variants of this theory are popular amongst social psychologists and sociologists and are consistent with the voluminous case study evidence that people engage in whistleblowing even where they know that the likelihood that they will be ‘heard’ - that their blowing the whistle will impact outcomes - is minimal.4

The psychological analysis of the need to conscience cleanse is related to the cognitive process of ‘doubling’ (Lifton (1986)), and thereafter to mental disfunction. In psychospeak all individuals have different components to their personalities, or multiple ‘selves’. Different individuals differ in the extent to which they are able to insulate their work self from their whole self, and to tolerate behavior from the former which would be un-

³Alford’s theory of ‘narcissism moralized’ is closely related to sense of shame. “Though not all whistleblowers use the language of shame, 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).

⁴It also bears a family resemblance to the proposal by Sunstein (2003) that peoples taste with regard to retribution are driven by an ‘outrage heuristic’:

“Peoples intuitions about punishment seem quite disconnected from the consequences of punishment. People want to punish the corporation as if it were a person. They do not inquire into the consequences of punishment. Punishment judgements are rooted in a heuristic, to the effect that penalties should be a proportional to the outrageousness of the act” (Sunstein (2003:5)).

Significantly outrageousness of an act may not be closely correlated with the forward-looking costs and benefits of intervention. He offers experimental evidence that subjects levy the same punitive damages in hypothetical contexts in which though the past violation was common, though the next period implications of penalty levels differ.
conscionable to the latter. “Doubling is a sophisticated emotional and cognitive act, one that whistleblowers have comparative difficulty performing. In this sense they might be regarded as dysfunctional actors in modern society” (Alford (2001: 73)).

An alternative and plausible hypothesis in this setting is that individual moral motivation could have a welfarist basis such that the whistleblowing decision will incorporate a cost-benefit element.

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. Our adoptive definition of whistleblowing implies that (a) is necessarily satisfied, whilst (b) implies both that the whistleblowing decision is forward-looking and that the agent applies some sort of cost-benefit test. This said, he muddies the divide between conscience cleansing and welfarist behavior: “...all the well-rehearsed arguments for and against utilitarian calculation could be invoked at this point. The moral codes of some may lead them to take the view that it reasonable to make a disclosure simply and only because ‘it is the right thing to do’ regardless of whether harmful consequences are 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).

What we are referring to as a welfarist motivation is consistent with the popular economic characterization of altruism as ‘other-regarding behavior’ - that the utility of others may enter as an argument in the agent’s own utility function - and the closely related notion of ‘warm glow’ (Andreoni (1990)).

Notice that in this case the disclosure decision is itself sensitive to the enforcement regime - in particular the likely responsiveness of the regulatory agency to information made available to it through this channel.

An third view - and one popular in the ‘airport’ literature - is that whistleblowers are punishment motivated. What Alford calls ‘hysterical malcontents’, primarily driven by a desire to discomfort the organization that employs them.

A certain proportion of staff might simply be ‘disgruntled employees’, unhappy for reasons unconnected with the firms planned non-compliance with the regulation, but opportunist in blowing the whistle when so doing creates sufficiently substantial discomfort (cost) for their employer. Alternatively, the agent’s desire to punish his principal may relate to the non-compliance decision itself. There is compelling experimental evidence that a portion of subjects 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 nor the interest of prospective social welfare (Fehr and Schmidt (1999), Fehr and Falk (2002)).

Arguments are sometimes made in favor of businesses writing an explicit code of conduct so that employees can see the difference between personal and business ethics. I am grateful for Dan Arce for bringing this point to my attention.
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. Our detailed reading of the non-economic literatures suggests three preponderant schools of thought regarding whistleblower motivation: (α) conscience-cleansing (an unwillingness or inability to be complicit with morally objectionable activity), (β) social motivation (based on a forward-looking calculation of the expected social costs and benefits of disclosure) and (γ) the desire to punish.\(^6\) Opting for one of the three schools for the purposes of policy analysis would be misleading. In what follows we convert each of α, β and γ into a behavioral heuristic, an algorithm that determines whistleblowing behavior, and explore policy design issues under each heuristic.

2 A Regulatory Setting

We propose a generic, single-shot compliance/enforcement setting in which whistleblowing can be embedded.

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-shot setting requires there be a lag between a decision to violate a law, and the moment at which the violation occurs and harm results. 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 waste illegally rather than using a licensed contractor. Before harm – in this case environmental damage – occurs 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 that the regulator has time to respond to the report and prevent the plans for violation being executed (as in Heyes (1996)).

Consistent with (2) 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. These are the unshredded memos, e-mail exchanges and taped conversations of high-profile whistleblowing cases. Alternatively, the agent may not himself hold evidence, but be able to direct the investigator to the place in the organization where the evidence can be found. This is analogous to the assumption made by Prendergast (2003) in the context of consumer complaints regarding treatment by public bureaucracy. The regulator can summarily dismiss as spurious a complaint not supported by evidence.

Consider the following stylization. Firms are required to adhere to a regulatory standard. Non-compliance by a firm imposes external damage \(d\), assumed to be the same

\(^6\)Arce (2005) offers an evolutionary story in which an individual who blow the whistle may be regarded within the organization as a ‘rat’ or a ‘hero’, depending upon (endogenous) organizational norms.
across all firms. Compliance costs, on the other hand, differ across firms: firm $i$’s cost of compliance $c_i$ is drawn from a probability distribution on the interval $[0, \infty)$, with distribution function $G(c)$. Here $G(c)$ denotes the probability that a randomly-drawn firm has compliance cost $c$ or lower.\(^7\)

At time $t = 1$ the firm (which we can equally think of as a single principal) makes a plan either to ‘comply’ or ‘violate’. That decision may be known to a number of employees, but to side-step complications due to multiple potential whistleblowers we assume that the content of the decision is known to a single agent. The agent holds evidence of the decision that is externally verifiable. To begin with, we assume that the agent also knows $c_i$ accurately, though information on costs is unverifiable. We consider the implications of relaxing this assumption of full knowledgeability in Section 3.

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. With probability $\pi$, the regulator ‘hears’ the report, and sends an inspector to ‘visit’ and with probability $(1 - \pi)$ it ignores the report. Here $\pi$ is a pre-determined and publicly-observed policy variable that captures the responsiveness of the regulator to whistleblowers. An enforcement ‘visit’ entails two things. First, the firm is required to comply. Second, it is subject to a fine $f$ for intent to violate. Notice that whilst we will assume our regulatory agency is welfare-motivated it is restricted in the instruments it has available. It is not allowed to condone and allow a violation to go ahead in exchange for a fee, for example. The two policy variables $\{\pi, f\}$ define the enforcement environment chosen by the regulator.

For simplicity we will proceed as if are no pecuniary costs or benefits to the whistleblower. This assumption warrants some discussion. There are numerous legal protections for whistleblowers in the US and Europe,\(^8\) and the most restrictive assumption we could make would be that those protections worked perfectly. That would not be realistic, however - there is plenty of evidence that legal protection is less than full and that whistleblowing imposes a significant costs on individuals. However, the key costs upon the individual are usually redistributive in nature – the nicer task, promotion, preferred office, salary increase etc.. The social costs appear slight. In terms of behavior, then, the private costs can be regarded as captured in the $\mu_i$, $\theta_i$, $\delta_i$ ‘thresholds’ that we will introduce in (2.2) through (2.4). In terms of welfare analysis they are transfers and can be ignored.

In the most general case we can allow the decision to disclose to depend upon $c_i$, $\pi$ and $f$, the only three ‘moving parts’ in the model. We can then define a general ‘motivation function’ that describes the probability that an individual drawn at random from the population of prospective agents will disclose when faced with a particular configuration

\(^7\)We assume that $G(d) < 1$, or that at least some firms have compliance costs that exceed $d$.

\(^8\)Examples of legislation that prohibit discharging or discriminating against an employee who reports an anticipated violation of the Act include the Clean Air Act Amendments (1977), 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.
of these variables, $\rho(c_i, \pi, f)$.

### 2.1 Policy Problem: General Form

We assume that the regulator acts to minimize social loss, defined as the sum of expected compliance costs and external damage.$^9$ His problem, then, is to choose $0 \leq \pi \leq 1$ and $f \leq F$ to minimize:

$$SL(\pi, f) = \int_{\text{Comply}} c_i dG(c) + \int_{\text{Violate}} [\rho(c_i, \pi, f)\pi c_i + (1 - \rho(c_i, \pi, f))\pi] dG(c)$$

(1)

where firm $i$ violates if and only if its cost of compliance exceeds the expected penalty associated with non-compliance:

$$c_i > \rho(c_i, \pi, f)\pi(c_i + f).$$

The notation attached to the first (second) integral in (1) denotes aggregation across the set of firms that makes an initial decision to comply (violate).

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

Analysis of the general problem (available in an earlier version) allows remarks to be made with respect to families of functions but at the cost of tractability. Here we choose to work with three alternative theories regarding whistleblower motivation, each associated with one of the ‘schools of thought’ identified in the previous section. We add a subscript $\rho_\Delta, \Delta \in \{\alpha, \beta, \gamma\}$, to denote the theory ‘in play’.

### 2.2 Theory $\alpha$: Whistleblowing is Conscience Cleansing

The conscience-cleansing school of thought proposes that the whistleblower’s decision is based solely upon the characteristics of the proposed offence and, in particular, on the moral defensibility of the employer’s decision.

How do we operationalise the ideas around ‘conscience cleansing’ that were introduced in Section 1? We need some measure of the moral defensibility of the act of violation. The prospective violator, recall, chooses to impose external environmental damage of value $d$ in order to save private costs $c_i$. Suppose that $c$ were a very large number, then it would require very large self-sacrifice on the part of the employer to avoid imposing the damage. A reasonable man or woman might take a view that non-compliance with the regulation in that case would be more defensible than in the case in which $c$ were a very small number.

$^9$There are alternative assumptions that we could have made. The regulator could for instance have been entrusted with a ‘stewardship’ role and so act to minimize expected external damage. But welfare maximization is a common approach, and allows us to focus on optimal policy.
The relationship between \( c \) and \( d \) embodies the rate at which the employer trades-off the interests of self versus others in his decision. Since the impact of non-compliance is fixed at \( d \), we can simply regard \( c_i \) as a measure of the defensibility of the firms decision to violate. We might then assume that an individual has some threshold of defensibility beyond which his conscience compels him to speak out, a threshold that can be expected to vary across individuals.

The construct is easiest understood by contemplating the extremes. If my employer chose not to comply despite having a \( c \) equal to zero then that to me (and perhaps most people) would be a particularly outrageous/inexcusable/indefensible decision - showing an unwillingness to act to protect the interests of others, even though such action would be personally costless - 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’. At the upper extreme, if \( c \) were to equal infinity, then non-compliance would be much more excusable - indeed unavoidable.

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

\[
c_i < \mu_i.
\] (2)

It is plausible to suppose that the threshold \( \mu \) will vary privately across individuals. Let the probability distribution of this attribute in the population of potential whistleblowers be given by \( \Phi(\mu) \). 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,

\[
\rho_\alpha(c_i) = 1 - \Phi(\mu(c_i)).
\]

It is important here - recalling the discussion of conscience cleansing presented in Section 1 - to note that the characteristics of enforcement environment, \( \{\pi, f\} \), do not impact the disclosure decision. The whistleblower takes a view on the morality of his employers choice, and decides whether or not his conscience allows him to ‘live’ or be complicit with that choice, without reference to the enforcement environment.

In Figure 1 we represent \( \rho_\alpha(c_i) \) as an ogive, consistent with \( \mu(.) \) being drawn from some continuous single-peaked distribution, but nothing rests on this.\(^{10}\) Significantly, the probability of disclosure does not depend upon the characteristics of the enforcement regime.

Firm \( i \) will comply voluntarily if

\[
c_i \leq \rho_\alpha(c_i)\pi(c_i + f).
\] (3)

\(^{10}\)We can pick out as a special cases 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. Note that this is NOT the same as an assumption that the disclosure decision is made to maximise welfare – that would require the individual take into account the probability that any disclosure would be acted upon, \( \pi \).
Compliance incentives – the net expected benefits from compliance – decrease monotonically 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 penalized. For values of $c_i$ close to zero net benefits must be positive. Therefore, the firm will comply if and only if its cost is less than some critical value $\hat{c}(\pi, f|\alpha)$, implicitly defined by

$$\rho_\alpha(\hat{c})\pi(\hat{c} + f) - \hat{c} = 0. \quad (4)$$

It is straight-forward to confirm that $\hat{c}$ is increasing in both $\pi$ and in $f$. 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 $c_i$. It equals 1 if $c_i \leq \hat{c}$, equals $\rho_\alpha(c_i)\pi$ otherwise.

Firms with compliance costs less than $\hat{c}$ comply voluntarily. Firms with costs above $\hat{c}$ comply only if coerced, which occurs with probability $\rho_\alpha(c_i)\pi$. Otherwise the firm will be left non-compliant, imposing external damage $d$. The regulator’s problem is to choose $0 \leq \pi \leq 1$ and $f \leq F$ to minimize:

$$SL(\pi, f|\alpha) = \int_0^{\hat{c}} c_i dG(c) + \int_{\hat{c}}^{\infty} [\rho_\alpha(c_i)\pi c_i + (1 - \rho_\alpha(c_i)\pi) d] dG(c). \quad (5)$$

We begin with the observation that a welfare-motivated regulatory agency will never choose to implement a regime that induces $\hat{c} \geq d$. Such a regime induces compliance beyond that which is justified by external damage $d$ and is welfare-reducing.\[11\] Given this, consider any pair $\{\pi', f'\}$ such that $\hat{c} < d$. Then,

$$\frac{\partial SL(\pi', f'|\alpha)}{\partial f} = \frac{\partial \hat{c}}{\partial f} (1 - \rho_\alpha(\hat{c})\pi)(\hat{c} - d) < 0. \quad (7)$$

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

$$\frac{\partial SL(\pi^*, F|\alpha)}{\partial \pi} = 0 = \frac{\partial \hat{c}}{\partial \pi} (1 - \rho_\alpha(\hat{c})\pi^*)(\hat{c} - d) + \int_{\hat{c}}^{\infty} \rho_\alpha(c_i)(c_i - d) g(c) dc \quad (8)$$

\[11\] A formal proof is by contradiction. Consider any regime $\{\pi', f'\}$ such that $\hat{c}(\pi', f'|\alpha) \geq d$. Differentiating (5) gives:

$$\frac{\partial SL(\pi', f'|\alpha)}{\partial \pi} = \frac{\partial \hat{c}}{\partial \pi} (1 - \rho_\alpha(\hat{c})\pi)(\hat{c} - d) + \int_{\hat{c}}^{\infty} \rho_\alpha(c_i)(c_i - d) dG(c). \quad (6)$$

The first term on the right-hand side is non-negative and the second is 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 $\hat{c}(\pi^*, f^*|\alpha) < d$. 


where $\hat{c} = \hat{c}(\pi^*, F|\alpha)$. The first composite term on the right-hand side of (8) is negative: An increase in inspection probability extends the interval of firms that plan to comply, which at the $\hat{c} < d$ margin increases welfare. The second term captures the social loss associated with the increased frequency with which compliance will be coerced at firms with compliance costs above $\hat{c}$, which must be positive in the vicinity of an interior solution.

If the expression in (8) 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 assuming 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** Assume that (a) whistleblowing is motivated by conscience cleansing and (b) the regulator’s enforcement instruments are potent enough that $\hat{c}(1, F|\alpha) > d$. Then optimal policy will be characterised by (i) a maximal penalty and (ii) 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), argues that to achieve any particular level of compliance the agency should set fines maximally and adjust inspection intensity correspondingly (see Heyes (2000) for a survey of the literature that has developed around the Becker model). His observation is based on inspections being costly, fines not. A similar result holds here, where inspections are costless, 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 qualitative pattern of actual compliance that results.

In our model the agency can increase compliance by raising either its responsiveness $\pi$ to whistleblowers’ or by raising the penalty $f$ for non-compliance. Doing so through increases in $f$ will always be beneficial provided $\hat{c}$ remains below $d$. An increase in $f$ has an effect 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. Equation (8) specifies that, with $f$ fixed at its maximal value $F$, the agency sets $\pi$ to trade-off the net benefits from increased voluntary compliance at the margin with the net disbenefits from increased coerced compliance.

Given that the pursuit of reports is costless in our model, 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. This serves to illustrate Remark 2.

### 2.3 Theory $\beta$: Whistleblowing is a Social Act

A key implication of conscience cleansing as a theory is that the prospective whistleblower’s decision to report is dependent only upon the characteristics of the offence itself. Importantly the whistleblower is not assumed to be balancing forward-looking costs and benefits – which will depend upon the characteristics of the enforcement regime – in making his disclosure decision.

The second behavioral theory proposed conceives of a utilitarian or social welfare basis for disclosure. We operationalize this by assuming that an individual will disclose if the expected social benefits from so doing are sufficiently large. **Assumption $\beta$**: Individual $i$ discloses planned violation if and only if

$$\pi(d - c_i) > \theta_i.$$  \hspace{1cm} (9)

Such an individual will report in cases where the social returns to so doing are sufficiently large. Observe that the potential whistleblower takes account of the likelihood that his report would lead to intervention (since that affects the welfare return) but does not take account of the penalty for non-compliance (which is a simple transfer). Again, $\theta_i \geq 0$ is an individual-specific threshold that captures what individual $i$ regards as ‘sufficient’. It might reflect the private costs associated with whistleblowing. These are not made explicit here (but recall penultimate paragraph in 2.0). If the distribution of the attribute $\theta$ in the population is given by $\Phi_\theta(.)$, the probability that planned violation by firm $i$ will be disclosed is given by $\Phi_\theta(\pi(d - c))$. We denote this as $\rho_\beta(c_i, \pi)$. 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.

Given that $\theta_i \geq 0$, we have $\rho_\beta(c_i, \pi) = 0$ for $c_i \geq d$: whistleblowers motivated by social welfare would not disclose planned violations where compliance costs exceed the external damage. More generally, $\partial \rho_\beta(c_i, \pi) / \partial c_i \leq 0$ so that firms with higher compliance costs face lower risk of disclosure. Further $\partial \rho_\beta(c_i, \pi) / \partial \pi \geq 0$: other things equal an increase
in the regulator’s responsiveness to disclosures increases the probability that decisions to violate will be reported.

Firm \( i \) will comply voluntarily if

\[
c_i \leq \rho_\beta(c_i, \pi)\pi(c_i + f)
\]  

(10)

As in (2.2) it is straightforward 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 \( \hat{c}(\pi, f) < d \). Once again, \( \hat{c} \) is increasing in \( \pi \) and in \( f \). 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 realized probability of compliance is non-increasing in \( c_i \): it equals 1 if \( c_i \leq \hat{c} \), and equals \( \rho_\beta(c_i, \pi)\pi \) otherwise.

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

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

(11)

The terms are analogous to those in (5). Recall that \( \rho_\beta = 0 \) for \( c_i > d \): a complaint from a whistleblower arises only arise when the cost of compliance is less than the damage avoided. As \( \rho_\beta \) is everywhere non-decreasing in \( \pi \) and invariant to \( f \) and as \( \partial \hat{c}/\partial \pi \) and \( \partial \hat{c}/f \) are positive, it is clear that the agency can do no better than set \( \pi = 1 \) and \( f = F \).

**Proposition 2** Assume that whistleblowing is motivated by social welfare. Then optimal policy will be characterised by (a) a maximal penalty and (b) a maximal inspection intensity.

The logic is straight-forward. Whistleblowers in this scenario have the same interests as the regulator – prospective social welfare. Whenever a report is forthcoming it will relate to a case in which \( c_i < d \) so the regulatory agency should pursue it (\( \pi^* = 1 \)). No firm with a compliance cost greater than \( d \) will ever be induced to comply because it knows it will never be reported. The regulator will wish to bring \( \hat{c} \) as close to \( d \) as he can, implying that he will set \( f \) as high as he can. Naturally this result might be altered if inspections are 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 3. 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 coerced.

### 2.4 Theory \( \gamma \): Whistleblowing as Punishment by Disgruntled Employees

Recall the third behavioral theory, namely that disclosure is driven by the desire of a subset of individuals to damage the firm that employs them. We operationalize this notion of punishment-motivated disclosure by proposing that the agent will disclose only if the damage he can do to the firm is sufficiently large.

**Assumption \( \gamma \):** An employee discloses planned violation if and only if the expected cost impact upon the firm is sufficiently large:

\[
\pi(c_i + f) > \delta_i \tag{12}
\]

Again, the threshold \( \delta_i \) can be interpreted as varying across individuals, with distribution function \( \Phi_{\delta}(\cdot) \). The probability that a planned violation will be disclosed equals \( \Phi_{\delta}(\pi(c_i + f)) \). We denote this as \( \rho_{\gamma}(c_i, \pi, f) \).

An immediate implication of this assumption is that the probability that a planned violation will be disclosed is increasing in the cost of compliance, or that \( \partial \rho_{\gamma}/\partial c_i \geq 0 \). It is also now increasing in both enforcement parameters: \( \partial \rho_{\gamma}/\partial \pi \geq 0 \) and \( \partial \rho_{\gamma}/\partial f \geq 0 \).

Prospective whistleblowers want to punish, and so increased expected penalty increases the likelihood of report. 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.

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 zero otherwise.

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

\[
\rho_{\gamma}(c_i, \pi, f)\pi(c_i + f) - c_i \tag{13}
\]

Recall that under Assumptions \( \alpha \) and \( \beta \), net benefits to compliance 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. However, here differentiating (13) with respect to \( c \) gives:

\[
\pi \left[ \frac{\partial \rho_{\gamma}(c_i, \pi, f)}{\partial c_i} (c_i + f) + \rho_{\gamma}(c_i, \pi, f) \right] - 1 \geq 0. \tag{14}
\]
An increase in $c$ makes pre-emptive compliance more costly, but now increases the likelihood that any given decision not to comply will be reported. The impact on net benefit is ambiguous.

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

In further contrast to the earlier cases:

**Proposition 3** Realized 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

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

In that case firms comply voluntarily if and only if

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

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 $c_i < \frac{\delta}{\delta - f}$ (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 $c_i$ (that is, where 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

$$\hat{c} = \rho_\gamma(\hat{c}, \pi, f)\pi(\hat{c} + f)$$  \hspace{1cm} (15)

An argument similar to that used earlier can establish that optimal policy will involve implementing some $\hat{c} < d$. We can no longer argue, 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
\[
\frac{\partial S L(\pi', f'|\gamma)}{\partial f} = \frac{\partial \hat{c}}{\partial f} (1 - \pi \rho(\hat{c}, \pi, f'))(\hat{c} - d) + \int_{\hat{c}}^{\infty} \pi \frac{\partial \rho(\hat{c}, \pi, f')}{\partial f}(\epsilon - d)
\]  

(16)

where \( \hat{c} = \hat{c}(\pi', f'|\gamma) < d. \) The first term, under Assumption \( \gamma, \) is negative. The second term is ambiguous in sign and will be positive in the vicinity of an interior solution to the regulator’s problem. Without further restriction on \( \rho \) this cannot be signed. This leads naturally to the following:

**Proposition 4** If whistleblowing is punishment motivated, optimal policy may involve (i) a less-than-maximal penalty and/or (ii) a less-than-maximal inspection intensity.

The unusual element here is that \( f \) may be less than maximal. Under theories \( \alpha \) 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 maximal 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 setting \( f \) at the highest level.

### 3 Whistleblowers with Noisy Information

Our analysis thus far assumes that whistleblowers are knowledgeable. This ignores the very real possibility that whistleblowers base their disclosure decision on faulty or partial information about the underlying firm-specific parameters. In terms of the moving parts of our model, it could be that whistleblowers under- or over-estimate the compliance cost of the firm. That way, for example, a conscience cleansing employee might mis-evaluate the egregiousness of his employer’s actions, and make his report/not report decision on the basis of that mis-evaluation.

We develop this possibility by assuming that the potential whistleblower’s observation of compliance cost is noisy: a whistle-blower observes

\[
\hat{c}_i = c_i + \varepsilon_i,
\]

(17)

where \( \varepsilon \) denotes noise. Negative values of \( \varepsilon \) reflect situations in which the whistleblower underestimates the true compliance cost while positive values reflect overestimation.\(^{12}\) For tractability, we assume that the distribution of the noise term, \( E(\varepsilon) \) is common knowledge and is independent of the firm’s true cost \( c_i. \) We also assume that the distribution of noise is independent of the prospective whistleblowers’s characteristics as captured by distributions \( \Phi_\mu, \Phi_\theta \) and \( \Phi_\delta. \)

\(^{12}\)We restrict the distributions of noise terms and true costs so that noisily observed costs are necessarily non-negative.
Whistleblowers’ disclosure decisions now depend on the noisily-observed value of compliance costs. Errors in observation that lead them to over- or under-estimate costs, depending on the scenario, alter the probability of disclosure. The likelihood of disclosure depends on the realization of the error term, so that in general we now have $\rho(c_i, \pi, f|\varepsilon_i)$.

The firm does not observe the specific realization of error terms but knows its distribution. For any level of compliance cost $c_i$, it evaluates the likelihood of disclosure as

$$\rho(c_i, \pi, f) = \int \rho(c_i, \pi, f|\varepsilon)dE(\varepsilon).$$

(18)

As we argue below, noisy observation of compliance cost may raise or lower the probability of disclosure relative to the case without noise. As before, a firm’s incentive to comply pre-emptively depends on its estimated risk of disclosure. It will comply voluntarily iff

$$c_i \leq \rho(c_i, \pi, f)\pi(c_i + f).$$

(19)

Social welfare varies with actual reports of violations, so depends directly on the realization of $\varepsilon$. For any realization of $\varepsilon$, define

$$SL(\varepsilon) = \int_{Comply} c_i dG(c) + \int_{Violate} [\rho(c_i + \varepsilon)\pi c + (1 - \rho(c_i + \varepsilon)\pi)d\varepsilon dG(c).$$

(20)

The regulator’s problem is to choose $0 \leq \pi \leq 1$ and $f \leq F$ to minimize expected social loss (that is, its expectation across possible realizations of the error term):

$$SL(\pi, f) = \int SL(\varepsilon)dE(\varepsilon).$$

(21)

We adapt this general framework to each of the environments studied earlier, allowing us to examine how noise in whistleblowers’ information affect the optimal enforcement policy.

3.1 Assumption $\alpha$

Under Assumption $\alpha$, an agent with noisy information reports a planned violation if and only if $\tilde{c}_i < \mu_i$. Given distribution $\Phi_\mu$ of characteristic $\mu$, the probability of disclosure conditional on noise $\varepsilon$ is

$$\rho_\alpha(c_i|\varepsilon_i) = 1 - \Phi_\mu(c_i + \varepsilon_i).$$

(22)

For this case a potential whistleblower who underestimates compliance cost is more likely to report a violation than one who overestimates it.

For any level of true compliance cost $c_i$, the firm computes the likelihood of disclosure by taking expectation over error terms

$$\rho_\alpha(c_i) = 1 - \int \Phi_\mu(c_i + \varepsilon)dE(\varepsilon).$$

(23)
In general, relative to the case without noise, the noisy observation of compliance cost may raise or lower the perceived likelihood of disclosure. Clearly, if the agent always underestimates compliance costs (that is, if the support of distribution $E$ is negative), the expected probability of disclosure is unambiguously greater. On the other hand, systematic overestimation of compliance costs lower the risk of disclosure. The case where errors are unbiased – consider a distribution of error terms with zero mean and positive variance – has ambiguous effects: If $\int \Phi_{\mu}(c_i + \varepsilon_i) dE(\varepsilon)$ is less than $\Phi_{\mu}(c_i)$ if $\Phi_{\mu}$ is locally concave. In that case, noise will cause expected probability of disclosure to rise. For convex ranges of the distribution $\Phi_{\mu}$, noise will cause expected likelihood of being reported to fall.

As in the case without noise, net expected benefit from compliance is decreasing in $c_i$ so that firm $i$ will comply voluntarily if and only if $c_i$ is less than some critical value $\hat{c}(\pi, f)$. This critical value depends on the distribution of noise terms but does not depend on the actual realization of $\varepsilon$. As before it is straightforward to confirm that $\partial \hat{c}/\partial \pi$ and $\partial \hat{c}/\partial f$ are both positive.

Social loss depends on reported violations, so varies with the realization of $\varepsilon$. The regulator must choose enforcement parameters $\{\pi, f\}$ to minimize $SL(\pi, f|\alpha) = \int SL(\varepsilon|\alpha) dE(\varepsilon)$ where

$$SL(\varepsilon|\alpha) = \int_0^{\hat{c}} c dG(c) + \int_{\hat{c}}^{\infty} [\rho_\alpha(c + \varepsilon)\pi c + (1 - \rho_\alpha(c + \varepsilon)\pi)] dG(c).$$

(24)

Replicating the familiar argument we can show that the welfare-motivated agency will never implement a regime that induces $\hat{c} \geq d$. Then, for any pair of enforcement parameters that satisfy the restriction $\hat{c} < d$, the partial derivative

$$\frac{\partial SL(\varepsilon|\alpha)}{\partial f} = \int \frac{\partial \hat{c}}{\partial f} [(1 - \rho_\alpha(\hat{c}(\varepsilon), \pi)(\hat{c} - d)] dE(\varepsilon) = \frac{\partial \hat{c}}{\partial f} [(1 - \rho_\alpha(\hat{c})\pi)(\hat{c} - d)]$$

is negative, so that the corner solution with the maximal penalty $f^* = F$ is optimal. An interior solution to the agency’s choice of $\pi$ is given by

$$\frac{\partial SL(\pi^*, F|\alpha)}{\partial \pi} = 0,$$

which now requires

$$\frac{\partial \hat{c}}{\partial \pi} [(1 - \rho_\alpha(\hat{c})\pi^*)(\hat{c} - d)] + \int_{\hat{c}}^{\infty} \rho_\alpha(c|\varepsilon)(c - d)dG(c) dE(\varepsilon) = 0.$$

(25)

As in the noiseless case, the first composite term in the expression is negative and the second must be positive at any interior solution for $\pi$. The logic of Proposition 1 – that if whistleblowing is motivated by conscience cleansing and if enforcement instruments are sufficiently potent, the optimal policy involves maximal penalty but less than maximal responsivenss to whistleblowing – carries over to the case where whistleblowing is based on noisy information.
3.2 Assumption $\beta$

Under Assumption $\beta$, an agent with a noisy observation $\tilde{c}_i = c_i + \varepsilon$ reports a planned violation if and only if $\pi(d - \tilde{c}_i) > \theta_i$. Given distribution of characteristics $\Phi_{\theta_i}$, the probability of disclosure conditional on $\varepsilon$ is

$$
\rho_\beta(c_i, \pi | \varepsilon_i) = \Phi_{\theta_i}(\pi(d - (c_i + \varepsilon_i))).
$$

The firm evaluates the expected likelihood of disclosure as

$$
\rho_\beta(c_i, \pi) = \int \Phi_{\theta_i}(\pi(d - (c_i + \varepsilon)))dE(\varepsilon).
$$

As under Assumption $\alpha$ errors in observing compliance costs may increase or decrease the expected likelihood of being reported relative to the case without noise, depending on the distribution of error terms and the local concavity /convexity of the distribution function $\Phi_{\theta_i}$.

Also, as under Assumption $\beta$ in the absence of noise, the net benefit of compliance is decreasing in $c$ so that voluntary compliance is rational for all $c$ less than some critical value $\hat{c}(\pi, f|\beta)$. This threshold value depends on the distribution of noise terms and not on the actual realization of $\varepsilon$. As before it is straightforward to confirm that $\hat{c}$ is increasing in $\pi$ and in $f$.

Social loss depends on actual disclosures, so varies with $\varepsilon$. Define

$$
SL(\varepsilon|\beta) = \int_0^{\hat{c}} c dG(c) + \int_{\hat{c}}^{\infty} \left[ \rho_\beta(c + \varepsilon, \pi ) \pi c + (1 - \rho_\beta(c + \varepsilon, \pi ) \pi ) d \right] dG(c).
$$

The regulator minimizes the expected social loss, $SL(\pi, f|\beta) = \int SL(\varepsilon|\beta)dE(\varepsilon)$. Revisiting arguments established earlier under Assumption $\beta$ in the absence of noise, we can confirm that (i) optimally-chosen enforcement parameters must implement a regime that has $\hat{c} < d$; (ii) the maximal penalty is still optimal; (iii) however, maximal responsiveness to whistleblowing may no longer be optimal.

To appreciate the last part, note that the regulator realizes that it is no longer the case – as was in the absence of noise – that complaints arise only for $c_i < d$. Specifically, there is a range of possibilities such $c_i > d$ so that optimally enforcement is not warranted, yet a complaint based on underestimated compliance costs (a negative realization of $\varepsilon$ such that $\hat{c} + \varepsilon < d$) coerces compliance. Inducing compliance in such cases lowers welfare. In setting the optimal enforcement parameters, the regulator must take account of this possibility, possibly by lowering the inspection probability $\pi$ below its maximal value.\footnote{Formally, in the relevant first-order condition

$$
\frac{\partial SL}{\partial \pi} = \frac{\partial \hat{c}}{\partial \pi} [(1 - \rho_\beta(\hat{c})\pi) (\hat{c} - d)] + \int \rho_\beta(c|\varepsilon)(c - d) dG(c) dE(\varepsilon)
$$

the first term is negative but the second is ambiguous in sign: if so, an interior maximum for $\pi$ cannot be ruled out.}
This is hardly surprising. Consider the case where the variance of the noise term is so large that the informational content of whistleblowing is low: it cannot be optimal for a welfare-motivated regulator to respond to every disclosure.

### 3.3 Assumption $\gamma$

Under Assumption $\gamma$, an agent with a noisy observation $\tilde{c}_i = c_i + \varepsilon$ reports a planned violation if and only if $\pi(\tilde{c}_i + f) > \delta_i$. Conditional on the noise term $\varepsilon$, the probability of disclosure now equals

$$\rho_\gamma(c_i, \pi, f|\varepsilon_i) = \Phi_\gamma(\pi(c_i + \varepsilon_i + f)),$$

For any level of compliance cost $c_i$, the firm evaluates the expected likelihood of disclosure as

$$\rho_\gamma(c_i, \pi, f) = \int \Phi_\gamma(\pi(c_i + \varepsilon + f))dE(\varepsilon).$$

The introduction of noise in estimates of compliance costs does not alter the qualitative findings of Section 2.4. As we found there, the net benefit from compliance may not be monotonic in $c$. Restricting attention to cases in which the net benefit is decreasing in compliance costs, there will be voluntary compliance for $c_i$ below some threshold $\hat{c}(\pi, f|\gamma)$. Assuming that the regulator must choose $\{\pi, f\}$ to minimize $SL(\pi, f|\gamma) = \int SL(\varepsilon|\gamma)dE(\varepsilon)$, as in the case without noise, the optimal policy may involve a less than maximal penalty and/or less than maximal inspection probability. In sum, Proposition 4 is robust to the introduction of noise.

### 4 Mixed Motivations and Lessons

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.

We have proceeded on the assumption that there were three theories as to what drove whistleblowing, one of which was correct. In any real world population, however, there could likely 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 more difficult to pin down. Allowing for mixed motives generates some mess without adding much to our understanding. Envisaging a population that is a mixture of $\alpha$-, $\beta$- and $\gamma$-types in some known proportions leads to policy implications that can be thought of as ‘weighted averages’ of the pure cases. But our analysis does suggest that policy advice that is heard - such as “we should always pursue good quality evidence of wrongdoing brought forward by whistleblowers to the fullest extent off the law” - has embedded in it implicit assumptions about prevalent motivations.
Though the model is particular, the analysis points to two general lessons. Both 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 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, in addition to the direct affect on compliance incentives. Again, the quantitative and qualitative response will depend upon whistleblower motives.

In Section 3 we explored the robustness of the results. The assumption that prospective whistleblowers have full knowledge about the ‘type’ of about their employer - and therefore about the context within which a particular decision to violate sits - was relaxed. Whilst this was a natural approach to take, there are other ways in which ‘noise’ of one sort or another could reasonably have been introduced into the framework. One could investigate the potential role of strategic behavior or lying on the part of whistleblowers, or the implications of potential for error in the evidence chain, or of slippage or imperfection in the enforcement process itself. Whilst each of these would render the model inconsistent with the strict definition of whistleblowing that we have adopted (that due to Glazer and Glazer (1989)), they would nevertheless be useful things to do, and exploring the robustness of our policy conclusions to these sorts of complications remains for further research. One might in addition consider a variety of other extensions. In a repeated setting a firms past performance, for example, may matter, though in ways that may not be straightforward 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’ or the related ‘betrayal heuristic’ (Sunstein (2003)) may make disclosure more likely. One could also consider employers differentiated along more dimensions than simply cost of compliance, or consider a budget-constrained or alternatively-motivated regulatory agency.

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 could/should be
combined with other enforcement instruments, such as random inspections.

5 Bibliography


Figure 1: Relationship between compliance cost and probability of disclosure under conscience cleansing

$\rho(\alpha(c_i))$
Figure 2: Realized probability of compliance (solid line) under optimal policy versus first best (dashed)
Figure 3: Realized probability of compliance in example