Optimal Review of Conduct with Informative Prior Audits

Claude Fluet*  
Université Laval, CRREP, CRED

Murat C. Mungan†  
George Mason University, Antonin Scalia Law School

March 2022

Abstract

We consider a principal-agent setting wherein the principal may reward (or punish) agents upon completing a review of noisy signals pertaining to the agent’s behavior. Only agents who are audited are reviewed. The audit process itself is (weakly) informative of agents’ behavior, because agents who act prosocially are (weakly) more likely to be audited. A reward generates reputational benefits in addition to its monetary value, and audits may trigger similar reputational impacts. We characterize the optimal review process, and identify the factors that affect how conservative or liberal the review process ought to be, including how visible the audit process is to third parties. We explain how our results pertain to important two-step review processes, including: arrests followed by trials; nominations followed by reviews; tax audits followed by compliance reviews; and suits followed by judicial review.

Keywords: Norms, social concerns, reputation, esteem, stigma, signaling, regulation, audits, reviews.

1 Introduction

In economic studies, penalties and rewards have been analyzed as tools to incentivize people to engage in prosocial behavior or refrain from harmful acts. These tools operate directly by attaching, in expectation, undesirable and desirable consequences to engaging in bad and good acts, respectively. This is the logic underlying, among others, the optimal deterrence literature wherein damages, fines, and imprisonment are used as a deterrent, as well as the economics literature investigating the optimal breadth and scope of patents to incentivize research and development. Many other settings have been analyzed

*Email: claude.fluet@fsa.ulaval.ca.
†Email: mmungan@gmu.edu
where economists sought to characterize optimal rewards and punishments in the form of taxes, subsidies, fines, sanctions, rewards, and other tools.

In addition to these direct incentive effects, rewards, liabilities, and punishments have a less obvious and less studied impact on behavior, which occurs due to their informational content. When carrots and sticks are conditioned on recipients’ accomplishments, failures, or behavior, they reveal information about the recipient to third parties. People may naturally alter their behavior towards reward recipients, or people who are punished, based on this information. Many familiar examples can be formulated. An academic hiring committee may be more willing to hire a person if she was the recipient of a *best scholar* award. Conversely, a news agency may be less likely to hire a journalist after it finds out that he was previously found liable for disseminating fake news. In the former case, the best scholar award may come with a monetary prize. In the latter case, the journalist may have paid monetary fines. However, in both cases the direct monetary consequence of these critical events may be dwarfed by their indirect, reputational, effects.

The significance of reputational effects depend, however, on the quality and importance of the information generated by the conveying of the reward or the infliction of the punishment. If, for instance, upon further examination the academic hiring committee finds out that the recipient of the best scholar award was determined by a lottery, the reward will be quite meaningless, as it will only convey that the candidate was lucky. Similarly, the journalist’s liability may be quite uninformative, if the news agency learns that he was found liable in a country with a despotic regime which has a reputation of bullying journalists. These two hypotheticals relate to the quality of the information revealed by the conveyance of the award and the liability.

How can one adjust the informational content of the reward or liability? So far, some of the most important works in the literature on this issue have surprisingly focused on a rather indirect method of altering the informational content of these devices, namely the magnitude of the reward or penalty (e.g. Benabou and Tirole (2006) and (2011), Adriani and Sonderegger (2019)). Altering the size of the direct incentive changes the set of individuals who engage in good and bad behavior, respectively, which leads third parties to reconsider their beliefs regarding the average characteristics of individuals in these sets. These beliefs shape the interactions between third parties and recipients (and non-recipients) of carrots and sticks. Thus, the size of the carrot or stick impacts the informational content associated with its receipt.

The *best scholar* and *journalist* examples, above, highlight an issue that is neglected in this type of analysis. In these examples, the presumed meaningfulness of an award and liability, respectively, vanished due to a consideration that has nothing to do with the size of the sanction, namely the process through which the recipient of the award and liability is determined, respectively. If instead of basing the receipt of rewards on the perceived characteristics of candidates, one simply flips a coin, the reward conveys very little information.

This relationship between review processes and the informational value of rewards and penalties has gone virtually unnoticed in the theoretical economics
literature, with very little exception.\textsuperscript{1} This may be perhaps because most of this literature implicitly or explicitly assumes either that people’s acts are perfectly detected (e.g. Benabou and Tirole (2006) and (2011)) or are imperfectly detected (e.g. audited with a probability less than one, as in the tax literature), but, once detected, are accurately perceived (e.g. Rasmusen (1996), Iacobucci (2014), Mungan (2016a)). In these settings, a ‘review process’ is not needed, simply because people’s actions are perfectly revealed to reviewers who decide whether to reward or punish them. However, as noted in literature, when false positives –providing an undeserving party a reward or penalty– are possible, they dilute the informational value of reward and punishment conferral (Mungan 2016b, 2017 and Fluet and Mungan 2022), and one can extend this logic to form a similar conjecture regarding the presence of false negatives. Thus, in realistic settings, offering rewards and punishments which carry large reputational impacts requires designing review processes that adequately balance the ‘informational costs’ associated with these two types of errors. Moreover, maximizing the incentives generated through reward and punishment mechanisms requires a further balancing between the reputational impact of the mechanism versus the direct incentive effects (e.g. due to the monetary prize or fine).

In this article, we study the interaction between these review processes and incentives generated by reputational and formal sanctions with the goal of identifying incentive maximizing review processes. We focus on the review procedures of a final decision authority (‘DA’ with pronoun it). This could be a court, a professional review board, an award committee, a professor grading an exam, a promotion committee, etc. The issue to be reviewed by the DA may be brought to its attention randomly, or through a procedure which is somewhat informative. In the context of tax compliance, for instance, one may assume that all individuals are audited randomly, in which case, a person’s tax documents may be brought to the attention of the IRS randomly. The professor who is grading exams can be conceived of as a special and extreme version of this ‘random audit’ case: all students’ exams are graded, regardless of how much they have studied. In other cases, the fact that a situation has been brought to the DA’s attention may itself be informative of the behavior of the person being reviewed. In the criminal context, for instance, police are more likely to arrest people who are involved in some criminal activity than they are to arrest people who are innocent; and in the tort context, a person who has failed to exercise due care is more likely to cause harm, and therefore be sued. Similarly, people with good research ideas may more likely be nominated for research grants than bad researchers. In the latter three cases, the fact that a person has been brought up for review is, to some extent, indicative of his behavior, whereas in the former two cases no such inference can be drawn. Thus, the informativeness of the ‘arrival process’ of the issue to be analyzed differs across contexts.

Another source of differentiation across contexts is the accuracy or the in-
formativeness of the final review process conducted by the DA. Intuitively, this feature refers to how well the DA can distinguish between people who have made an effort to obtain the reward or avoid the punishment from others who have not. In the exam context, for instance, this could be measured by the professor’s skill in writing questions that successfully sort out student’s who have a good command of the material from those who do not. In the criminal justice context, this could be measured by the ability of the court to sort out the guilty from the innocent.

One would suspect that these two features, i.e. the informativeness of the arrival process and the final review process, as well as others, ought to naturally affect the design of review processes vis-à-vis the maximization of the incentive effects of rewards and penalties. Here we operationalize this idea by formalizing the difficulty of the review process, i.e. how much evidence there must be for the reviewer to extend a reward or a penalty to the reviewee. In the legal context, the difficulty of the review process corresponds to the standard of proof, i.e. how much evidence there must be for a finding of liability or guilt. In the more familiar context of academic exams, the difficulty of the review process corresponds to the passing grade, e.g. $x$ points out of 100.

Our preliminary analysis reveals some intuitive, and other not-so-intuitive results. For instance, one may conjecture that if the final review process is very noisy, but the arrival process is quite informative, it makes little sense to have any kind of review process at all, at least for purposes of maximizing the informational content of rewards and penalties. This conjecture holds in many cases, and is particularly important in the legal context where a ‘no-review’ process corresponds to a ‘strict liability’ and a ‘review process’ corresponds to ‘fault-based liability’ regimes, e.g. negligence. However, a relatively subtle effect causes reviews to increase the reputational effects of rewards and penalties even in some cases where they are less informative than the arrival process. This is because making the reward or penalty subject to a final review reduces the number of rewards/penalties imposed by the DA. This causes these conferrals to be even rarer events, which, in turn, reduces the variance of the conferral, and makes the process more informative. The intuition behind this effect can be conveyed through a simple mathematical problem.

Suppose that there are two bags $A$ and $B$ containing $R_A \geq R_B$ red chips, respectively, and $100 - R_A$ and $100 - R_B$ white chips, respectively. The bags are not labeled, so that one does not know which one is which. A bag is randomly selected from which a chip is drawn. Obviously a red draw is more indicative of having drawn from bag $A$, and a white draw is more indicative of having drawn from bag $B$. But, how large is the discriminatory power of a single draw? Alternatively, how much hinges on a draw from these two bags? Obviously, the answer depends on $R_A$ and $R_B$, and the greater the difference is between these two values, the more we can infer. Moreover, the smaller the variance of our draw, the more informative it is, because there is less uncertainty. These are the standard considerations that affect the reliability of statistical tests of differences in means: the difference in the means of two groups and the standard deviation. Specifically, if we use the abbreviations $r_i = \frac{R_i}{100}$ for $i \in \{A, B\}$ and
\( r = \frac{r_A + r_B}{2} \); then the difference in our estimates of the probability that we drew our chip from bag \( A \), conditional on having drawn a red versus a white chip, is

\[
P(A|R) - P(A|W) = P(A) \left[ \frac{P(R|A)}{P(R)} - \frac{P(W|A)}{P(W)} \right] = \frac{1}{2} \left( \frac{r_A}{r} - \frac{1 - r_A}{1 - r} \right) = \frac{1}{4} \frac{r_A - r_B}{r(1 - r)}
\]

Here the numerator represents the difference in ‘red representativeness’ in bag \( A \) versus bag \( B \) and the denominator is the variance in the distribution of chips. Thus, when red chips are uncommon to begin with, i.e. \( r < 0.5 \), taking out an equal number of red chips from each bag and replacing them with white ones would increase the informational value of our draw (by keeping the numerator constant and reducing the denominator). A reduction in rewards, which causes fewer people to make efforts to get rewards, has an effect similar to reducing \( r \). As we explain here, implementing final reviews can have a similar effect by reducing the number of people who get rewards, and, thus, enhance reputational effects. This example and its explanation reveal the importance of a third factor which is pertinent to identifying informational content maximizing review processes, namely the frequency with which rewards and penalties are provided. As we note in section 3, below, this consideration informs an existing debate in legal scholarship on whether the use of strict liability should be confined to circumstances in which the risky activity is uncommon. It cuts against the existing doctrine, and suggests that strict liability is even less likely to be socially desirable when the act is uncommon.

Although these dynamics are interesting and may explain, in some contexts, how review processes ought to be designed, they abstract from a very important consideration. Specifically, they are built on the premise that the DA has monopoly over information dissemination, i.e. that all information receivable by third parties must come from the DA. In many circumstances, this assumption is clearly violated. Many arrest records, for instance, are public information. Thus, even absent a court’s verdict regarding a defendant’s actions, third parties can receive information regarding the defendant’s arrest record. In these situations, third parties can receive all the information that is capturable from the arrival process, even when the DA conducts no review. Thus, in such circumstances the informational content of a final review is diminished. What does this imply with respect to how the final review process ought to be designed for purposes of maximizing the informational content of the review process as a whole? We show, perhaps counter intuitively, that in these circumstances having a final review process always enhances the informational content of rewards and penalties. This is because, since third parties have independent access to information pertaining to the arrival process, conducting a final review never interferes with what they can infer from this process. Thus, even very noisy final review processes can be used to enhance the informational content of re-
wards and penalties, as long as they provide even the smallest amount of new information to third parties.

Our analysis also highlights a potential conflict that may arise in choosing the review standard that pertains to the provision of formal incentives and reputational incentives. Specifically, when the final review process is more informative than the arrival process, formal incentives are maximized by the implementation of a review standard which corresponds to the \textit{preponderance of the evidence standard} in the legal context, which need not maximize reputational incentives. In these cases, splitting formal liability determinations from what we call \textit{‘conduct announcements’} can help in the maximization of incentives. This can be done by using different standards of review in determining whether to reward or punish individuals from the standard of review used to make an announcement about the DAs classification of the person’s act (e.g., commendable or shameful). This approach, in essence, allows the DA to maximize formal and reputational incentives independently from each other.

To summarize our findings and to contrast them with results obtained in the prior literature, we interpret them in the tort context where no-review corresponds to strict liability and a review-based process corresponds to a negligence regime. Absent reputational considerations, it is optimal to use strict liability if, and only if, the final review process is more informative than the arrival process. However, when findings of liability can trigger reputational losses, the final review process being less informative than the arrival process is a necessary, but not sufficient, condition for the optimality of strict liability, unless the imposition of liability is not the norm. Therefore, the presence of reputational concerns limits the optimality of strict liability. Moreover, everything else equal, the presence of individuals who can obtain information regarding the occurrence of accidents independently from lawsuits cuts against the optimality of strict liability. Thus, the presence of reputational concerns, as well as the possibility of third parties obtaining information independently from courts narrows the conditions under which strict liability can be optimal, further.

In section 2, below, we build a fairly generalized model to incorporate the considerations outlined above. Subsequently, in section 3, we use this model to analyze review processes that can be designed to maximize the informational content of rewards and penalties in a setting where the difficulty of the review process can influence actors’ behavior. We conclude in section 4.

2 Model

2.1 Overview

We consider a continuum of individuals who each choose an act \( a \in \{0, 1\} \) where 1 denotes a prosocial act and 0 an antisocial act which leads to an expected social loss of \( h > 0 \) relative to the prosocial act. Neither act is directly observable, but may trigger an audit, with probabilities \( p_1 \) and \( p_0 \) respectively, which is in place
to incentivize the commission of act 1 versus 0.² If audits are triggered upon suspicion of antisocial acts to be punished, then it is natural to assume \( p_1 < p_0 \). Conversely, if audits are triggered upon suspicion of prosocial behavior to be rewarded, then it is natural to assume \( p_1 > p_0 \). To reduce repetition, throughout our analysis, we focus on cases in the first category. However, because the analysis is equally applicable in the second context, we define concepts in a neutral way, when possible, e.g. by referring to sticks and carrots, collectively, as incentives.

Conditional on audit, the DA observes a noisy signal emitted by the reviewee, \( x \in [x, \pi] \), which we call evidence.³ The evidence pertains to the behavior of the reviewee, and based on its relationship with what the DA deems to be sufficient evidence, the DA chooses whether to impose a monetary liability (denoted \( s \)) upon the reviewee. As we explain in the next section, the ‘sufficiency’ of evidence criterion that the DA uses to decide cases can be completely summarized by the likelihood, denoted \( \alpha \), with which the DA incorrectly punishes a reviewee who has engaged in the prosocial act. For any criterion that leads to such a probability, there is a corresponding probability \( \beta(\alpha) \), with which the DA correctly punishes a reviewee who has engaged in the antisocial act. Thus, a person who engages in the prosocial (respectively antisocial) act is eventually punished with probability \( p_1 \) (respectively \( p_0 \)). In addition to suffering this punishment, the person’s relationships with third parties may be impacted by the fact that he was audited and/or punished. Thus, a person can be deterred from engaging in antisocial behavior due to the presence of both formal sanctions, as well as informal sanctions, which take the form of reduced value from interactions with third parties.

A person makes decisions by comparing the formal and informal sanctions he expects to suffer against the gains he obtains (or costs he avoids incurring) by engaging in the antisocial rather than the prosocial act. This gain, denoted \( g \in [0, \bar{g}] \), differs from person to person, and its distribution among the population is summarized by the probability density function \( k(g) \) and the corresponding cumulative distribution function \( K(g) \) with support \([0, \bar{g}]\). We assume that the deterrence of antisocial acts is desirable, but that they cannot be completely deterred by formal sanctions, i.e.,

\[
h > \bar{g} > s
\]

A person chooses to engage in the antisocial act if his gains are sufficiently high. Specifically, his net gains from the antisocial and pro-social acts are, respectively,

\[
g + E(v|a, a = 0) - p_0 \beta(\alpha)s \quad \text{and} \quad E(v|a, a = 1) - p_1 \alpha s
\]

where \( v \) denotes the value of interactions with third parties, whose behavior towards the person is impacted by the information they receive through the

²See Shavell (1991) and Mookherjee and Png (1992) who also treat these audit probabilities as given, since they are determined by factors pertaining not only to the specific decision process they analyze but by a host of many other considerations.

³The bounds of the support need not be finite.
regulatory regime, which consists both of audits as well as the DA’s review. Thus, the person engages in the antisocial act if:

\[ (p_0\beta(\alpha) - p_1\alpha)s + E(v|\alpha, a = 1) - E(v|\alpha, a = 0) < g \text{ or } \]

\[ g > \hat{g}(\alpha, \hat{\Omega}(\alpha, g^e)) = (p_0\beta(\alpha) - p_1\alpha)s + \hat{\Omega}(\alpha, g^e) \]  

(3)

where \( \hat{\Omega} \) denotes anticipated level of expected stigma, i.e. the reduction in the expected value of a person’s interactions with third parties, due to the different information that third parties may receive as a result of the person’s choice to engage in the antisocial act rather than the prosocial act. The critical threshold \( \hat{g} \) defined in (3) characterizes individuals’ best responses in terms of engaging in the antisocial versus prosocial act as a function of their anticipated level of stigma \( \hat{\Omega} \) and the differential punishment risk, i.e. \( p_0\beta(\alpha) - p_1\alpha \). The degree of stigma that third parties impose is, of course, also responsive to the behavior of individuals and the review procedures chosen by the DA. Specifically, in addition to \( \alpha \), it depends on the gain threshold defining the proportion of individuals who engage in the antisocial act, and, thus, we denote the best response of third parties as \( \hat{\Omega}(\alpha, g^e) \) for any such threshold \( g^e \in (0, \hat{g}) \). The Bayesian equilibrium threshold gain, denoted \( g^e(\alpha) \), and the equilibrium level of expected stigma, \( \Omega(\alpha) \equiv \hat{\Omega}(\alpha, g^e(\alpha)) \), are obtained when individuals’ choices over acts lead third parties to stigmatize individuals with expected costs that lead them to behave exactly in that manner. In symbols, an equilibrium is obtained when

\[ \hat{g}(\alpha, \hat{\Omega}(\alpha, g^e)) = g^e \]  

(4)

which implicitly defines the equilibrium threshold as a function of \( \alpha \), which we henceforth express as \( g^e(\alpha) \). We note that multiple equilibria are in general possible. But, because they can be ruled out through plausible restrictions, we assume a unique equilibrium for each \( \alpha \) such that \( g^e(\alpha) \) is a well-defined function.

The above description is meant to give an overview of our model and to illustrate the various mechanics in play, which we explain in more detail in the next section. Specifically, in section 2.2, we describe the evidence generation and review process, i.e. how \( \alpha \) and \( \beta(\alpha) \) emerge from the evidence reviewed by the DA. In section 2.3, we explain how third parties receive and process information from the regulatory mechanism, i.e. how \( \hat{\Omega} \) is related to \( \alpha \) and any given threshold \( g^e \) describing the individuals’ behavior. After introducing these elements, in the subsequent section, we characterize equilibria, analyze the properties of reputational sanctions and discuss how review processes can be designed to maximize the incentives provided to the population.

Before proceeding further, we highlight some aspects and assumptions underlying our analysis. We are almost exclusively interested in the DA’s review process, summarized by \( \alpha \), as a choice variable. Thus, we take all other parameters, which could naturally be impacted by other policy choices, as exogenously determined. Specifically, the audit process (summarized by \( p_0 \) and \( p_1 \)) and the magnitude of the formal incentive (summarized by \( s \)) are both taken as given.
This approach allows us to focus entirely on the review process, and can also be justified by the fact that in many cases the DA is not in a position to alter the process through which it receives cases and the penalty that it imposes (e.g., in the criminal setting, law enforcers and legislators choose the audit process and the applicable sanctions in the criminal setting, respectively, but the standard of proof is a legal concept implemented by courts).

In addition, we assume that these exogenous factors are such that formal sanctions, alone, are insufficient to deter all individuals, i.e. \( \max_{\alpha} (p_0 \beta(\alpha) - p_1 \alpha) s < \bar{g} \). The standard Beckerian insight suggests that under-deterrence should emerge as an optimal solution, because when full-deterrence is obtained, the probability of audit can be reduced to save valuable enforcement resources. Therefore, if acting efficiently, the authorities whose behavior we do not analyze here would choose small enough audit probabilities to achieve this result.\(^4\) Moreover, in the alternative case, one could achieve any desired level of deterrence through formal sanctions alone, suppressing all information flow from the regulatory regime to third parties, in which case there would not be much to analyze regarding the optimal review process.

### 2.2 Evidence and Decision Rules

The probability with which an individual emits any signal \( x \in [\underline{x}, \bar{x}] \) depends on whether he engaged in act 1 or 0, and is described by the probability density functions \( f_1 \) and \( f_0 \). These functions satisfy the monotone likelihood ratio property (MLRP) with \( f_0 / f_1 \) monotonically decreasing, i.e. a large \( x \) suggests that the person is more likely to have engaged in act 1. \( F_0 \) and \( F_1 \) are the corresponding cumulative distribution functions and it naturally follows that \( F_1 \) first-order stochastically dominates \( F_0 \).

An intuitive decision rule is for the DA to find liability only if \( x \) is below a critical value, denoted \( x^* \), and which we refer to as the evidence threshold. By doing so, the DA implicitly fixes the probability of erroneously imposing liability on a reviewee who has engaged in pro-social behavior. We refer to this as the probability of type-1 error (or “false positive”), conditional on a review, which equals

\[
\alpha = F_1(x^*)
\]

The probability of correctly finding a given reviewee liable is

\[
\beta = F_0(x^*)
\]

so that the probability of type-2 error (“false negative”), conditional on review, is \( 1 - \beta \). This type of threshold decision rule is efficient in the sense that the probability of type-2 error is minimized for any level of type-1 error.

Although liability is assessed on the basis of a critical evidence threshold, it is convenient to conduct the analysis as if the DA directly chooses \( \alpha \in [0, 1] \),

\(^4\)See, e.g., Fluet and Mungan (2017), where this result is proven in more detail.
implying the evidence threshold \( x^s = F_1^{-1}(\alpha) \). Choosing the type-1 error yields a \( \beta \) satisfying
\[
\beta(\alpha) = F_0(F_1^{-1}(\alpha)) \tag{7}
\]
It is easily seen that \( \beta(0) = 0 \), \( \beta(1) = 1 \), and that the derivative satisfies
\[
\beta'(\alpha) = \frac{f_0(F_1^{-1}(\alpha))}{f_1(F_1^{-1}(\alpha))} \tag{8}
\]
and \( \beta''(\alpha) < 0 \), which follows from the MLRP of \( f_0/f_1 \). These properties imply \( \beta(\alpha) \) for all \( \alpha \in (0, 1) \).

It is important to note that \( \alpha = 1 \) and \( \alpha = 0 \) correspond to cases where the DA has a policy of not conducting reviews. When \( \alpha = 0 \), there is neither a review nor any formal incentives provided to individuals. When \( \alpha = 1 \), the regime provides "Unconditional Incentives", i.e. strict liability upon the audit triggering event. In all other circumstances, i.e. when \( \alpha \in (0, 1) \), we say that the DA follows a "Review-Based Incentives" regime. The comparison between the behavioral effects of the Unconditional Incentives regime and Review-Based Incentive regimes depends much on the precision of the evidence about care, as we shall demonstrate. Therefore, we proceed by describing how the properties of \( \beta(\alpha) \) capture the informativeness of the evidence generating process (henceforth EGP).

**Accuracy of the Evidence**

Any pair of densities \( \{f_0, f_1\} \) defines an information system. According to a well known criterion, a system is more informative than another if it yields a smaller type-2 error for any given level of type-1 error.\(^6\) For our purpose, it will be useful to allow for all possibilities ranging from completely uninformative to nearly perfectly informative evidence.

Possible information systems are represented by the set of continuously differentiable and strictly increasing concave functions \( \beta(\alpha), \alpha \in [0, 1] \), satisfying \( \beta(0) = 0 \) and \( \beta(1) = 1 \). From this set, one can extract families whose elements are ordered in terms of the relation ‘more informative than’.

**Definition 1** \( \{\beta_\gamma\}_{\gamma \in \mathbb{R}_+} \) is an ordered family with system \( \gamma' \) more informative than system \( \gamma \) if \( \gamma' > \gamma \) implies \( \beta_{\gamma'}(\alpha) > \beta_{\gamma}(\alpha) \) for all \( \alpha \in (0, 1) \), with \( \beta_0(\alpha) \equiv \alpha \) and \( \lim_{\gamma \to -\infty} \beta_\gamma(\alpha) = 1 \) for \( \alpha > 0 \).

At the lower bound \( \gamma = 0 \), the EGP is completely uninformative and the likelihood ratio of anti- versus pro-social act satisfies \( f_0/f_1 \equiv 1 \). At the other end, for \( \gamma \) very large, \( f_0/f_1 \) ranges from very large values to nearly zero. From

\(^5\)Conversely, for any function \( \beta(\cdot) \) with the above properties, there exists signals with density functions \( f_0 \) and \( f_1 \) that generate \( \beta(\cdot) \). An example is the uniform distribution \( F_1(x) = x \), for \( x \in [0, 1] \), together with \( F_0(x) = \beta(x) \).

\(^6\)This criterion is discussed in Blackwell and Girschick (1954) among other equivalent conditions. The criterion is equivalent to the likelihood ratio \( f_0/f_1 \) having more dispersion (in the sense of a mean preserving spread) in more informative systems; see Demougin and Fluet (2001) for a simple proof and Jewitt (2007) for a comprehensive analysis.
(8), this implies that the slope $\beta'(\alpha)$ is very large for small values of the type-1 error and approaches zero as $\alpha$ approaches unity. We will repeatedly use the latter property.

**Lemma 1** In an ordered family of information systems, $\lim_{\gamma \to \infty} \beta'(1) = 0$.

Henceforth, when explicitly comparing information systems, we assume they belong to an ordered family as defined above. Otherwise, we omit reference to the index $\gamma$.

An important consideration is whether the EGP provides useful information compared to the mere knowledge that the reviewee was audited, i.e. the arrival process. Because $p_0 > p_1$, an audit is more consistent with antisocial behavior. Loosely speaking, the EGP provides valuable incremental information compared to the arrival process if it is capable of producing a body of evidence that is more indicative of pro-social behavior than the audit is indicative of antisocial behavior. We call such EGPs *more informative* than the arrival process. Conversely, we call EGPs that cannot fulfill this function, *less informative* than the arrival process. This concept is defined formally as follows.

**Definition 2** An evidence generating process is more informative than the arrival process if

$$\frac{p_0 f_0(x)}{p_1 f_1(x)} < 1 \text{ for some realizations } x \quad (9)$$

Whether the EGP is more informative than the arrival process is closely related to whether the incremental liability risk that one faces from engaging in antisocial behavior can be maximized by a review-based incentive regime. This relationship can be summarized as follows.

**Lemma 2** The evidence generating process is more informative than the arrival process if and only if $p_0(\beta(\alpha) - p_1\alpha)$ has an interior maximum.

The function $p_0(\beta(\alpha) - p_1\alpha)$ is the differential liability probability between engaging in pro-social and antisocial behavior under a liability regime with type-1 error equal to $\alpha$. The differential is depicted in Figure 1 for three information systems from an ordered family with $\gamma'' > \gamma'' > \gamma'$. The positively sloped dotted line is the differential liability risk in the limiting case (represented by $(p_0 - p_1)\alpha$) of completely uninformative evidence (i.e., $\gamma = 0$). The negatively sloped dotted line is the upper bound (represented by $p_0 - p_1\alpha$), never reached, of completely informative evidence ($\gamma = \infty$). Under system $\gamma'$ (thin), the EGP is less informative than the arrival process. The differential liability risk then

---

7See Demougin and Fluet (2005, 2006) for a similar discussion.
8We are slightly misusing the usual terminology here. Under the standard definition (see footnote 6), the EGP is ‘more informative’ than the arrival process if the likelihood ratio $f_0(x)/f_1(x)$ has more dispersion than the likelihood ratio induced by the signals ‘audit’ and ‘no audit’. One can show that this implies the existence of realizations $x''$ such that $(1-p_0)f_0(x'') > (1-p_1)f_1(x')$ or of realizations $x''$ such that $p_0 f_0(x'') < p_1 f_1(x'')$, or both. The latter inequality is the focus of Definition 3.
Figure 1: Liability probability differentials under different information systems: $p_0 = 0.6$, $p_1 = 0.2$, $\beta(\alpha) = \frac{\alpha^\gamma}{1 + \alpha^\gamma}$; $\gamma' = 1$, $\gamma'' = 4$, $\gamma''' = 99$.

reaches a corner maximum at $\alpha = 1$, i.e., unconditional incentives. The two other systems (medium for $\gamma''$ and thick for $\gamma'''$) are more informative than the arrival process, so the curves have an interior maximum. Obviously, (9) holds if the evidence is very informative (i.e., a large $\gamma$) and it does not hold if the evidence is sufficiently poor (i.e., a small $\gamma$).

In the sequel, we take the EGP as given for the category of cases considered. However, comparative statics statements about the optimal legal regime may depend on the quality of the potential evidence. We use the phrase *sufficiently informative* evidence to mean that there is a sufficiently large $\gamma$ in some ordered family of information systems such that the statement is true. Similarly, *sufficiently uninformative* means that there is a sufficiently small $\gamma$ in some such family such that the statement holds.

### 2.3 Third Party Information

In the overview of the model, we noted that third parties may receive information from the regulatory process, which includes audits and reviews, which they may use in shaping their interactions with people. This is what gives rise to expected reputational costs associated with acting antisocially, which we denoted as $\Omega$ in the overview. Here, we explain the various components that go into $\Omega$, and subsequently explain how each component reacts to various changes.

There are different possibilities with respect to what information third parties have access to in forming their opinions about individuals. Recall that there
are two steps in the regulatory process. First, an audit takes place, and, subsequently, and conditional on the audit, a review takes place based on evidence with a final decision reached by the DA on whether to punish/reward the auditee. In many contexts, it could be argued that only 'successful reviews', i.e. those where there are sufficient evidence for punishment/reward, are salient, and, thus, only they become public knowledge. For instance, in the context of tax compliance, one may claim that tax audits seldom become public knowledge unless they end with the defendant being found liable for non-compliance. In such cases, third parties would be able to partition the set of individuals, based on the information they received, into two groups as illustrated in Table 1.

Table 1: Partitioning by Non-observers

<table>
<thead>
<tr>
<th></th>
<th>Audited</th>
<th>Not Audited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient Evidence</td>
<td>Punishment/Reward</td>
<td>No Action</td>
</tr>
<tr>
<td>Insufficient Evidence</td>
<td>No Action</td>
<td>No Action</td>
</tr>
</tbody>
</table>

One may argue, however, that in other contexts third parties may receive information regarding audits, independently. For instance, criminal trials may become public knowledge regardless of whether they lead to convictions. In financial markets, it may be known that a firm is under investigation by the SEC for possible infringement of financial regulations. In these cases, third parties would be able to more finely partition individuals into various groups. This is illustrated by Table 2.

Table 2: Partitioning by Observers

<table>
<thead>
<tr>
<th></th>
<th>Audited</th>
<th>Not Audited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient Evidence</td>
<td>Punishment/Reward</td>
<td>No Action</td>
</tr>
<tr>
<td>Insufficient Evidence</td>
<td>Failed to Punish/Reward</td>
<td>No Action</td>
</tr>
</tbody>
</table>

It will often be the case that some third parties are able to partition groups as in Table 2, while others partition them as in Table 1, and perhaps some individuals are completely unaware of legal outcome. We ignore the third type, since
the presence of completely uninformed individuals is irrelevant to our analysis, because these people impose no reputational costs which are responsive to the review mechanism. Thus, to allow a comprehensive analysis, we consider cases where $q \in [0, 1]$ represents the proportion of third parties who have independent information about audits, as in Table 2, and we call these third parties Observers. We call the remaining third parties Non-Observers; these third parties receive information only about whether a person was the recipient of a punishment/reward. Although we treat $q$ as exogenously determined, we will comment on the social desirability of increasing $q$, if it were possible. This framework naturally allows us to consider the extreme assumptions where all parties are observers (i.e. $q = 1$) and where all parties are non-observers (i.e. $q = 0$).

As the above tables illustrate, a person may be perceived by observers and non-observers as belonging to one of four categories:

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category $P$:</td>
<td>Punished/Rewarded</td>
</tr>
<tr>
<td>Category $\tilde{P}$:</td>
<td>not Punished/Rewarded</td>
</tr>
<tr>
<td>Category $\tilde{A}$:</td>
<td>not Audited</td>
</tr>
<tr>
<td>Category $A\tilde{P}$:</td>
<td>Audited and not Punished</td>
</tr>
</tbody>
</table>

We use the letters listed above for references to each category throughout our analysis. In addition to these four groups, we also consider the category of auditee’s (i.e. the union of categories $P$ and $A\tilde{P}$), which we denote by the letter $A$. This will be helpful in distinguishing between the various reputational costs that one may suffer.

### 2.4 Reputational Sanctions and Behavior

**Observer and Non-observer Stigma**

Using the foregoing, we denote by $w_i$ the gains that a person gets from interacting with a third party who perceives him to be in category $i \in \{P, \tilde{P}, \tilde{A}, A\tilde{P}\}$. Here, we take the $w_i$’s as given. In the next sub-section, we explain how these gains are derived as functions of actors’ behavior profile. By invoking one of two equivalent assumptions, we may use the $w_i$’s to calculate actors’ expected third party interaction pay-offs. First, it may be assumed that each actor is randomly matched with a third party who may be an observer, or non-observer, with probabilities $q$ and $(1 - q)$ respectively. Alternatively, we may assume that each actor interacts with a continuum of third parties with proportions $q$ and $1 - q$ of observers and non-observers respectively. In explaining the dynamics, we will refer to interpretations that are consistent with the first assumption, because this eases descriptions of key terms. However, in either case, the expected third party interaction benefits for an actor, conditional on his behavior,

\footnote{We also disregard the possibility that the review board may decide to keep its decision confidential, as with Non-Prosecution Agreements. This would raise the possibility that Observers know that a person was audited, but do not know the decision of the review board. See Mungan (2019).}
is given by:

\[ E(v|\alpha, a = 1) = q\{p_1 \alpha w_P + p_1 (1 - \alpha) w_{A\beta} + (1 - p_1) w_A \} \]
\[ + (1 - q)\{p_1 \alpha w_P + (1 - p_1) w_{\beta} \} \]

and

\[ E(v|\alpha, a = 0) = q\{p_0 \beta w_P + p_0 (1 - \beta) w_{A\beta} + (1 - p_0) w_A \} \]
\[ + (1 - q)\{p_0 \beta w_P + (1 - p_0) w_{\beta} \} \]

Thus, we may express the reputational sanctions that an actor expects to be subjected to by non-observers and observers, respectively, as

\[ \hat{\Omega}_N = (p_0 \beta - p_1 \alpha)(w_{\beta} - w_P) \text{ and} \]
\[ \hat{\Omega}_O = (p_0 - p_1)(w_A - w_{A\beta}) + (p_0 \beta - p_1 \alpha)(w_{A\beta} - w_P) \]

such that

\[ \hat{\Omega} = E(v|\alpha, a = 1) - E(v|\alpha, a = 0) = q\hat{\Omega}_O + (1 - q)\hat{\Omega}_N \]

represents the total expected reputational cost introduced in equation (3) of the overview section. The total cost depends on the stigmas imposed by observers and non-observers weighted by the probability of being matched with a third party of either group.

**Third Party Beliefs as a Function of Behavior**

The above describes the stigma imposed by observers and non-observers, for any given third party expectations of quality as expressed by the \( w_i \)'s. Here we analyze how third parties calculate these expressions based on the actors’ behavior and the review process, i.e. how \( \hat{\Omega}(\alpha, g^c) \), through its components \( \hat{\Omega}_N(\alpha, g^c) \) and \( \hat{\Omega}_O(\alpha, g^c) \), can be expressed as a function of \( \alpha \) and \( g^c \). We focus on how changes in \( \alpha \) affect reputational sanctions. For the time being, \( g^c \) is taken as given.

As we noted, third parties use the information available to them to estimate the likelihood with which the person committed the antisocial act. We assume that they care about this fact, because they believe that people who are more likely to have engaged in such behavior are worse interaction partners (e.g. worse employees). We call the aspect that third parties care about the quality of a person. Since \( g \) captures a person’s tendency to commit antisocial acts, we assume that \( \phi(g) \) with \( \phi' < 0 \) captures third parties’ assessed quality of a person with gains \( g \). Thus, in a behavior profile where actors commit the antisocial act if, and only if, \( g > g^c \), the average quality of individuals who have committed the antisocial and prosocial act can be calculated as

\[ \lambda(g^c) = \frac{\int_{g^c}^{\infty} \phi(g) k(g) dg}{1 - K(g^c)} \quad \text{and} \quad \Delta(g^c) = \frac{\int_{g^c}^{\infty} \phi(g) k(g) dg}{K(g^c)} \]

where

\[ \Delta(g^c) = \frac{\int_{g^c}^{\infty} \phi(g) k(g) dg}{K(g^c)} - \frac{\int_{g^c}^{\infty} \phi(g) k(g) dg}{1 - K(g^c)} \]

\[ 15 \]
is the difference between the average qualities of individuals in the two groups, and is positive for all \( g^c \in (0, \bar{g}) \) since \( \varphi'(g) < 0 \).

We assume that each third party chooses the scope of his interaction with another individual based on his estimate of his quality, and that this provides benefits to individuals which are proportional to the same estimate. Let \( \Pr(i|a) \) denote the probability that an individual ends up in category \( i \) when he chooses action \( a \). Then

\[
\mu_i \equiv K(g^c) \Pr(i|a = 1) + (1 - K(g^c)) \Pr(i|a = 0)
\]

and

\[
\pi_i \equiv K(g^c) \Pr(i|a = 1),
\]

are, respectively, the total measure of individuals in group \( i \) and the measure of individuals in group \( i \) who have behaved prosocially. The third party benefits an individual obtains from being associated with group \( i \), denoted as \( \hat{w}_i(\alpha, g^c) \), can then be expressed as follows:

\[
\hat{w}_i(\alpha, g^c) = \frac{\pi_i}{\mu_i} [\lambda(g^c) + \Delta(g^c)] + (1 - \frac{\pi_i}{\mu_i}) \lambda(g^c)
\]

\[
= \lambda(g^c) + \frac{\pi_i}{\mu_i} \Delta(g^c)
\]

Thus, the interaction benefits associated with all groups contains the same baseline level \( \lambda(g^c) \) which is adjusted upwards by \( \Delta(g^c) \) proportionally with the fraction of individuals who have acted prosocially in that group. We also note that the average quality of individuals who are audited (i.e. \( \hat{w}_A \)) is a convex combination of \( \hat{w}_{AP} \) and \( \hat{w}_P \). Specifically,

\[
\hat{w}_A = \phi \hat{w}_P + (1 - \phi) \hat{w}_{AP}, \quad \text{where}
\]

\[
\phi \equiv \frac{\mu_P}{\mu_{AP} + \mu_P}
\]

is the proportion of individuals who are punished conditional on being audited.

Stigmatization occurs as a result of being associated with one group rather than another. We therefore use equation (16) to calculate the differences in average quality between any two groups \( i \) and \( j \), as follows

\[
\hat{w}_i - \hat{w}_j = \left( \frac{\pi_i}{\mu_i} - \frac{\pi_j}{\mu_j} \right) \Delta(g^c)
\]

Plugging in the definitions of \( \pi_i \) and \( \mu_i \) from (14) and (15), we can re-write (18) as

\[
\hat{w}_i - \hat{w}_j = \{ \Pr(i|a = 1) \Pr(j|a = 0) - \Pr(i|a = 0) \Pr(j|a = 1) \} \frac{\tau(g^c)}{\mu_i \mu_j}
\]

where the notation

\[
\tau(g^c) \equiv K(g^c)(1 - K(g^c)) \Delta(g^c)
\]
is used to abbreviate expressions further.

We can now derive the stigmas imposed by observers and non-observers by substituting from (19) in (10) and (11), plugging in the relevant conditional probabilities. For non-observers, straightforward re-arrangements then yield

\[ \hat{\Omega}_N(\alpha, g^c) = \frac{(p_0 \beta(\alpha) - p_1 \alpha)^2}{\mu_p(\alpha, g^c)(1 - \mu_p(\alpha, g^c))} \tau(g^c) \]

For observers, we first use (17) to rewrite (11) as

\[ \hat{\Omega}_O(\alpha, g^c) = (p_0 - p_1)(\bar{w}_A - \bar{w}_A) + (p_0(\beta(\alpha) - \phi) - p_1(\alpha - \phi))(\bar{w}_A - \bar{w}_p) \]

Substituting from (19),

\[ \hat{\Omega}_O(\alpha, g^c) = \left\{ \frac{(p_0 - p_1)^2 \tau(g^c)}{\mu_A(g^c)} \right\} + \left\{ \frac{(\beta(\alpha) - \alpha)^2}{\phi(\alpha, g^c)(1 - \phi(\alpha, g^c))} \frac{(p_0 p_1)^2 \tau(g^c)}{\mu_A(g^c)} \right\} = \frac{\zeta_A(g^c)}{\mu_A(g^c)} + \frac{\zeta_R(\alpha, g^c)}{\mu_A(g^c)} \]

where the last two terms correspond to the expressions in the two squiggly brackets. This decomposition of reputational sanctions imposed by observers is particularly useful, because it allows us to separate out the impacts of the audit process (the term \( \zeta_A \)) and the review process (the term \( \zeta_R \)).

**Lemma 3** For \( g^c \in (0, \gamma) \), \( \hat{\Omega}_O \geq \hat{\Omega}_N > 0 \) for all \( \alpha \in (0, 1) \) with at most a unique \( \tilde{\alpha} \neq 1 \) such that \( \hat{\Omega}_O = \hat{\Omega}_N \). Moreover, \( \hat{\Omega}_N(0, g^c) = 0 \) and \( \hat{\Omega}_O(0, g^c) = \hat{\Omega}_O(1, g^c) = \hat{\Omega}_N(1, g^c) > 0 \).

Figure 2 provides an illustration. In the limiting case where \( \alpha = 0 \), non-observers obviously generate no reputational incentives because they receive no information about the individuals' behavior. This is not so for observers who can independently distinguish between audited and non audited individuals. The reputational sanctions imposed by observers are then the same as in the other limiting case defined by \( \alpha = 1 \), i.e. the Unconditional Incentives regime. In that regime, an audit entails liability, so observers and non-observers receive the same information and, therefore, impose the same reputational sanctions. For any Review-Based Incentives regime, observers generate greater reputational incentives than non-observers, except possibly at some critical value of the type 1 error. This arises only when the EGP is very informative.\(^{10}\) Note that in any case the total reputational sanction \( \Omega \) is increasing in \( q \) for almost all \( \alpha \).

As drawn in Figure 2, the \( \hat{\Omega}_N \) and \( \hat{\Omega} \) curves exhibit an interior maximum with respect to \( \alpha \). This is not necessarily so. It depends on the informativeness of the EGP and the proportion of antisocial versus pro-social actions.

\(^{10}\)There must be realizations \( x \) such that \( p_0 f_0(x)/p_1 f_1(x) < (1 - p_0)/(1 - p_1) \), a condition stronger than (9). The \( \hat{\Omega}_O \) and \( \hat{\Omega}_N \) are then tangent at \( x \) solving \( p_0 \beta(x)/p_1 = (1 - p_0)/(1 - p_1) \). At \( x \), the categories \( \hat{\bar{A}} \) and \( \hat{\bar{P}} \) provide the same information, so \( \hat{\bar{w}}_{\hat{\bar{A}}} = \hat{\bar{w}}_{\hat{\bar{P}}} \). For \( \alpha > x \), being audited and not punished is more favorable news than not being audited, so \( \hat{\bar{w}}_{\hat{\bar{A}}} > \hat{\bar{w}}_{\hat{\bar{P}}} \).
Figure 2: Reputational Sanctions
Lemma 4 For \( g^c \in (0, \overline{g}) \):

(i) \( \partial \Omega_N(\alpha, g^c) / \partial \alpha > 0 \) if the EGP is sufficiently uninformative, or if it is less informative than the arrival process and \( \mu_P(\alpha, g^c) \geq 0.5 \); \( \partial \Omega_N(1, g^c) / \partial \alpha < 0 \) if the EGP is sufficiently informative, or if it is more informative than the arrival process and \( \mu_P(1, g^c) \leq 0.5 \).

(ii) \( \Omega_N \) has an interior maximizer; so does \( \Omega \) if either \( q \) is sufficiently large, the EGP is sufficiently informative, or if it is more informative than the arrival process and \( \mu_P(1, g^c) \leq 0.5 \).

Note that \( \mu_P(1, g^c) = p_0(1 - K(g^c)) + p_1 K(g^c) \). Therefore, a sufficient condition for \( \mu_P(1, g^c) \leq 0.5 \) is \( p_0 \leq 0.5 \). When only a minority of individuals are audited and the EGP is more informative than the arrival process, the \( \Omega \) and \( \Omega_N \) stigma curves are as in Figure 2, irrespective of the actual proportion of antisocial acts.

Discussion

The foregoing bears a relation to the “honor-stigma model” of Bénabou and Tirole (2006, 2011); see also Mazyaki and van der Weele (2019) and Adriani and Sonderegger (2019). In that literature, individuals choose between two actions and actions are directly (and perfectly) observable by third parties. With perfectly observable actions, using our notation, reputational incentives reduce to \( (g^c) \) as defined in (13). An important issue is whether \( (g^c) \) is a decreasing or increasing function of the threshold \( g^c \). When the function is increasing, behavior exhibits strategic complementarity, i.e. reputational incentives to act pro-socially increase as more individuals do so. Conversely, there is strategic substitutability when the function is decreasing.\(^{11}\)

In our analysis, however, actions are not perfectly observable by third parties, who can only observe the outcome of public enforcement policies, and so far we have taken \( g^c \) as fixed. We focused on how decision rules for finding liability, as captured by \( \alpha \), affect reputational sanctions by modifying the information content of signals about the individuals’ behavior. Keeping behavior constant, there are two effects which can go in the same or in opposite directions. Consider the expression for \( \Omega_N(\alpha, g^c) \) in equation (21). When \( p_0 \beta(\alpha) - p_1 \alpha \) is marginally increasing in \( \alpha \), a slightly less stringent requirement for finding liability increases the discriminant power of guilty verdicts, in the sense that antisocial acts become relatively more likely to be punished compared to pro-social acts. However, there is also a composition effect. A marginal increase in \( \alpha \) increases the proportion \( \mu_P(\alpha, g^c) \) of people found liable. When \( \mu_P(\alpha, g^c) \) is less than one half, the composition effect makes liability a noisier signal of antisocial behavior. The overall effect of less stringent evidence for finding liability may then be to increase or decrease reputational sanctions. By contrast, when \( \mu_P(\alpha, g^c) \) is greater than one half, the composition effect reinforces the discriminant effect. The argument is the same for the stigma imposed by observers in equation (23).

\(^{11}\)In our set-up, this depends on the probability distribution of \( \varphi(g) \) across types, which in turn depends on the form of the density function \( k(g) \) and the curvature of the function \( \varphi(g) \).
except that the discriminant effect is conditional on the individual being audited and reduces to \(\beta(\alpha) - \alpha\); accordingly, the composition effect depends on the proportion \(\phi(\alpha, g^c)\) of individuals found liable among the audited.

### 3 Welfare and Optimal Review Standards

The formal sanctions at the DA’s disposal are assumed to be monetary. Moreover, following the prior literature we assume that reputational benefits are completely transferable (see, e.g., the discussion in Fluet and Mungan 2022). Therefore, welfare is given by

\[
Z = \int_{g^c(\alpha)}^{\tilde{g}} (g - h)k(g)dg
\]

where \(g^c(\alpha) < \tilde{g}\) is the equilibrium threshold defined by (4). The inequality follows from our assumption that formal sanctions, alone, are insufficient to completely deter antisocial behavior, i.e. \(\tilde{g} > s\) (as noted in (2)), and the fact that reputational benefits vanish as complete deterrence is approached, i.e., \(\lim \Omega(\alpha, g^c) = 0\). Since antisocial behavior is inefficient, i.e. \(h > g\) (as noted in (2)), it follows that the optimal review process is that which maximizes deterrence. Therefore, we focus on characterizing review standards that maximize deterrence.

#### 3.1 Optimal Review Standards

**Proposition 1** There exists a decreasing function \(\hat{q}(\gamma)\) with \(q(0) = 1\) and \(q(1) = 0\) such that the optimal review policy is to conduct no review (i.e. \(\alpha^* = 1\)) if \(q < \hat{q}(\gamma)\) and to conduct a review if \(q > \hat{q}(\gamma)\).

Proposition 1 can be summarized by figure 3, below, and has two broad implications. First, for any given non-extreme evidence generating process, it follows that whether a final review is needed depends solely on the composition of third parties with respect to how they obtain information about the actor’s behavior. If most third parties have information about whether the actor was audited, then it is always optimal to conduct further review, i.e., it is optimal to have review-based incentives. On the other hand, if most parties obtain information from the DA, then it is optimal, given a non-extreme evidence generating process, to use unconditional incentives. The rationale behind this counterintuitive result is that imposing the sanction without a review reveals information regarding whether the actor was audited in a less noisy manner.

Next, we note a result related to the frequency with which incentives are used (i.e. \(\mu_p\)) and optimality of review based incentives.

---

\(^{12}\) We note that the welfare function may differ from (24) when, for instance, the formal sanction used is endogenous and not monetary. However, even then, it may become optimal to use review standards that maximize reputational concerns, because it may be more cost effective to use symbolic sanctions (Fluet and Mungan 2022).
\[ q \]

\[ q(\gamma) \]

Review Based Incentives

Unconditional Incentives

\[ \gamma' \]

\[ \gamma'' \]

\[ \hat{\gamma} \]

\[ \gamma \]
**Proposition 2** The optimal regime involves review based incentives, if it involves infrequently employed incentives (i.e. $\mu_P \leq 0.5$) and the evidence generating process is more informative than the arrival process.

The proposition establishes a relationship between the commonality of incentives, and the optimal regime when the EGP is more informative than the arrival process. An implication of the proposition is that the uncommonality of incentives are more consistent with review based incentives being optimal than frequently used incentives.

**Discussion**

The result summarized by proposition 2 is related to the ‘uncommonality requirement’ in tort law. In the United States, courts generally hold an injurer strictly liable in tort law if two requirements are met: “(1) the injurer’s activity must generate a highly significant danger even when undertaken with reasonable care; and (2) the injurer’s activity must be uncommon” (Shavell 2017, p. 1 citing Restatement (Third) of Torts). The wisdom behind the second of these requirements has been recently questioned in Shavell (2018), which argues that all acts that meet the first requirement ought to be regulated through strict liability.\(^\text{13}\) Shavell’s primary claim is that, by making uncommonality a requirement, the legal system forgoes the opportunity to increase the deterrence of acts that are dangerous but common, and that this causes reductions in welfare.

Our model can be used to discuss this requirement by interpreting $p_1$ and $p_0$ as the probabilities with which the actor causes harm by engaging in an act after taking due care versus not taking due care. Thus, liability is imposed with these probabilities under a regime of strict liability. On the other hand, liability is conditional on the finding of fault (with probabilities $\alpha$ and $\beta(\alpha)$, respectively) under a fault-based liability regime (e.g. simple negligence), and the review standard corresponds to the standard of proof used in trial in determining whether the actor was at fault. The uncommonality requirement then corresponds to the equilibrium level of deterrence (i.e., $K(g^e)$) being large. As proposition 2 illustrates, review based incentives are optimal under a broader set of conditions when liability is infrequent compared to when it is frequent. Liability is more frequent, in turn, when the level of deterrence is low. Thus, our analysis suggests that, contrary to the legal doctrine in the United States, fault-based liability enjoys a comparative advantage over strict liability when the act is uncommon. The rationale behind this result is that when liability is infrequent, switching from strict liability to fault based liability causes a further reduction in the frequency of liability. The end result is a more extreme separation between the groups of individuals who take care and who do not take care, and, thus a finding of liability has a greater signal value. Therefore, fault-based liability generates additional deterrence effects. Hence, our analysis suggests that there is a rationale to use fault-based liability more often when the act is uncommon.

\(^\text{13}\)Shavell (2018) also notes that the American approach appears to be an outlier, because no other countries seem to adopt a similar requirement for the imposition of strict liability.
On the other hand, proposition 1 can be related to the overtness of the harm in the tort context, which is captured by the parameter $q$. The proposition suggests that the benefits of strict liability are limited further when the harm is committed overtly. Contrary to our finding, in tort law, the overtness of the harm is not a factor that is relevant in whether liability ought to be strict or fault-based. This may be due to the fact that the overtness of harm is often correlated with other considerations which may affect the desirability of using strict liability, including the dangerousness of the act (captured by the pair $p_0,p_1$).

3.2 Optimal Review Standards with Split Liability and Announcements

We have assumed so far that third parties receive information about the DAs review only based on the DAs final determination on whether to penalize the actor. Here, we consider the possibility of the DA uncoupling its two primary functions: providing a formal incentive, and informing third parties regarding its review. To do so, the DA chooses two thresholds of evidence $x_f$ and $x_a$ (the subscripts are abbreviations for formal and announcement) and imposes a formal sanction if $x < x_f$ and announces its belief that the actor behaved antisocially if $x < x_a$. To ensure that its formal liability determination does not influence reputational incentives, the DA does not publicly announce whether the actor has received a formal punishment, and only publicly announces whether it believes that the actor behaved antisocially. Thus, the differential formal liability associated with non-compliance versus compliance is $p_0 \beta(\alpha_f) - p_1 \alpha_f$ where $\alpha_f = F(x_f)$, and the observer and non-observer stigma functions are obtained by replacing $\alpha$ and $\beta$ in (22) and (21) with $\alpha_a = F(x_a)$ and $\beta(\alpha_a)$. Thus, a person commits the antisocial act if

$$g > \hat{g}(\alpha_f,\hat{\Omega}(\alpha_a, g^c)) \equiv (p_0 \beta(\alpha_f) - p_1 \alpha_f)s + \hat{\Omega}(\alpha_a, g^c)$$

An immediate implication of Lemma 2 is that formal incentives are maximized by setting $\alpha_f = [\beta']^{-1}(p_0/p_1)$ if the final review process is more informative than the arrival process, and $\alpha_f = 1$ otherwise. The maximizer of $\Omega(\alpha_a, g^c)$, on the other hand, is interior for $q = 1$. Thus, when formal sanctions can be uncoupled from informal sanctions, it follows that the review pertinent to ‘shaming announcements’ can never be based on the arrival process, unless some third parties are non-observers.

4 Conclusion and Extensions

The reputational consequences of rewards and punishments, including legal sanctions, has received substantial interest in legal as well as economic scholarship. Despite this, many potential determinants of reputational considerations have been ignored in the literature. For instance, while much of the economics literature focuses on the interactions between formal sanctions and reputational
considerations, the process through which sanctions are imposed has received less attention. In recent work (Fluet and Mungan 2022), we noted this relationship and described how review standards can affect the informational content of incentives. On the other hand, existing legal scholarship on the subject assumes that a negligence regime must always produce superior information than a strict liability regime (Jacob and Shapira 2022), contrary to existing theories (e.g. Demougin and Fluet 2005, 2006). Thus, to gain a better understanding of the likely information production effects of liability regimes, we built a fairly general model that identifies a host of additional considerations (including the relative informativeness of the arrival and final review processes as well as heterogeneity in third parties’ knowledge regarding events).

Our analysis revealed some positive implications (see section 2) and normative implications (see section 3), which are conceptually separable from each other. From a positive perspective, our analysis revealed that, ceteris paribus, the overtness of harm as well as the informativeness of the review process increases the importance of reputational incentives. Moreover, reputational sanctions are decreasing in the strictness of the review standard when third parties must rely on the reviewer to obtain information about actors, and when the reviewer performs poorly in distinguishing between different types of actors. In all other cases, increasing the strictness of the review standard enhances reputational considerations when the review standard is very weak to begin with. From a normative perspective, we noted that in cases where the objective is to maximize reputational considerations, one ought to use review based incentives when the review process is sufficiently informative or when harms are sufficiently overt. Applied to the tort context, these results provide strong deterrence-based rationales for limiting the use of strict liability to exceptional cases.

In concluding, we note specific ways through which our preliminary analysis can be extended to study additional problems.

First, to focus on how review processes can impact reputational considerations, we took the monetary incentive \( s \) as given. Three questions related to formal sanctions naturally arise: (i) How does the optimal review standard (captured by \( a \)) respond to changes in \( s \)?; (ii) When the monetary incentive is a choice variable how is its optimum characterized?; and (iii) How do changes in other parameters impact the optimal standard and monetary incentive? Relatedly, we have not investigated how the optimal review standard responds to parameters (e.g. \( q \), \( p_0 \) and \( p_1 \)). These questions can be answered by conducting the necessary comparative statics and endogenizing some of the variables which we have already incorporated.

Second, to provide a general theory of how various considerations may affect reputational considerations, we imposed as few restrictions on the structure of our model as possible. Future work focusing on more specific contexts with more restrictive assumptions can generate more illuminating results. For instance, when studying random audits in administrative contexts it would be reasonable to set \( p_0 = p_1 \) to incorporate the non-informative selection for review. In this case, all reviews would be more informative than the ‘arrival process’, and therefore the statements in our lemmas and propositions that refer to this
requirement would be met.

Third, we have considered a single period setting, which implies that the reputational sanctions only affect the incentives of actors in committing a single act. In the criminal setting, for instance, this assumption is problematic, because the legal sanction stigmatizes offenders, which can in turn cause an increase in ex-offenders’ propensities to recidivate. This consideration can be incorporated by extending our analysis to a multi-period setting, in which a trade-off may emerge between deterring people without prior offenses and those with prior convictions (see, e.g. Mungan (2017)). Moreover, our observation regarding the possibility of splitting of liabilities and conduct announcements can be particularly relevant in this context.

As our very brief discussion of these potential extensions illustrates, there are many additional dimensions of reputational concerns that need to be studied. Our objective here is to move the theoretical analysis of reputational concerns one step further by identifying considerations which have thus far been neglected in the literature.

5 Appendix

Proof of Lemma 1: Define \( \beta_\gamma'(1) \) as the left-derivative at \( \alpha = 1 \). Because \( \beta_\gamma \) is increasing and concave, for \( \alpha > 0 \),

\[
0 \leq \beta_\gamma'(1) \leq \frac{1 - \beta_\gamma(\alpha)}{1 - \alpha}
\]

The result then follows from \( \lim_{\gamma \to \infty} \beta_\gamma(\alpha) = 1 \). □

Proof of Lemma 2: Condition (9) cannot hold for all \( x \) because \( p_0 > p_1 \) and

\[
\int_x^\infty \frac{f_0(x)}{f_1(x)} f_1(x) \, dx = 1
\]

Hence, (9) is true if and only if, for some \( x^* \in (x, \infty) \),

\[
p_0 f_0(x^*) / p_1 f_1(x^*) = 1
\]

Then (9) holds for \( x > x^* \) and the reverse inequality holds for \( x < x^* \). Defining \( \alpha_P \equiv F_1(x^*) \), it follows that \( \alpha_P \in (0, 1) \) and from equation (8)

\[
p_0 \beta'(\alpha_P) / p_1 = \frac{p_0 f_0(x^*)}{p_1 f_1(x^*)} = 1
\]

implying that \( p_0 \beta(\alpha) - p_1 \alpha \), a concave function, is maximized at \( \alpha_P \). □

In the proof that follows, we use the following notation:

\[
\hat{\zeta}_A \equiv \hat{w}_A - \hat{w}_A \quad (\text{Audit stigma})
\]

\[
\hat{\zeta}_P \equiv \hat{w}_A - \hat{w}_P \quad (\text{Punishment stigma conditional on audit})
\]

\[
\hat{\zeta}_E \equiv w_{AP} - \hat{w}_A \quad (\text{Stigma from non-exoneration; conditional on audit})
\]

\[
\hat{\zeta}_N \equiv w_P - \hat{w}_P \quad (\text{Unconditional punishment stigma})
\]

25
Proof of Lemma 3:

By using (22) $\hat{\Omega}_O(\alpha, g^r)$ can be expressed as:

$$\hat{\Omega}_O = (p_0 - p_1)\hat{\xi}_A + (p_0\beta - p_1\alpha)\hat{\xi}_P - (p_0(1 - \beta) - p_1(1 - \alpha))\hat{\xi}_E$$  \hspace{1cm} (26)

$$= (p_0 - p_1)(\hat{\xi}_A - \hat{\xi}_E) + (p_0\beta - p_1\alpha)(\hat{\xi}_P + \hat{\xi}_E)$$  \hspace{1cm} (27)

Recalling that $\hat{\Omega}_N = (p_0\beta - p_1\alpha)\hat{\xi}_N$, we have:

$$\hat{\Omega}_O - \hat{\Omega}_N = (p_0 - p_1)(\hat{\xi}_A - \hat{\xi}_E) + (p_0\beta - p_1\alpha)(\hat{\xi}_P + \hat{\xi}_E - \hat{\xi}_N)$$  \hspace{1cm} (28)

Next, note that

$$\hat{\xi}_P + \hat{\xi}_E - \hat{\xi}_N = (\hat{w}_A - w_P) + (\hat{w}_{AP} - \hat{w}_A) - (\hat{w}_P - \hat{w}_A) = \hat{w}_{AP} - \hat{w}_P$$  \hspace{1cm} (29)

where the three parentheses in (29) correspond to $\hat{\xi}_{\in\{P, E, N\}}$, openly written.

Substituting $\hat{w}_P = \frac{\mu_A}{\mu_A + \mu_{AP}} + \frac{\mu_{AP}}{\mu_A + \mu_{AP}} = \frac{(1 - \mu_A)}{(1 - \mu_A)}\hat{w}_A + \frac{\mu_{AP}}{(1 - \mu_A)}\hat{w}_{AP}$ into (29) we have that

$$\hat{\xi}_P + \hat{\xi}_E - \hat{\xi}_N = \frac{1 - \mu_A}{1 - \mu_P}(\hat{w}_{AP} - \hat{w}_A) = \frac{1 - \mu_A}{1 - \mu_P}(\{\hat{w}_{AP} - \hat{w}_A\} - \{\hat{w}_A - \hat{w}_A\}) = \frac{1 - \mu_A}{1 - \mu_P}(\hat{\xi}_E - \hat{\xi}_A)$$  \hspace{1cm} (30)

where the squiggly brackets in (30) correspond to the definitions of $\hat{\xi}_E$ and $\hat{\xi}_A$, respectively. Thus, letting

$$\chi \equiv (\hat{\xi}_A - \hat{\xi}_E)\frac{1 - \mu_A}{\mu_{AP}} = \{p_0 - p_1\} - (p_0\beta - p_1\alpha) + (\beta - \alpha)p_0p_1$$ \hspace{1cm} (31)

(where the equality is obtained by using the formula (19) we can re-write (28) to note that

$$\hat{\Omega}_O - \hat{\Omega}_N = \chi\frac{\tau}{(1 - \mu_A)\mu_{AP}} \left\{p_0 - p_1 - \frac{1 - \mu_A}{1 - \mu_P}(p_0\beta - p_1\alpha)\right\}$$ \hspace{1cm} (32)

Next, we demonstrate that $\chi$ and $\left\{p_0 - p_1 - \frac{1 - \mu_A}{1 - \mu_P}(p_0\beta - p_1\alpha)\right\}$ have the same sign. To see this, we note that

$$(p_0 - p_1) - \frac{1 - \mu_A}{1 - \mu_P}(p_0\beta - p_1\alpha) \geq \chi = (p_0 - p_1) - (p_0\beta - p_1\alpha) + (\beta - \alpha)p_0p_1$$  \hspace{1cm} (33)

and subsequently we re-write the terms in the squared brackets as convex combinations of two terms with different weights of $K$ and \((1 - K)\), as follows:

$$\mu_{AP} - p_1(1 - \mu_P) \geq (1 - K)p_0(1 - \beta) - p_1p_0\beta + K(p_1(1 - \alpha) - p_1 + p_1^2\alpha)$$  \hspace{1cm} (34)

and subsequently we re-write the terms in the squared brackets as convex combinations of two terms with different weights of $K$ and \((1 - K)\), as follows:

$$= (1 - K)(p_0(1 - \beta) - p_1p_0\beta) + K\{p_1(1 - \alpha) - p_1 + p_1^2\alpha\}$$

$$= (1 - K)\{(p_0 - p_1) - (1 - p_1)p_0\beta - K\{(1 - p_1)p_1\alpha\}$$

26
and

\[ \mu_A - p_0(1 - \mu_p) = (1 - K) \{ p_0(1 - \beta) - p_0 + p_0^2 \beta \} + K \{ p_1(1 - \alpha) - p_0 + p_0 p_1 \alpha \} \]
\[ = -(1 - K)(1 - p_0) p_0 \beta + K \{ (p_1 - p_0) - (1 - p_0) p_1 \alpha \} \]

Thus, the inequality in (33) can be re-written by collecting terms multiplied by \( K \) and \((1 - K);\) as follows:

\[ (1 - K) \{ p_0 \beta [(p_0 - p_1) - (1 - p_1) p_0 \beta] + p_1 \alpha (1 - p_0) p_0 \beta \} \]
\[ = -K \{ p_1 \alpha [(p_1 - p_0) - (1 - p_0) p_1 \alpha] + p_0 \beta (1 - p_1) p_1 \alpha \} \geq 0 \]

Collecting the common terms (of \( p_0 \beta \) in the first line and \( p_1 \alpha \) in the second line) we have:

\[ (1 - K) p_0 \beta \{ [(p_0 - p_1) - (1 - p_1) p_0 \beta] + p_1 \alpha (1 - p_0) \} \]
\[ = -K \{ p_1 \alpha [(p_1 - p_0) - (1 - p_0) p_1 \alpha] + p_0 \beta (1 - p_1) \} \geq 0 \]

Re-arranging terms reveals that this inequality is equivalent to

\[ \{ (1 - K) p_0 \beta + K p_1 \alpha \} \{ (p_0 - p_1) - (p_0 \beta - p_1 \alpha) + p_1 p_0 (\beta - \alpha) \} = \mu_p \chi \geq 0 \]

Thus, \( \{ (p_0 - p_1) - \frac{1 - \mu_A}{1 - \mu_p} (p_0 \beta - p_1 \alpha) \} \geq \chi \) whenever \( \chi \geq 0, \) which implies that \( \Omega_O - \bar{\Omega}_N > 0 \) whenever \( \chi \neq 0, \) and \( \Omega_O - \bar{\Omega}_N = 0 \) whenever \( \chi = 0. \)

Finally, we note that since \( \chi = (p_0 - p_1) - (p_0 \beta - p_1 \alpha) + (\beta - \alpha)p_0 p_1, \) it follows that \( \chi(0) = p_0 - p_1 > 0 = \chi(1), \) and \( \chi' = -p_0 \beta'(1 - p_1) + p_1 (1 - p_0) < 0 \) if \( \frac{p_1 (1 - p_0)}{p_0 (1 - p_1)} < \beta'(\alpha). \) Thus, if \( \beta'(1) \geq \frac{p_1 (1 - p_0)}{p_0 (1 - p_1)}, \) then \( \chi' < 0 \) for all \( \alpha \) implying that \( \chi > 0 \) for all \( \alpha < 1. \)

On the other hand, if \( \beta'(1) < \frac{p_1 (1 - p_0)}{p_0 (1 - p_1)}, \) then there exists \( \alpha' \) such that \( \chi'(\alpha') \geq 0 \) iff \( \alpha \leq \alpha'. \) This, in turn, implies the existence of a unique \( \bar{\alpha} \in (0, 1) \) such that \( \chi(\alpha) \geq 0 \) iff \( \alpha \leq \bar{\alpha}. \)

Thus, we have that if \( \beta'(1) \geq \frac{p_1 (1 - p_0)}{p_0 (1 - p_1)}, \) then \( \frac{\partial \Omega}{\partial g} > 0 \) for all \( \alpha \in [0, 1). \) Otherwise, there exists a unique \( \bar{\alpha} \in (0, 1) \) such that \( \frac{\partial \Omega}{\partial g}|_{\alpha = \bar{\alpha}} = 0. \)

**Proof of Lemma 4:** (i) Note that \( \frac{\partial \Omega_N(\alpha, g^\alpha)}{\partial \alpha} \) has the same sign as

\[ \frac{\partial}{\partial \alpha} \left( \frac{p_0 \beta(\alpha) - p_1 \alpha}{\mu_p(\alpha, g^\alpha)(1 - \mu_p(\alpha, g^\alpha))} \right) \]

which in turn has the same sign as

\[ Q(\alpha, g^\alpha) \equiv [p_0 \beta'(\alpha) - p_1] \mu_p(\alpha, g^\alpha)(1 - \mu_p(\alpha, g^\alpha)) \]
\[ - [p_0 \beta(\alpha) - p_1 \alpha] \frac{\partial \mu_p(\alpha, g^\alpha)}{\partial \alpha}(1 - 2 \mu_p(\alpha, g^\alpha)) \]

(39)
which leads to total expected costs of

\[ Q(\alpha, \gamma) = [p_0 \beta(\alpha, \gamma) - p_1] \mu_p(\alpha, \gamma) (1 - \mu_p(\alpha, \gamma)) \]

\[ - [p_0 \beta(\alpha, \gamma) - p_1 \alpha] \frac{\partial \mu_p(\alpha, \gamma)}{\partial \alpha} (1 - 2 \mu_p(\alpha, \gamma)) \]  

(40)

where \( \mu_p(\alpha, \gamma)(\alpha, \gamma) = p_0 \beta(\alpha, \gamma)(1 - K) + p_1 \alpha K \) and therefore \( \frac{\partial \mu_p(\alpha, \gamma)}{\partial \alpha} > 0 \).

The first claim. If the EGP is less informative than the arrival process, then \( p_0 \frac{\partial \beta(\alpha, \gamma)}{\partial \alpha} > 1 \), hence \( Q(\alpha, \gamma) < 0 \) if \( \mu_p(\alpha, \gamma) > 1/2 \). Next we show that \( Q(\alpha, \gamma) > 0 \) for all \( \alpha > 0 \), irrespective of \( \mu_p(\alpha, \gamma) \), if the EGP is sufficiently uninformative. It suffices to show that this is true for \( \gamma = 0 \), in which case \( \beta(\alpha, 0) = \alpha \) and therefore \( \psi(\alpha, 0) = \alpha \left[ p_0 (1 - K) + p_1 K \right] \). Substituting in (40) then yields

\[ Q(\alpha, 0) = (p_0 - p_1) [p_0 (1 - K) + p_1 K]^2 \alpha^2 > 0 \] for all \( \alpha > 0 \)

The second claim. By Lemma 2, if the EGP is more informative than the arrival process, then \( p_0 \frac{\partial \beta(\alpha, \gamma)}{\partial \alpha} < 1 \), hence \( Q(1, \gamma) < 0 \) if \( \mu_p(1, \gamma) \leq 1/2 \). Next we show that \( Q(1, \gamma) < 0 \), irrespective of \( \mu_p(1, \gamma) \), if the EGP is sufficiently informative. By Lemma 1, \( \lim_{\gamma \to \infty} \frac{\partial \beta(1, \gamma)}{\partial \alpha} = 0 \) and therefore \( \lim_{\gamma \to \infty} \frac{\partial \mu_p(1, \gamma)}{\partial \alpha} = p_1 K \). Substituting in (40) evaluated at \( \alpha = 1 \) then yields

\[ \lim_{\gamma \to \infty} Q(1, \gamma) = -p_0 p_1 \left[ p_0 (1 - K) - p_1 K \right] \]

\[ - p_1 (1 - p_0) [p_0 (1 - K) + p_1 K] K \]

\[ < 0 \]

which completes the proof.

(ii) From (23) it follows that \( \hat{\Omega}_O(0, g^c) = \hat{\Omega}_O(1, g^c) = \zeta_A(g^c) \) and \( \hat{\Omega}_O(\alpha, g^c) = \zeta_A(g^c) + \zeta_R(\alpha, g^c) \) with \( \zeta_R(\alpha, g^c) > 0 \) for all \( \alpha \in (0, 1) \). Thus, \( \hat{\Omega}_O(\alpha, g^c) \) has an interior maximizer, and \( \hat{\Omega}(\alpha, g^c) \) also has an interior maximizer for large \( q \), since \( \hat{\Omega}(\alpha, g^c)_{q=1} = \hat{\Omega}_O(\alpha, g^c) \).

Finally, \( \hat{\Omega} \) also has an interior maximizer if the EGP is more informative than the arrival process and \( \mu_p(\alpha, \gamma) \leq 1/2 \), or if the EGP is sufficiently informative, since then \( \hat{\Omega}_N \), like \( \hat{\Omega}_O \), has an interior maximizer as shown in part (i).

Proof of Proposition 1:

The unique unconditional incentive equilibrium threshold is given by \( g^s \) such that

\[ \hat{g}(1, \hat{\Omega}(1, g^s, q, \gamma)) = g^s \]

which leads to total expected costs of

\[ (p_0 - p_1) s + \hat{\Omega}(1, g^s, q, \gamma) \]

for all \( q \) and \( \gamma \)

since \( \hat{\Omega}_q(1, g^s, q, \gamma), \hat{\Omega}_s(1, g^s, q, \gamma) = 0 \).

Note that \( \frac{dg^s(\alpha, g^s, q, \gamma)}{dq} \geq 0 \) for all \( \alpha, q, \gamma \), since \( \frac{\partial \hat{\Omega}(\alpha, g^s, q, \gamma)}{\partial q} = \hat{\Omega}_O(\alpha, g^s, \gamma) - \hat{\Omega}_N(\alpha, g^s, \gamma) \geq 0 \) as note in lemma 3.
Thus, if \( \arg \max_{\alpha} g^\alpha(\alpha, 0, \gamma) < 1 \), then \( g^\alpha(\alpha^*, 0, \gamma) > g^\alpha \) for some \( \alpha^* < 1 \), and therefore \( g^\alpha(\alpha^*, \gamma, \gamma) > g^\alpha \) for all \( q \). If this is the case, let \( \hat{q}(\gamma) = 0 \). (In which case \( \alpha^*(q, \gamma) < 1 \) for all \( q \)).

But, if \( \arg \max_{\alpha} g^\alpha(\alpha, 0, \gamma) = 1 \), then note that (per lemma 3), \( \arg \max_{\alpha} g^\alpha(\alpha, q, \gamma) < 1 \) for all \( q > \hat{q} \), and \( \arg \max_{\alpha} g^\alpha(\alpha, q, \gamma) = 1 \) for all \( q \leq \hat{q} \). (In which case \( \alpha^* < 1 \) if \( q > \hat{q}(\gamma) \) and \( \alpha^* = 1 \) if \( q < \hat{q}(\gamma) \).)

Similarly, note that \( \frac{\partial g^\alpha(\alpha, q, \gamma)}{\partial q} \geq 0 \) for all \( \alpha, q, \gamma \). Therefore, \( \gamma' < \gamma'' \), implies that \( \hat{q}(\gamma') \geq \hat{q}(\gamma'') \).

Finally, note that \( \Omega_0(\alpha, g^\alpha, 0) = \zeta_A(g^\alpha) \) and \( \frac{\partial \Omega_0(\alpha, g^\alpha, 0)}{\partial \alpha} > 0 \). Thus, \( \frac{\partial \Omega(\alpha, \omega, \alpha, 0)}{\partial \alpha} > 0 \) for all \( \alpha \) and \( g^\alpha \) and for all \( q < 1 \). Therefore, when \( \gamma = 0 \), unconditional incentives are optimal for all \( q \) (and unconditional incentives lead to the same equilibrium threshold as any other \( \alpha < 1 \) in the exceptional case where \( p_0 = p_1 \)).

Thus, \( \hat{q}(0) = 1 \).

Similarly, note that \( \lim_{\gamma \to 0^+} \{ \arg \max_{\alpha} g^\alpha(\alpha, q, \gamma) \} = 0 \) for all \( q \). Therefore, \( \hat{q}(1) = 0 \).

**Proof of Proposition 2:** The proof is by contradiction. \( \alpha^* = 1 \) implies that \( (p_0 p(1) - p_1)s < 0 \), due to the EGP being more informative than the arrival process, and from Lemma 3, \( \frac{\partial \hat{q}(1, \omega(\gamma', 1))}{\partial \alpha} < 0 \) when \( \mu_p \leq 0.5 \) and the EGP is more informative than the arrival process. This is a contradiction, since the optimality of unconditional incentives requires \( \frac{\partial \hat{q}(1, \omega(\gamma', 1))}{\partial \alpha} \geq 0 \).

**References**


