A Theory of “Crying Wolf”: The Economics of Money Laundering Enforcement

Előd Takáts*
International Monetary Fund

The article shows how excessive reporting, called “crying wolf”, can dilute the information value of reports and how more reports can mean less information. Excessive reporting is investigated by undertaking the first formal analysis of money laundering enforcement. Banks monitor transactions and report suspicious activity to government agencies, which use these reports to identify investigation targets. Banks face fines should they fail to report money laundering. However, excessive fines force banks to report transactions which are less suspicious. The empirical evidence is shown to be consistent with the model’s predictions. The model is used to suggest implementable corrective policy measures, such as decreasing fines and introducing reporting fees. (JEL G28, K23, L51, M21)

1. Introduction
Money laundering gained notoriety: The FBI estimates money laundering volumes up to $1500 billion (Schroeder 2001). Furthermore, terrorists have joined the ranks of money launderers increasing national security concerns. Responding to that, law enforcement agencies paid increasing attention to money laundering offenses. Most importantly, they required banks to report and identify suspicious activities and threatened with fines if they fail to do so. Banks responded by exploding the number of reports, and many times they report innocent activities. The Simpson (2004a), for example, covers the case of the

*International Monetary Fund, 700 19th Street, NW Washington, DC 20431. Email: etakats@imf.org.

I am grateful for comments from Dilip Abreu, Patrick Bolton, Markus Brunnermeier, Avinash Dixit, Henry Farber, Robert Flood, Luis Garicano, Linda Goldberg, Peter Reuter, Hélène Rey, Esteban Rossi-Hansberg, David Skeie, Ádám Szeidl, Edwin Truman, Gábor Virág, and two anonymous referees. I am also grateful for comments from conference participants at the European Economic Association meeting (Budapest), the Econometric Society, European Winter Meeting (Istanbul), the Offshore and Underground Finance conference, and the Lawless Finance conference (Università Bocconi, Milan). I am thankful for comments for seminar participants at Princeton University, University of Michigan Ross School of Business, Universitat Pompeu Fabra, Tilburg University, European Central Bank, Federal Reserve Board, Federal Reserve Bank of New York, the Central Bank of Hungary, the International Monetary Fund, and the World Bank. I am also grateful for the advice from many experts from the Federal Reserve System, the Financial Crime Enforcement Network, and the banking industry. All remaining errors are mine.

The Journal of Law, Economics, & Organization
doi:10.1093/jleo/ewp018

© The Author 2009. Published by Oxford University Press on behalf of Yale University. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
falsely reported former presidential candidate and then Senate majority leader Bob Dole. The question of the article is: Is this reporting system efficient? And if not, how could it be made more efficient?

The article shows formally how excessive reporting in this situation fails to identify what is truly important by diluting the information value of reports. The intuition can be best understood through an analogy with the tale: “The boy who cried wolf.” In the tale, the boy cried wolf so often that his cries became meaningless. Similarly, excessive reporting, which will be referred to as “crying wolf,” fails to identify what is truly relevant. This crying wolf problem will be shown to be relevant in other economic situations, such as product information and auditing as well. More generally, the model shows that information is not only data but also able and expert identification of truly important data.

The reporting problem is investigated through the first formal analysis of money laundering enforcement. Choosing money laundering enforcement as the leading example is motivated by the fact that the identification role of reports is particularly strong. Furthermore, money laundering is an economically significant crime. Several hundred billion dollars are washed through the financial sector in the United States, and money laundering facilitates crimes as harmful as drug trafficking and terrorism, as detailed in Appendix A.

Money laundering enforcement relies on bank reporting to law enforcement and government agencies as reviewed in Reuter and Truman (2004). The suspicious activity report (SAR), a discretionary report introduced in 1996, is the most important reporting tool. The SAR is filed to the Financial Crime Enforcement Network (FinCEN) for any activity that the bank considers to be “suspicious.” The vague definition of what is suspicious does not allow money launderers to structure transactions so as to surely avoid reporting. SARs became prominent exactly because simple rule-based reporting can be avoided relatively easily. For instance, banks file the rule-based currency transaction report (CTR) for any cash transactions exceeding $10,000. Of course, money launderers can “smurf,” that is, break down large cash deposits over $10,000 into smaller deposits in order not to generate CTRs. However, the bank might spot and report several connected transactions just below $10,000 and identify them as suspicious because they hint at smurfing.

The reports summarize but do not fully capture why the bank believes the transaction to be suspicious. There are two main reasons for this coarseness in communication. On one hand, the formatting of reports does not allow infinitely fine elaboration on the considerations to file. The bank combines in the report its judgment, fine knowledge of the industry, and the specificities of the client’s area and line of business to identify what is suspicious. These considerations are left out by necessity. On the other hand, the bank might not want to identify certain reasons for filing a report fearing future liability or bad reputation. For instance, banks might prefer not to identify ethnicity of the client as a reason for finding the transaction suspicious.

Banks incur costly screening, monitoring, and reporting because of the threat of sanctions. Sanctions include nominal fines (civil money penalties),
the costs of public law enforcement actions (cease and desist orders or written agreements), the costs of private law enforcement actions (memoranda of understanding), and also implicit reputation costs. Failing to file SARs led to fines of $25 million for Riggs Bank (FinCEN 2004a), $24 million for Arab Bank (FinCEN 2005c), $50 million for AmSouth Bank (FinCEN 2004b; Braverman 2005), and $80 million for ABN-AMRO Bank (FinCEN 2005d). Most importantly, sanctions or fines are levied for false negatives, that is, for not reporting transactions which are later prosecuted as money laundering or judged to be suspicious ex post.\(^1\) Banks are not fined for false positives, that is, for reporting legal transactions as money laundering. This “safe harbor” provision further strengthens banks’ incentive to report.

The model explores theoretically the agency problem between the bank and government law enforcement agencies closely following the observed reporting setup. The bank monitors transactions and reports suspicious activity to the government, which identifies targets for investigations based on these reports. The bank undertakes costly monitoring and reporting because the government fines it if money laundering is successfully prosecuted and the bank did not report the transaction. Though Masciandaro (1999) abstracted from this agency problem in the first economic analysis of money laundering, it is crucial as the later survey by Masciandaro and Filotti (2001) shows.

The formal model builds on five main economic building blocks. First, communication is coarse between the bank and the government as the bank cannot communicate in a short report all the local information it has. This communication problem is similar in spirit to the information hardening problem in Stein (2002), though here the problem is not with verifying the information but rather with telling it precisely. Second, the bank’s incentives to report are coarse; the bank is fined only for false negatives, that is, for not reporting transactions which are prosecuted later as money laundering. Third, the bank is always uncertain about the transaction’s true nature, that is, every transaction can be potential money laundering. Fourth, the bank faces dual tasks: it has to monitor all transactions in order to report the suspicious ones. Fifth, the bank’s information, that is, its signal on the transaction, is not verifiable ex post because the local information at the time of the judgment cannot be reproduced later.

The model shows that harmful excessive reporting, called crying wolf, can arise in this setup. As the bank cannot share its signal with the government, the government must make decisions based on whether or not it observes the report. Intuitively, if the bank identifies all transactions as suspicious, then it fails to identify any one of them—exactly as if it would not have identified a single one. Thus, crying wolf can fully eliminate the information value of reports. Crying wolf can arise because excessively high fines for false negatives force the uncertain bank to err on the safe side and report transactions which are less

\(^1\) Regulatory language is somewhat different: banks are fined if suspicious activity detection and reporting procedures are not in place. Yet, the test of these policies is whether banks are able to identify and report those transactions which are considered to be suspicious ex post.
suspicious. In the extreme case, the bank is forced to report all transactions, thereby fully diluting the information value of reports.

The article shows that the model’s findings are consistent with the available empirical evidence in the United States. Fines have increased in the last 10 years, especially so after the USA Patriot Act. In response, banks have reported an increasing number of transactions. However, the number of money laundering prosecutions has fallen—even though the estimates of money laundering volumes have been stable. Furthermore, regulatory agencies have identified “defensive filing” which exhibits striking similarities with what happens under crying wolf.

The model also provides implementable policy implications on how to stop crying wolf and thereby increase the efficiency of money laundering enforcement. First, the model calls for reduced fines to cease crying wolf as optimal and not maximal fines are needed. The bank needs some fines in order to monitor and report, but excessively strong ones result in crying wolf. The intuition behind deviating from Becker’s (1968) seminal proposal of maximal deterrence is that banks are not criminals but rather honest informants. Thus, excessively strong incentives do not deter banks from committing a crime but rather distort their information provision.

Second, reporting fees might be needed to elicit optimal reporting. As in Holmström and Milgrom (1991), single dimensional incentives might not be able to elicit efficient two-dimensional banking effort to monitor and report. Reporting fees provide an implementable second dimension of incentives by punishing the bank for false positives. Furthermore, reporting fees can be thought of as pricing reporting externalities. Each report dilutes the value of all other reports, and reporting fees would make banks internalize these externalities.

Third, the model’s most important comparative static result shows that fines should decline in the harm caused by money laundering. The intuition is that the bank’s effective incentives have two components: government investigation to identify false negatives and nominal fines. As money laundering becomes more harmful, optimal government investigation increases as the marginal benefit of prosecuting money laundering increases. However, the bank’s incentives should be constant so as not to trigger crying wolf. Thus, the model shows that the fine increase of the USA Patriot Act, which reacted to the new threat of terrorism financing, could have been an intuitive but mistaken measure.

As it was argued earlier, crying wolf is a general economic problem. The main economic building blocks identified in money laundering enforcement can be found in many other situations. For instance, product information provision is very similar to suspicious activity reporting. Firms must use their expertise to identify the most relevant dangers related to using their product. There, coarse incentives are provided by the legal structure: omissions of warnings can result in damages and following lawsuits. False positives have no such easily identifiable victims. However, there is a real damage from crying wolf because information is lost as customers disregard warnings and accept contracts without reading the fine print.
A Theory of “Crying Wolf”

The agency setup of the model is very similar to that of the auditing problems analyzed first in Tirole (1986). The government uses both auditors and banks to obtain information about their clients. More precisely, the government is interested in learning if there are problems such as accounting omissions or signs of money laundering. Naturally, both auditors and banks are reluctant to provide such negative information, which creates an agency problem. These similarities in the agency setup allow building on the auditing model of Kofman and Lawarrèe (1993) in setting up the action set. A major difference from the crying wolf problem is, however, that the auditing literature focused on the disclosure of verifiable or certifiable information as reviewed in Verrecchia (2001).

The model’s focus on coarse communication of unverifiable information is particularly relevant to investigate the increased role of auditors after the Sarbanes-Oxley Act. Auditors not only are supposed to disclose verifiable information but also have to identify material transactions, that is, transactions that fundamentally affect the firm’s value. Identifying material transactions is very similar to identifying suspicious activities because in both cases unverifiable and uncertain information is identified through coarse communication. Furthermore, auditors are also sanctioned for false negatives, that is, for not disclosing transactions which later substantially affect the firm’s value. Thus, excessive fines might make auditors report more transactions as material, thereby failing to identify the truly important ones. Lengthening auditing reports and firm disclosures, documented in Gordon (2006), might well signal crying wolf.

The model can be extended to other settings. It questions the understanding in corporate finance that more disclosure is always better, which could be tested empirically following La Porta et al. (2006). A particularly interesting application of the crying wolf problem can arise in intelligence settings following the research started in Garicano and Posner (2005).

The rest of the article is organized as follows. Section 2 sets up the model. Section 3 solves the model and demonstrates crying wolf. Section 4 analyzes comparative statics. Section 5 links the model to available empirical evidence. Section 6 discusses the policy implications and other contexts where the model could be used. Section 7 concludes. Appendix A details most proofs and formal extensions such as numerical estimation of a nonlinear version of the model.

2. Model Setup

The model explicitly investigates the agency problem between the government and the bank, where the government needs information from the bank to investigate the transaction which may or may not be money laundering. The setup is illustrated in Figure 1.

If the bank monitors the transaction, then it receives information, a signal about the transaction. The informed bank that has received the signal is able to inform the government by filing an SAR. The government provides incentives, by means of fines, for the bank to monitor and report suspicious transactions.
2.1 Economy and Players

The economy of the model consists of a single money transfer. The transaction is either money laundering or a legitimate transfer. The prior probability that the transaction is money laundering is \( \alpha \in (0, 1/2) \). Money laundering causes harm \((h > 0)\) to society.\(^2\)

Two players are modeled explicitly: the government and the bank, who form a principal-agent relationship. Both players are risk neutral. The bank maximizes private profit, and the government maximizes social welfare.

The costs of monitoring and reporting, which are defined formally later, decrease the bank’s private profit. The profit is naturally decreased by the fine \((F)\). The transaction fee charged to the user and the bank’s cost of undertaking the transaction are normalized to zero. Hence, they do not affect the bank’s profit.

The harm caused by money laundering \((h)\) decreases social welfare. However, prosecuting money laundering increases social welfare by \(\rho h\), where \(\rho > 0\). Parameter \(\rho\) represents the recovery rate, the portion of harm prevented by the prosecution of money laundering. Utility \(\rho h\) can be interpreted as a

---

\(^2\) Money laundering is a crime of facilitation because it makes other crimes more lucrative, but it is not per se harmful. Hence, money laundering makes combating (and deterring) predicate crimes costlier. This effect is summarized by harm parameter \(h\).
A Theory of “Crying Wolf”

reduced-form representation of utilities from asset seizure, deterring money laundering and predicate crime, and from preventing future crimes by intercepting the money flow. Naturally, prosecuting money laundering is more useful the more harmful money laundering is. Fines do not affect social welfare as they represent simple transfers from the bank to the government. Finally, social welfare is decreased by government investigation costs and bank monitoring and reporting costs.

The agency problem arises because the bank does not internalize the social gains stemming from the prosecution of money laundering, \( \rho h \). Thus, to implement socially desirable policies (and to make the bank monitor and report), the government uses fines.

2.2 Timing

There are five periods in the model:

1. Nature selects the true nature of the transaction (legal or money laundering).
2. The bank exerts monitoring effort to learn the signal.
3. The informed bank, which has observed the signal, decides to report. (If the bank has not observed the signal, it cannot report.)
4. The government observes the report and sets investigation effort.
5. The government fines the bank if money laundering was prosecuted and the bank did not report the transaction.

2.3 Signal Structure

The bank might receive an informative signal \( \sigma \) about the transaction. The signal is assumed to be binary, and it takes either high (1) or low (0) value, \( \sigma \in \{0, 1\} \). In case of money laundering, the signal takes the high value 1 with probability \( \delta \) and the low value 0 with probability \( 1 - \delta \). For the legal transaction, the signal takes the low value 0 with probability \( \delta \) and the high value 1 with probability \( 1 - \delta \). Under symmetry, probability \( \delta \) can be interpreted as the precision of the signal. The higher \( \delta \), the more likely that the high signal indicates money laundering and the low signal indicates a legal transaction. The probabilities are summarized below:

Probability \( \delta \) is restricted to being more than one-half: \( \delta \in (1/2, 1) \). This implies that money laundering is more likely to trigger the high signal and legal transaction the low signal. Thus, observing a high signal implies that the transaction is more likely to be money laundering.

The posterior probabilities of money laundering can be determined through straightforward Bayesian updating. The posterior probabilities, \( \beta_0 \) and \( \beta_1 \), denote, respectively, the likelihood of money laundering given that the signal is low (\( \beta_0 \)) or high (\( \beta_1 \)):

\[
\beta_0 = \Pr(ML|\sigma = 0) = \frac{\alpha(1 - \delta)}{\alpha + \delta - 2\alpha\delta},
\]
\[
\beta_1 = \Pr(ML|\sigma = 1) = \frac{\alpha\delta}{1 - \alpha - \delta + 2\alpha\delta} > \beta_0.
\]
2.4 Action Sets

The government investigates transactions and fines the bank. The investigation effort \((I)\) exerted by the government determines the probability that the government discovers the truth about the transaction, that is, whether the transaction is legal or money laundering. Uncovered money laundering is prosecuted. As investigation effort \(I\) represents a probability, the government’s effort choice is naturally constrained to the unit interval. The investigation effort is costly, and it is assumed to be quadratic in the investigation effort with parameter \(k > 0: kI^2\). Furthermore, the government can condition its investigation effort only upon receipt of the bank’s report. The investigation effort chosen given no reporting is indexed as \(I_0\), and the investigation effort given reporting is indexed as \(I_1\). The government is able to commit to equilibrium, best response investigation actions. Thus, when multiple equilibria are possible the state can implement the equilibrium with the highest social welfare.

The government also imposes fines \((F \geq 0)\) on the bank. As discussed earlier, fines are socially costless to impose as fines represent welfare transfers from the bank to the government. The government imposes fines if the bank does not report the transaction, which is later prosecuted as money laundering. The government is able to commit to levying preset fines. In sum, the action set of the government has three elements: \((I_0, I_1, F) \in ([0, 1]^2 \times [0, \infty))\).

The bank monitors the transaction and reports to the government. Bank monitoring effort \((M)\) is assumed to be binary, and it takes either high \((M = 1)\) or low values \((M = 0)\), thus \(M \in \{0, 1\}\). The monitoring effort determines—as in Kofman and Lawarrée (1993)—the probability that the bank receives its signal \((\sigma)\). Thus, with high effort the bank always observes the signal, whereas with low effort it never observes the signal. The bank’s monitoring cost for high effort \((M = 1)\) is normalized to \(m > 0\) and for low effort \((M = 0)\) to zero.

If the bank did not receive the signal, it cannot report. If the bank received the signal, it sets the reporting threshold \(T \in \{0, 1\}\) and reports all signals weakly higher than \(T\) and does not report signals below it. In sum, the bank’s actions set has two elements: \((M, T) \in \{(0, 1)^2\}\).

Given that there is no time inconsistency in the game, the players could be thought of taking their action simultaneously at the beginning of the game.

---

3. In a more general model proposed in Appendix A, monitoring is handled as a probability taking values on the \([0, 1]\) interval. In the simpler setting proposed here, it can be viewed as a lump-sum investment.

4. The assumption relies on the empirical fact that banks need to provide detailed information in order to file the three-page SAR.

5. Notice that constraining \(T \in \{0, 1\}\) is without the loss of generality. Allowing for \(T \in [0, 1]\) would produce exactly the same equilibrium reporting thresholds, that is, either zero or one.

6. The bank’s actions set could be written more generally and allowing for reporting the low signal but not reporting the high signal transaction. As it is discussed in Appendix A, the more general problem is equivalent to the one presented here.
2.5 Final Assumptions

Six probability shortcuts are introduced to present the model in a concise form. The probability notations conditional on reporting threshold $T$ are summarized below:

- $p_T$: probability of reporting;
- $q_{0T}$: probability of money laundering given no reporting;
- $q_{1T}$: probability of money laundering given reporting.

The probability that the informed bank reports is denoted by $p_T$, where $T \in \{0, 1\}$. If $T = 0$, then the informed bank reports all signals, and thus $p_0 = 1$. If $T = 1$, then the bank only reports the high signal, thus $p_1 = 1 - \alpha - \delta + 2\alpha\delta$.

The probability of money laundering conditional on reporting is denoted by $q_{1T}$ and conditional on no reporting by $q_{0T}$. The probabilities depend on the reporting threshold $T$. Unit value of reporting threshold ($T = 1$) implies that the report perfectly conveys the bank’s signal. This would mean that the fact of reporting implies that the signal is high, thus $q_{11} = \beta_1$. Similarly, no reporting implies that the signal is low, consequently $q_{01} = \beta_0$. However, zero reporting threshold ($T = 0$) renders reporting uninformative. Thus, the probability of money laundering is the same with or without the report and equals the unconditional probability $q_{10} = q_{00} = \alpha$.

Four assumptions finalize the problem setup. First, the solution focuses on pure strategy subgame perfect Nash equilibria in order to ease interpretation. Second, it is assumed that parameters are such that in the first best equilibrium, when the agency problem is abstracted away, the bank monitors and reports. This assumption is consistent with the basic premise behind the money laundering enforcement regime, that is, banks need to monitor their clients. The necessary and sufficient parameter restrictions for the second assumption are

$$\frac{[\alpha(1-\delta)\rho h]^2}{4k(\alpha + \delta - 2\alpha\delta)} + \frac{(\alpha\delta\rho h)^2}{4k(1-\alpha-\delta+2\alpha\delta)} - \frac{(\rho h)^2}{4k} \quad > \quad (\alpha + \delta - 2\alpha\delta)c + m. \quad (1)$$

The parameter restriction implies that bank monitoring and reporting (right-hand side) are relatively cheap compared to the gains derived from prosecuting money laundering (left-hand side).

Third, it is assumed that government investigation is sufficiently costly to ensure interior investigation solutions. The formal parameter restriction is

$$\frac{\alpha\delta\rho h}{2(1-\alpha-\delta+2\alpha\delta)} < k. \quad (2)$$

In other words, restriction (2) implies that the government never sets first best investigation to unity. The reason for this is that investigation is so costly that some uncertainty in prosecution is preferred to spending the resources required to establish the truth with certainty.

---

7. Both equations (1) and (2) follow from the proof of Proposition 1 in Appendix A.
Finally, for tie-breaking it is assumed that whenever they are indifferent, both players take actions which are better for the other player. This can be interpreted as a weak understanding of common goals.

3. Solving the Model

The model is solved in three steps. The first subsection derives the benchmark first best equilibrium. The second subsection characterizes second best equilibria, where the bank’s incentive problem is analyzed. The third one analyzes the second best game under exogenously set fines. Most importantly, this subsection demonstrates the crying wolf problem. Namely, that excessively high fines trigger excessive reporting, which dilutes information and hinders money laundering prosecutions.

3.1 First Best Benchmark

The first best case, when the bank’s incentive problem is abstracted away, is investigated first. Using the probability shorthands derived before, the first best problem can be set as a social welfare \( W \) maximization problem:

\[
\max_{I_0, I_1, F, M, T} W = M[(1 - p_T)(q_{0T}I_0\rho h - kI_0^2) + p_T(q_{1T}I_1\rho h - kI_1^2 - c) - m] + (1 - M)[\alpha I_0\rho h - kI_0^2] - \alpha h. \tag{3}
\]

Interpreting equation (3) is straightforward. The first term shows welfare, when the bank is informed, which happens with probability \( M \). The informed bank, however, does not report with probability \( (1 - p_T) \). In this case, the probability that the transaction is money laundering is \( q_{0T} \). Money laundering is prosecuted then with probability \( I_0 \), which yields utility \( \rho h \). The government’s investigation cost is \( kI_0^2 \). The informed bank reports with probability \( p_T \). Then the probability that the transaction is money laundering is \( q_{1T} \) and money laundering is prosecuted with probability \( I_1 \). The government’s investigation cost is \( kI_1^2 \), and the bank incurs reporting cost \( c \). Finally, the bank has to incur cost \( m \) to be informed.

The second term depicts welfare when the bank is uninformed, which happens with probability \( (1 - M) \). Then the probability that the transaction is money laundering is the unconditional probability \( \alpha \). Money laundering is prosecuted with probability \( I_0 \). The government’s investigation cost is \( kI_0^2 \). The third and last term summarizes the social losses \( (h) \) stemming from money laundering, which happens with probability \( \alpha \).

Equation (3) immediately shows two properties of the first best problem. First, if there is no money laundering enforcement regime in place, that is, the bank does not monitor or report and the government does not investigate, then social welfare is \( -\alpha h \). Second, in the first best problem redistributive fines \( (F) \) do not play any role.

The first best solution can be determined in an intuitive manner by focusing on the limited number of bank actions. The government investigation best
responses to these bank actions can be determined straightforwardly. Furthermore, these government best responses allow easy quantification of the welfare resulting from each bank action pair. The bank action pair \((M = 1, T = 0)\) can be immediately excluded as it cannot be the first best equilibrium. Under this action pair, the bank provides useless information (reports all signals) at a positive monitoring \((m)\) and reporting \((c)\) cost. Welfare could be trivially improved by not reporting and not monitoring.

This leaves two possible first best equilibria. First, the equilibrium where the bank does not monitor \((M = 0)\) and does not provide any information to the government \((T \in \{0, 1\})\). Note that if the bank is uninformed, then the reporting threshold choice is irrelevant because the bank never reports. Second, the bank monitors \((M = 1)\) and reports only with high threshold \((T = 1)\). However, the parameter restriction set in equation (1) implies that the equilibrium with bank monitoring and high reporting threshold provides higher social welfare. Thus, it is the unique pure strategy first best equilibrium. Proposition 1 formalizes the argument.

**Proposition 1.** In the first best equilibrium, given condition (1) the bank monitors \((M = 1)\) and reports only the suspicious, high signal transaction \((T = 1)\). The government’s best responses conditional on no reporting \((I_0^*)\) and reporting \((I_1^*)\) are determined in terms of parameters in the following:

\[
I_0^* = q_{01} \rho h 2k = \frac{\alpha(1 - \delta) \rho h}{2k(\alpha + \delta - 2\alpha \delta)} < I_1^*, \tag{4}
\]

\[
I_1^* = q_{11} \rho h 2k = \frac{\alpha \delta \rho h}{2k(1 - \alpha - \delta + 2\alpha \delta)} < 1. \tag{5}
\]

The government does not levy fines \((F = 0)\) in the first best equilibrium.

Finally, in order to ease further discussion, equilibria that implement the first best welfare and actions are defined formally.

**Definition 1 (First Best equilibrium).** The equilibrium that implements the first best social welfare (denoted as \(W^*\)) and actions derived in Proposition 1, by using fines, is called First Best equilibrium with fine \(F\).

### 3.2 Second Best

In the second best problem, the government aims to maximize social welfare as in the first best problem (3) subject to the bank’s profit maximization (6) and (7):

\[
\Pi(M, T, I_0, I_1) = -M[(1 - p_T)q_0TF + p_Tc + m] - (1 - M)\alpha I_0F, \tag{6}
\]

\[
\text{IC} \quad \{M, T\} = \arg \max \Pi(M, T, I_0, I_1). \tag{7}
\]

Banking profit \((\Pi)\) shown in equation (6) has two main terms. First, the bank is informed with probability \(M\). The informed bank does not report with
probability \((1 - p_T)\); however, in this case the transaction might still be money laundering with probability \(q_{0T}\). The government uncovers money laundering with probability \(I_0\) and fines the bank with \(F\). The bank reports with probability \(p_T\) at filing cost \(c\). Finally, in order to be informed, the bank incurs monitoring cost \(m\).

Second, the bank is uninformed with probability \((1 - M)\) and then it cannot report. The transaction is money laundering with probability \(\alpha\), and the government prosecutes money laundering with probability \(I_0\), thereby fining the bank with \(F\).

Obviously, the government prefers to implement the First Best equilibrium, which requires actions \((M = 1, T = 1, I_0 = I_0^*, I_1 = I_1^*)\). The problem with implementing the first best actions is that incentive compatibility requires both weak and strong fines. On the one hand, fines should be high enough that the bank would not deviate to no monitoring \((M = 0)\). On the other hand, fines should be sufficiently low that the bank would not deviate and report all signals \((T = 0)\). If fines are set sufficiently low and high, then the first best is implemented because the government can commit to First Best equilibrium investigations. Lemma 2 summarizes the results.

**Lemma 2.** The First Best equilibrium can be implemented with fines on the \([F^*, F^{**}]\) interval only if \(F^* \leq F^{**}\), where

\[
F^* \equiv 2k(\alpha + \delta - 2\alpha\delta)^2 \left( (1 - \alpha - \delta + 2\alpha\delta)c + m \right) / \alpha^2\delta(1 - \delta)\rho h,
\]

\[
F^{**} \equiv 2kc(\alpha + \delta - 2\alpha\delta)^2 / \left[ \alpha(1 - \delta) \right]^2\rho h.
\]

**Proof.** The first best welfare level is implemented by equilibrium actions \((M = 1, T = 1, I_0 = I_0^*, I_1 = I_1^*)\). These actions are implementable in the second best if they are incentive compatible and maximize the bank’s profit. Formally,

\[
\text{IC}_1 \quad \Pi(M = 0, T \in \{0, 1\}, I_0 = I_0^*, I_1 = I_1^*) \leq \Pi(M = 1, T = 1, I_0 = I_0^*, I_1 = I_1^*),
\]

\[
\text{IC}_2 \quad \Pi(M = 1, T = 0, I_0 = I_0^*, I_1 = I_1^*) \leq \Pi(M = 1, T = 1, I_0 = I_0^*, I_1 = I_1^*).
\]

Starting with \(\text{IC}_1\):

\[-\alpha I_0^* F \leq - (1 - p_1)q_{01}I_0^* F - p_1 c - m,\]

\[
F^* \equiv \frac{p_1 c + m}{(\alpha - (1 - p_1)q_{01})I_0^*} = 2k(\alpha + \delta - 2\alpha\delta)^2 \left( (1 - \alpha - \delta + 2\alpha\delta)c + m \right) / \alpha^2\delta(1 - \delta)\rho h \leq F.
\]
Then IC2:

\[ -c - m \leq - (1 - p_1) q_0 I_0^* F - p_1 c - m, \]

\[ F \leq \frac{c}{q_0 I_0^*} = \frac{2kc(\alpha + \delta - 2\alpha \delta)^2}{[\alpha(1-\delta)]^2 \rho h} \equiv F^{**}. \]

Finally, the First Best equilibrium is surely implemented with fines such that \( F^* \leq F \leq F^{**} \) because the government can commit to \((I_0 = I_0^*, I_1 = I_1^*)\) investigation levels.

Fines can be both sufficiently weak and strong if reporting is sufficiently costly \((c)\) compared to monitoring costs \((m)\). Intuitively, if reporting is overly cheap, then fines strong enough to make the bank monitor (and incur cost \(m\)) will also force them to report all the signals (at cost \(c\)). Thus, fines can take intermediate values if the monitoring costs are not too large compared to filing costs. The reasoning is formalized in Lemma 3.

**Lemma 3.** The First Best equilibrium is implementable if and only if reporting is sufficiently costly compared to bank monitoring

\[ m \leq \frac{(1-\alpha)(2\delta-1)}{(1-\delta)} c. \]  

\[ (8) \]

**Proof.** Follows from Lemma 2 and noting that condition (8) is equivalent with \( F^* \leq F^{**} \). \( \Box \)

Lemma 3 shows that the reporting cost penalizes for false negatives and thus discourages the bank from crying wolf. Consequently, it provides one of the model’s main policy insights, namely that increasing reporting costs might be crucial in implementing the First Best equilibrium. Reporting fees could raise the right-hand side of condition (8) until the first best is implementable.

In order to ease the following discussion, note that all possible equilibria with zero bank monitoring \((M = 0)\) share a number of properties and, most importantly, yield the same social welfare as summarized in Corollary 4.

**Corollary 4.** All equilibria with no bank monitoring \((M = 0)\) yield the same social welfare \((W^{**})\), and government investigation conditional on no reporting \((I_0)\) is

\[ I_0 = I^{**} \equiv \frac{\alpha \rho h}{2k}. \]

Bank reporting threshold and government investigation conditional on reporting are undetermined, \(T \in \{0,1\}\) and \(I_1 \in [0,1]\). Furthermore, the resulting welfare \((W^{**})\) is lower than the First Best welfare: \(W^{**} < W^*\).

**Proof.** Follows from the proof of Proposition 1. \( \Box \)

The equilibria characterized in Corollary 4 are called Second Best equilibria. The reason is that, as it will be proven in Proposition 5, one of these equilibria will prevail if the First Best is not implementable.
Definition 2 (Second Best equilibria). Equilibria with no bank monitoring ($M = 0$) and implemented by fine $F$ are called Second Best equilibria with fine $F$.

Notice that the definition allows to distinguish between several different second best equilibria based on the fine level.

Next, it is shown that Second Best equilibria indeed prevail if the First Best is not implementable. The intuitive reason is that if the First Best is not implementable, then bank monitoring ($M = 1$) implies that the bank reports all signals ($T = 0$). However, if the SAR is filed irrespective of the signal, then it does not convey any information. Thus, the government rather implements no bank monitoring ($M = 0$) to save on effort costs. Furthermore, the government then sets fines to zero as this maximizes the profit of the bank and does not change the social welfare function. Proposition 5 summarizes the results.

Proposition 5. The First Best equilibrium is implementable (and it is implemented) with $F = F^*$ if condition (8) holds. Otherwise, Second Best equilibria are implemented with zero fines ($F = 0$).

3.3 Second Best with Exogenous Fines

In this subsection, the second best incentive compatible game is analyzed under exogenous fines. The question is asked: “How does fine setting affect the efficiency of the money laundering enforcement regime?” For the sake of tractability, this subsection focuses on the parameter setup where the first best is implementable, that is, when inequality (8) holds. Equilibria for the parameter setup when the first best is not implementable are detailed in Appendix A.

The equilibria under exogenous fine setting follow mostly from the characterization of the Second Best equilibria. First, as Lemma 2 shows, fines on the $[F^*, F^{**}]$ interval implement the First Best equilibrium. Second, the Second Best equilibria are implemented with low fines, when the bank prefers paying fines to exerting monitoring and reporting effort. Sufficiently low fines are set formally in Corollary 6.

Corollary 6. Second best equilibria prevail if fines are sufficiently low: $F \leq F'$, where

$$F' \equiv \min \left\{ \frac{2k(1 - \alpha - \delta + 2\alpha\delta)c + m}{\alpha\delta\rho h}, \frac{2k(\alpha + \delta - 2\alpha\delta)(c + m)}{\alpha^2\rho h} \right\}.$$

(9)

Third, if neither the First Best nor the Second Best equilibria can be implemented, then only one more pure strategy bank action pair remains: when the bank monitors ($M = 1$) and reports all transactions ($T = 0$). As it turns out, all equilibria in which the bank monitors and reports are characterized by the same social welfare and similar government investigation levels. Corollary 7 summarizes the result.
Corollary 7. All equilibria in which the bank monitors \((M = 1)\) and reports all transactions \((T = 0)\) yield identical welfare \((W^{**})\). Moreover, in all such equilibria the government investigates, given reporting, with effort \(I_1 = I^{**}\). Furthermore, the resulting welfare \((W^{**})\) is lower than that of Second Best equilibria: \(W^{***} < W^{**} < W^{*}\).

Proof. The proof follows from the proof of Proposition 1. □

The similarities among the equilibria entailing bank action \((M = 1, T = 0)\) allow these equilibria to be called Third Best according to their welfare rank.

Definition 3 (Third Best equilibria). The equilibria in which the bank monitors \((M = 1)\) and reports all transactions \((T = 0)\), and which are implemented with fines \(F\) and with investigation effort \(I_0\) in case of no reporting, are called Third Best equilibria with \(F\) and \(I_0\).

Third best equilibria are implemented by high fines. Intuitively, strong enough fines cause the bank not only to monitor but also to report all signals. Lemma 8 formalizes the argument.

Lemma 8. Sufficiently large fines \(F \geq F^{**}\) implement the Third Best equilibria if the First Best equilibrium is implementable. Then government investigations given no reporting are set larger than \(I'_0(F)\), \(I_0 \geq I'_0(F)\), where

\[
I'_0(F) = \max \left\{ \frac{(\alpha + \delta - 2\alpha\delta)c}{\alpha F}, \frac{c + m}{\alpha F} \right\}.
\]

Furthermore, \(I'_0(F) < 1\) for \(\forall F \geq F^{**}\).

The above results allow for characterizing the Second Best equilibria with exogenous fine setting. Proposition 9 summarizes the results of the subsection and characterizes the equilibria with exogenous fines.

Proposition 9. Fines implement the following equilibria if the First Best equilibrium is implementable:

\[
F \leq F' \quad \text{Second Best equilibria}
\]

\[
F^* \leq F \leq F^{**} \quad \text{First Best equilibrium}
\]

\[
F^{**} < F \quad \text{Third Best equilibria with } I_0 \geq I'_0(F).
\]

There is no pure strategy equilibrium with fines \(F' < F < F^*\).

Proof. The proof follows directly from Lemma 2, Corollary 6, and Lemma 8. □

Finally, there is no pure strategy equilibrium with fines in the \((F', F^*)\) region. Fines that are too high for the second best are too low for the first best. The reason is that fines represent incentives only with government investigation conditional on no reporting. In Second Best equilibria, investigation conditional on no reporting is higher than in the First Best equilibrium.
(I^*_0 > I^{**}). Thus, fines that are too strong with high investigation (I^*_0) turn out to be too weak with low investigation (I^{**}).

3.3.1 Crying Wolf. Proposition 9 clearly shows that fines first increase, but later on decrease welfare. Moreover, very high fines lead to lower social welfare (W^{**}) than no fines at all (W^{**}).

This subsection shows that the effects are graver than a simple reduction in welfare. Even the prosecution rate, that is, the probability that money laundering is prosecuted, decreases with excessive fines. Prosecution rates are the highest in the First Best equilibrium. However, the rate is the same in the Second and the Third Best equilibria. Intuitively, what matters for the success of government investigation is information supplied by the bank. However, reporting no transaction or reporting all transactions is equally uninformative in identifying the most likely suspects. Lemma 10 formalizes the result.

**Lemma 10.** Fines first increase and later decrease the likelihood that money laundering is prosecuted (χ) if the First Best equilibrium is implementable. The expected prosecution rate in the three possible equilibria is set as

\[
F \leq F' \quad \text{Second Best equilibria} \quad \chi^{**} = \alpha I^{**}
\]

\[
F^* \leq F \leq F^{**} \quad \text{First Best equilibrium} \quad \chi^* = \alpha(1 - \delta)I^*_0 + \alpha \delta I^*_1 > \chi^{**}
\]

\[
F^{**} < F \quad \text{Third Best equilibria with } I_0 \geq I'_0(F) \quad \chi^{**}
\]

Lemma 10 highlights the crying wolf problem. In the case of coarse communication, some scarcity of reporting is desirable. Providing more reports can be detrimental because the additional reports dilute the value of existing ones. As shown in the model, if the bank reports all transactions, then the information value of the reports is eliminated. This dilution of information is called crying wolf and is defined formally below.

**Definition 4 (Crying Wolf).** Crying wolf arises when excessive reporting dilutes the information value of reports. In the extreme case of crying wolf, reports become completely uninformative.

Formally, the extreme version of crying wolf arises in the Third Best equilibria when the bank monitors and also reports all transactions as reports become completely uninformative.

3.3.2 Information Laffer Curve. The result of Lemma 10 can also be interpreted as the information Laffer curve. Prosecutions can be drawn as a function of fines. Lemma 10 shows that the number of expected prosecutions first rises with fines, but eventually it falls back to the original low level. This is called the information Laffer curve. The qualitative results are illustrated in Figure 2.
In order to illustrate the workings of the model, the expected reporting figures are also charted as a function of fines.

Due to the discrete structure of the model, the information Laffer curve shows jumps. However, in a continuous model, investigated and numerically simulated in Appendix A, the curve not only survives but also becomes smooth continuous. As Figure 8 of Appendix A shows, increasing fines monotonically increase the expected number of reports, but the expected number of convictions shows a continuous hump-shaped curve.

The information Laffer curve shows an unusual property of the model. In most law enforcement problems, increasing incentives increases prosecution. Thus, if fighting a particular type of crime is important, uncertain decision makers might want to err on the safe side and introduce strong incentives. In the money laundering enforcement problem, this logic produces counterproductive results. In fact, as Second Best equilibria welfare dominate Third Best equilibria, uncertain decision makers should actually prefer lower fines, if anything.

4. Comparative Statics

4.1 First Best Investigation: $I_0^*$ and $I_1^*$

As expected, investigation efforts ($I_0^*$ and $I_1^*$) increase both in the harm caused by money laundering ($h$) and in the recovery rate ($\rho$). Intuitively, the marginal benefit from the prosecution of money laundering is higher with higher $h$ or $\rho$, hence the government prefers to undertake more investigation. Similarly, investigation efforts (again both $I_0^*$ and $I_1^*$) decrease in the cost of government investigation ($k$). Clearly, if the marginal cost of investigation is higher, then less investigation is undertaken.

More interestingly, investigation efforts also depend on the prior likelihood of money laundering ($\alpha$) and the precision of the signal ($\delta$). The more likely money laundering is, the higher both investigation efforts are. The intuition is that higher likelihood of money laundering, ceteris paribus, increases the marginal benefit of investigation as more laundering is prosecuted with the
same effort expenditure. Formally,
\[
\frac{\partial I_0^*}{\partial \alpha} = \frac{\delta (1 - \delta) \rho h}{2k(\alpha + \delta - 2\alpha \delta)^2} > 0,
\]
\[
\frac{\partial I_1^*}{\partial \alpha} = \frac{\rho h}{2k(1 - \alpha - \delta + 2\alpha \delta)^2} > 0.
\tag{10}
\]

However, the precision of the signal affects the two investigation efforts differently:
\[
\frac{\partial I_0^*}{\partial \delta} = -\frac{\alpha (1 - \alpha) \rho h}{2k(\alpha + \delta - 2\alpha \delta)^2} < 0,
\]
\[
\frac{\partial I_1^*}{\partial \delta} = \frac{\alpha (1 - \alpha) \rho h}{2k(1 - \alpha - \delta + 2\alpha \delta)^2} > 0.
\tag{11}
\]

The intuition is that the more precise the signal is, the more reliable first best reporting is. Thus, the government investigates more vigorously given reporting and less so in the absence of it.

4.2 Minimal Optimal Fine: \( F^* \)

Two properties of the minimal optimal fine are crucial to understand the comparative statics. First, the minimal optimal fine provides incentives for the bank to stay informed and not deviate from monitoring \( (M = 1) \), as can be seen from the IC1 constraint in the proof of Lemma 2. Second, the minimal optimal fine depends on the first best investigation effort conditional on no reporting \( (I_0^*) \). The reason is that fines represent incentives only together with \( I_0^* \) as the bank is fined only if money laundering is prosecuted. Equation (12) depicts fines as a function of first best investigation \( (I_0^*) \), bank costs \( (c, m) \), prior probability of money laundering \( (\alpha) \), and signal precision \( (\delta) \):

\[
F^* \equiv \frac{(1 - \alpha - \delta + 2\alpha \delta)c + m}{\alpha \delta I_0^*}.
\tag{12}
\]

Evidently from equation (12), the minimal optimal fine decreases in the first best investigation level \( (I_0^*) \), which has three unexpected consequences.

First, the minimal optimal fine decreases in the harm caused by money laundering \( (h) \) and in the recovery rate \( (\rho) \). Equation (4) shows that higher harm and higher recovery rates increase investigation. Thus, increased investigation provides additional incentives for the bank to stay informed and lower fines suffice.

Second, the minimal optimal fine also decreases in the prior probability of money laundering \( (\alpha) \). Two factors simultaneously decrease the optimal fine. First, increasing \( \alpha \) increases the likelihood of uncovering money laundering should the bank choose to be uninformed. This means that smaller fines suffice, which is demonstrated in the following equation:

\[
\frac{\partial}{\partial \alpha} \frac{(1 - \alpha - \delta + 2\alpha \delta)c + m}{\alpha \delta} = -\frac{(1 - \delta)c + m}{\alpha^2 \delta} < 0.
\]
Second, increasing the prior probability of money laundering increases the investigation effort, which is apparent in equation (10). Increasing investigation, as before, further reduces the fine level. Thus, both effects cause the minimal optimal fine to decrease in the prior likelihood of money laundering.

Third, the minimal optimal fine increases in investigation costs $k$. Higher investigation costs decrease first best investigation by equation (4). Thus, higher fines are needed to preserve the bank’s incentives to stay informed.

The minimal optimal fine also decreases in the precision of the signal ($\delta$). However, the intuition of this result is not immediately obvious. On the one hand, as shown in equation (11), higher precision of the signal implies lower investigation effort. This effect requires increased fines to preserve incentives. On the other hand, higher precision also increases the value of being informed for the bank. The informed bank is more certain about its information, and the likelihood of fines, given the low signal, is lower. Formally, this can be demonstrated as

$$\frac{\partial}{\partial \delta} \frac{(1-\alpha-\delta+2\alpha\delta)c+m}{\alpha\delta} = -\frac{(1-\alpha)c+m}{\alpha\delta^2} < 0.$$ 

In order to evaluate the relative strength of the contradicting forces, $F^*$ is formally differentiated:

$$\frac{\partial F^*}{\partial \delta} = \frac{-1}{(\alpha\delta)^2 I_0^*} \left[ \frac{\alpha(1-\alpha)(2\delta-1)}{(1-\delta)(\alpha+\delta-2\alpha\delta)} c + \frac{\alpha(2\delta-1)+\delta^2(1-2\alpha)}{(1-\delta)(\alpha+\delta-2\alpha\delta)} m \right] < 0.$$ 

The result shows that the second effect is dominant, and the optimal fine decreases in signal precision.

The remaining comparative statics are straightforward. The optimal minimal fine increases in the bank’s costs of monitoring ($m$) and reporting cost ($c$). This is intuitive as the greater the effort costs are, the higher the incentives must be.

4.3 Maximal Optimal Fine: $F^{**}$

Two properties of the maximal optimal fine are crucial in understanding the comparative statics. First, maximal optimal fine shows how large the fine can be before the bank deviates to reporting every transaction ($T = 0$), as can be seen from the IC$_2$ constraint in the proof of Lemma 2. Second, maximal optimal fine depends on the first best investigation effort given no reporting ($I_0^*$) because fines represent incentives only with investigation. Thus, fine $F^{**}$ can be written formally as follows:

$$F^{**} = \frac{(\alpha+\delta-2\alpha\delta)c}{\alpha(1-\delta)I_0^*}. \quad (13)$$

Equation (13) allows the analysis of $F^{**}$ similarly to that of $F^*$. The results are very similar and show the same three unexpected consequences.

First, maximal optimal fine $F^{**}$ decreases in the harm caused by money laundering ($h$) and in the recovery rate ($\rho$). The reason, in both cases, is that equilibrium investigation $I_0^*$ increases in $h$ and $\rho$ as equation (4) shows. Thus,
the bank needs smaller nominal fines so as not to cry wolf because the probability of prosecuting money laundering is higher.

Second, maximal optimal fine also decreases in the prior probability of money laundering \((\alpha)\). The reason is two-fold. First, higher prior likelihood of money laundering increases investigation \(I^*_0\) as equation (10) shows. Second, higher prior money laundering probability implies that even with the low signal the transaction is more likely to be money laundering. Thus, ceteris paribus, it is more likely that the bank will be fined if it does not report. So, the bank needs smaller fines so as not to report all transactions. This effect is demonstrated by

\[
\frac{\partial}{\partial \alpha} \left( \frac{\alpha + \delta - 2\alpha \delta}{\alpha (1 - \delta)} \right) c = \frac{-\delta c}{\alpha^2 (1 - \delta)} < 0.
\]

Thus, both effects decrease maximal optimal fine \(F^{**}\) in the prior likelihood of money laundering.

Third, maximal optimal fine increases in investigation costs \(k\). Higher investigation costs affect maximal optimal fines through decreasing investigation \(I^*_0\). Thus, higher fines are needed to keep incentives constant.

Maximal optimal fine also increases in the precision of the signal \((\delta)\). First, investigation \(I^*_0\) decreases in \(\delta\) by equation (11). Intuitively, the more precise the signal, the less investigation is needed if no report was filed in the First Best equilibrium. Second, higher precision decreases the likelihood that the low signal activity is indicative of money laundering. Thus, the bank has lower incentives to report low signal transactions. This latter effect is demonstrated formally by

\[
\frac{\partial}{\partial \delta} \left( \frac{\alpha + \delta - 2\alpha \delta}{\alpha (1 - \delta)} \right) c = \frac{(1 - \alpha) c}{\alpha (1 - \delta)^2} > 0.
\]

Thus, both effects increase maximal optimal fine \(F^{**}\) in signal precision.

The remaining results are trivial. Maximal optimal fine increases in the cost of reporting \((c)\) as more costly reporting provides disincentives against crying wolf. Furthermore, it is unaffected by changes in the monitoring cost \(m\) because the bank keeps monitoring even if it deviates from the first best to cry wolf.

Finally, the comparative statics are summarized concisely on the table below.

<table>
<thead>
<tr>
<th>Money laundering ((\alpha))</th>
<th>Legal transaction ((1 - \alpha))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low signal ((0))</td>
<td>(1 - \delta)</td>
</tr>
<tr>
<td>High signal ((1))</td>
<td>(\delta)</td>
</tr>
</tbody>
</table>

5. Empirical Evidence

The model provides two predictions that can be investigated empirically:

**Prediction 1 (Fines)** The number of reporting is monotonically increasing in fines. However, the number of prosecutions shows a hump-shaped curve: prosecutions first increase later decline in fines.
Prediction 2 (Harm) *Even with constant fines, increasing harm of money laundering can also trigger more reporting and fewer prosecutions.*

Data are only available on a national level. In order to link national-level data to the single bank reporting model, the available data are interpreted as the aggregate outcome of many reporting problems. The predictions are evaluated by analyzing the estimates of money laundering volumes, the evolution of fines, the perceived harm, the number of reports, and prosecutions.

5.1 Money Laundering Volumes

Money laundering volume estimates are shown as a rudimentary control to allow focusing on reports, fines, and prosecutions. If money laundering volumes do not change substantially, as it is the case, then changes in reporting and fines might explain the evolution of prosecution figures.

Money laundering volumes need to be estimated as hard data are obviously not available. Estimates fall into two broad categories. First, microeconomic estimates use micro-level data and link it to money laundering or related proxies. Second, macroeconomic estimates use macrovariables to estimate money laundering or its proxies.

Estimating criminal earnings follows the microeconomic approach. Criminal earnings need to be “laundered”; hence, they are useful proxies for money laundering volumes. The estimation is done in two steps: First, earnings related to particular kind of crimes (such as an act of tax evasion or a drug dealing) are estimated. Second, the number of criminal acts is estimated using prosecution data and latency estimates. Figure 3 demonstrates that criminal earnings estimates show a remarkable stability in the United States over a long period of time.

Estimating the size of the underground economy by cash demand follows the macroeconomic approach. The size of the underground economy is a proxy...
for the amount of monies laundered because underground incomes need to be “laundered” eventually. Again, there are no hard data on the underground economy. However, the amount of cash demanded and used in an economy (relative to its development and characteristics of financial intermediation) can be used to gauge the size of the underground economy (Tanzi 1996). The size of the underground economy is estimated by using variations in the tax rate: Higher tax rate provides stronger incentives for informality. Hence, the difference between the actual cash demand and the cash demand estimate under a hypothetical zero tax schedule is a proxy for the size of the underground economy. Using this methodology, Schneider (2003) estimates the US underground economy to be around 8.8% of Gross Domestic Product (GDP), which is remarkably close to the criminal earnings estimate above.

In sum, both methodology estimates money laundering volumes, or its proxies, to be stable. In spite of the caveats above, there does not seem to be strong trends in money laundering volumes. Hence, the focus on the interaction between fines, harms, reporting, and the prosecution might be justified.

5.2 Fines
Fines have increased strongly between 1994 and 2006. However, analyzing fines is not straightforward due to fragmented databases. Hence, a number of potential fine measures are investigated here.

Fine and law enforcement databases are fragmented along with US banking and anti money laundering (AML) regulating agencies. Reflecting this fragmented regulatory setup, banks have been fined for AML violations by many different institutions such as FinCEN or the Office of the Comptroller of the Currency (OCC). For instance, ABN AMRO Bank was fined for $30 million by FinCEN and for further $50 million by other government agencies (FinCEN 2005d). Unfortunately, there is no comprehensive database to assess, which is comparable across time.

The civil money penalties levied by FinCEN provides a consistent database, and it is displayed on Figure 4. Unfortunately, the database only spans the 1999–2005 period. Though FinCEN was established in 1990 and was granted...
regulatory responsibilities in 1994, it issued fines first in 1999. However, the data span most of the period since the introduction of SARs in 1996. In any case, the FinCEN database is consistent with rapidly increasing fines during the observed period.

Another potential channel to assess fines is to investigate all fines and restitutions. This source unfortunately includes fines not only on banks but also on nondepository institutions and individuals. Reuter and Truman (2004) report that all fines and restitutions increased six-fold between 1996 and 2001. The increase is monotonic and the increase comes from larger, not more frequent fines and restitutions further confirms larger fines. This piece of evidence further supports the increasing fines hypothesis.

Furthermore, some qualitative statements and examples might be examined. The reason is that easily quantifiable civil money penalties constitute only a small part of all fines, and substantial elements of fines are difficult to quantify. For instance, the costs of public law enforcement actions (cease and desist orders or written agreements) for banks are not straightforward to calculate. The costs associated with unobserved private law enforcement actions (memoranda of understanding) are even harder to discern. Hence, there is a role of qualitative assessments. Braverman’s (2005) report along with many interviewed industry professionals shows that there has been substantial tightening of punishments for AML violations.

Finally, the example of the Bank of New York provides further evidence for increasing fines. The Bank of New York was not fined in 1998. However, AmSouth and ABN AMRO Bank received stiff penalties for very similar offenses in 2004 and 2005, respectively.

In sum, though there is no perfect database for money laundering enforcement fines, every available piece of evidence points toward strong and monotonic increase to fines and punishments since the introduction of SARs in 1996.

5.3 Harms
The perceived harm of money laundering has also increased considerably. Increases of fines themselves show the growing impatience of regulators with money laundering offenses. Most importantly, the 9/11 attacks have highlighted and increased awareness of terrorism as a predicate crime. Money laundering regulation was strengthened substantially in the USA Patriot Act in order to increase safety.

5.4 Reporting
Reporting has also increased steeply. Reporting data are published in FinCEN (2006). The number of SARs has exploded in the last 10 years as evidenced in Figure 5. The figure normalizes reporting by the population increase of the United States in order to correct for natural growth.8 Furthermore, charting

---

8. Population data were taken from the Economist Intelligence Unit.
only the total number of SARs is misleading as the range of reporting institutions has widened substantially. Hence, SARs filed by depository institutions are displayed separately. As it is evident from Figure 5, even depository SARs grew exponentially as fines grew, and the growth does not seem to saturate. Furthermore, the growth seems to have accelerated after 9/11.9

5.5 Prosecution

Prosecution figures show a hump-shaped curve, a marked difference from the rapid increase of reporting or fines. Prosecution can be measured in two main ways.10 First, the number of money laundering cases filed provides a measure how efficient SARs are in tipping off law enforcement agencies to prosecute cases. Figure 6 shows the evolution of cases filed—normalized by population numbers. The chart clearly shows that money laundering cases filed have increased until 1999 and declined thereafter. Second, the number of money laundering convictions measures how efficient SARs are in providing useful evidence to convict criminals. Figure 7 shows the same hump-shaped curve as Figure 6, again normalized by population numbers.

---

9. Between 1996 and 2000, the number of SAR filings by depository institutions grew by 27,663 per year on average. Between 2001 and 2005, depository SAR filings grew on average by 71,987 per year. Total SAR filing grew even faster, from 27,758 report-increase/year to 151,209.

10. Both kinds of prosecution data are obtained from the Federal Justice Statistics Resource Center Web site (http://fjsrc.urban.org/index.cfm) query system until the latest date available: 2005. The query system reports the title and section of the most serious terminating offense. It also implies that the data are not directly comparable to the Money Laundering Special Report (Motivans 2003), which uses the title and section of the most serious filing offense.
5.6 Qualitative Evidence

Finally, official testimonies also support the model’s predictions. FinCEN (2005b) describes “defensive filing” as strikingly similar to the effects of crying wolf. First, the number of SARs filed skyrockets. Second, banks file SARs regardless of the level of suspicion. Third, law enforcement efforts are compromised by the large number of filings. A FinCEN (2005a: 3) quote illustrates the point:

We estimate that if current filing trends continue, the total number of Suspicious Activity Reports filed this year will far surpass those filed in the previous years. [...] it fuels our concern that
financial institutions are becoming increasingly convinced that the key to avoiding regulatory and criminal scrutiny under the Banking Secrecy Act is to file more reports, regardless of whether the conduct or transaction identified is suspicious. [...] If this trend continues, consumers of the data—law enforcement, regulatory agencies, and intelligence agencies—will suffer.

In sum, the available empirical evidence is consistent with the model’s predictions. The increases in fines and in the perceived harm of money laundering have triggered an explosion in reporting. Increasing reporting initially helped but eventually hindered money laundering prosecution. In particular, the empirical evidence shows that the United States’ anti-money laundering regime is potentially facing a crying wolf problem, where the increasing number of reports is detrimental to the efficiency of the anti-money laundering regime.

6. Discussion

6.1 Implementable Policy Implications

This article identifies the crying wolf problem as potentially relevant to the money laundering enforcement regime in the United States. In addition, the model is also able to identify implementable policy solutions. However, as the empirical evidence is scattered, the policy implications are suggestions for future policy debates rather than strict prescriptions.

The model identifies that the crying wolf problem might be remedied by reducing fines. If the first best is implementable, then, as Proposition 9 shows, decreasing fines will implement it. There are some signs that the problem of excessive fines is understood. Fines could be decreased, for instance, by centralizing the prosecution of banks for Banking Secrecy Act violations as it happened with the modification of the US Attorneys’ Manual (2005). According to industry opinions, the changes were motivated by the case of AmSouth Bank which was criminally prosecuted by a Mississippi state attorney. The prosecution disturbed banks because they felt threatened to be fined by too many organizations, such as Federal Reserve Banks, OCC, FinCEN, and state attorneys. Furthermore, an internal OCC (2005) memorandum, denying zero tolerance policies, became available in the public domain, which could also be interpreted as a signal for lowering fines.11

The model also highlights the potential role of reporting fees. As Lemma 3 has shown, the first best is always implementable if reporting is sufficiently costly. Furthermore, an extension of the model with continuous signals and monitoring efforts, discussed in Appendix A, demonstrates that reporting fees are generally necessary to implement the first best. Moreover, changing the

---

11. An important caveat applies to decreasing fines. Fines might be necessary to stop banks from competing for criminal money. Hence, fines should not be decreased so much to elicit such a competition. However, industry experts maintain that current fines—and the cost of nonmonetary penalty—are more than sufficient for this goal.
cost of reporting is easily implementable as regulatory authorities can charge a reporting fee.

The economic argument for reporting fees could also be understood through the externalities caused by reporting. Each report dilutes the value of all the other reports. However, banks do not fully internalize these reporting externalities, that is, the shadow costs of information dilution, and reporting fees could be used to price these dilution costs.12

Alternatively, banks could be given incentives by “privatizing” money laundering prosecution, that is, rewarding banks for useful reports and charging them for wasted law enforcement efforts on the others. There are three main problems with this proposal. First, banks might not want to appear to be reporting their clients for rewards, a kind of “blood money.” Second, such a system would need a separate pricing mechanism for each bank, which might make it unfeasible. Third, banks might not find acceptable to depend on the results of uncontrollable investigation efforts.

The comparative statics also provide lessons for setting policy in the future. Three examples are discussed here. First, increased harm caused by money laundering should lead to decreased fines. Second, increased likelihood of money laundering should also lead to decreased fines. Surprisingly, the graver the problem, that is, the more likely or the more harmful money laundering is, the weaker the incentives should be. Third, changes in information technology are likely to decrease reporting costs which can lead to crying wolf. In order to avoid this, policy makers might want to introduce or further increase reporting fees.

6.2 Applications in Product Information and Auditing

6.2.1 Product Information. Product information on potential hazards is important in identifying for customers the relevant risks. Crying wolf is harmful because it leaves consumers less well informed about relevant risks.

In this setting, all the building blocks of crying wolf are present. Communication is coarse because the manufacturer has much more information than could be conveyed in a simple manual. Manufacturer’ incentives are coarse because injured consumers might litigate for damages if they have not been properly informed. Thus, manufacturers are punished for false negatives only. Finally, uncertainty is naturally present as no product is completely fool-proof.

In this setting, crying wolf can arise if damages awarded (or fines imposed) by courts are excessively high. Crying wolf can take the form of excessive small prints in product manuals or contracts. McDonald’s warning that the hot chocolate is indeed hot is an example of crying wolf. Under crying wolf, customers cannot distinguish between relevant warnings and those that are

---
12. Reporting fees might be seen as putting undue burden on the banking industry. This is, however, not necessarily so. Reporting fees are not needed for their redistributive effect (which might be the major concern of the industry) but rather for their incentive effect. For instance, fees might be distributed back to the banking industry according to some statistics (such as assets, deposits, or transfers) eliminating the redistributive effect and still maintaining the incentives.
listed “defensively.” Thus, in order to save on costly precautionary actions, they might disregard all the information provided. For instance, online customers commonly sign or click that they have read and understood the terms of service without even looking at them.

6.2.2 Sarbanes-Oxley Act: Reporting Material Transactions. The model can be used to further think about analyzing reporting material transactions as prescribed by the Sarbanes-Oxley Act. Auditors reporting material transactions face a similar problem as banks disclosing suspicious activities. First, communication is coarse: auditors identify a transaction as material but cannot communicate all the premises of their judgment. Second, incentives are coarse: auditors face fines should they fail to identify and report material transactions. Third, auditors evaluating the importance of the transactions are uncertain about their true importance. Finally, the validity of auditors’ professional judgment on the importance of the transaction cannot be verified ex post. These similarities are sufficient to give rise to the crying wolf problem if fines are excessively high.

This new angle on material transactions highlights a potential danger of the Sarbanes-Oxley Act. The act has increased auditing fines on auditors much like the USA Patriot Act has increased fines on banks. Clearly, the threat of fines was insufficient before the accounting scandals. The question is how high auditing fines are now: “Are fines already excessively high resulting in crying wolf or are they in the optimal range?” The question deserves further investigation. Especially, since some preliminary findings hint that the crying wolf problem is potentially relevant. For instance, Gordon (2006) reports that the average length of Fortune 500 firms’ 10-K reports has grown from 16 pages in 1950 to 126 pages in 2000, even before the Sarbanes-Oxley Act.

There are, of course, some caveats before directly applying the model based on money laundering enforcement to auditing situations. First, as auditing reports are more detailed and less numerous than SARs, communication might be less coarse in auditing. Second, though auditing firms have specific information about their clients, it is rarely as specific as banks’ information on their clients. Regulators can always ask a second auditor to double check the findings of the first one. Third, individual clients are arguably more important for auditing firms than for banks. In addition, auditing reports are known to the client, in contrast to SARs. Thus, auditors might have much stronger disincentives to report than banks.

7. Conclusion

The article identifies the crying wolf problem in money laundering enforcement. This is important for three main reasons. First, the article is the first to formally analyze money laundering enforcement. The research helps to understand the specificities of how this harmful financial crime is fought.

Second, the article provides implementable policy implications, many of which seem counterintuitive at first sight. The article calls for decreased fines against banks found to be negligent in carrying out their anti-money laundering
duties. Furthermore, the model shows that fines against banks should decrease as the harm caused by money laundering increases. The article also makes the case for introducing reporting fees, that is, charging banks for informing law enforcement agencies. The article also highlights how well intentioned and seemingly sensible money laundering regulations, such as the USA Patriot Act, could have backfired.

Third, the article explores the general problem of crying wolf. For instance, the model can be applied to analyze product information provision. Furthermore, the crying wolf model can be used to analyze how auditors disclose material transactions prescribed by the Sarbanes-Oxley Act. Most importantly, the article questions the conventional wisdom in corporate finance that more disclosure implies better information.

The model could also act as a basis for further analysis of potential problems of crying wolf in governance settings. The model can be extended to encompass dynamic reporting considerations, when reporting agents care about their career and reputation as in Prendergast and Stole (1996). For instance, such a dynamic model could be used to analyze how optimal catastrophe and terrorism warning systems should work. Furthermore, reporting and potential crying wolf in hierarchies could be examined in hierarchies following Prendergast (1993).

Crying wolf might be particularly relevant in intelligence settings. Intelligence reports to decision makers are potentially susceptible to crying wolf. Communication is coarse because all the special information cannot be transmitted. Uncertainty about the validity of the information is naturally present. Furthermore, the specific information of the intelligence agencies cannot usually be verified ex post. Thus, following Garicano and Posner (2005), the efficiency of different organizations and incentives in curbing crying wolf and the relevant trade-offs could be investigated.

Finally, the problem of crying wolf sheds some light on the fundamental question of what qualifies as information. The article focused on the identification role of reporting and showed formally how excessive reporting fails to identify and thus inform. This identification role highlights that in certain situations filtering, in other words withholding specific data, is necessary to efficiently inform. Such filtering of data is becoming increasingly important in the economy as data are increasingly easier to store and forward with the development of information technology. This suggests that crying wolf problems might surface in many other applications in the foreseeable future.

Appendix A

Money Laundering

Money laundering is defined as an illicit money transfer. There are two main kinds of illicit money transfers. First, traditional money laundering entails transferring illegally obtained funds to conceal their origins and make them appear legal. For example, drug dealers deposit cash revenues in banks and later transfer them until the funds appear to originate from legitimate sources. Second, terrorism financing entails transferring mostly legal funds for illegal
purposes. For instance, legal charity donations are transferred to fund terrorist attacks. In sum, both forms of money laundering are characterized by illicit and socially harmful fund transfers. Money laundering causes social harm because it facilitates crime and enables criminals to enjoy criminal revenues.

Money laundering can happen through various intermediaries. Bank transfers, both by wire and by check, are the most common channels for illicit money transfers as described in Reuter and Truman (2004). Money transmitting businesses, such as Western Union, are also used for money laundering as detailed in The Simpson (2004b). These businesses are typically franchised or owned by individuals, who might have stronger incentives to turn a blind eye to money laundering than bank branch managers. In the greyer area of finance, informal value transfer systems provide money transmitting services usually without a proper paper trail. The *hawala* or *hindi* systems used by different ethnic communities are described, for instance, in El-Qorchi (2002).

Money laundering is an economically significant crime, though precise estimates are hard to obtain. According to Camdessus (1998), the consensus range of money laundering volume is between 2% and 5% of the global GDP. The FBI (Schroeder 2001) estimates the volume of globally laundered funds as falling between $600 billion and $1.5 trillion. Laundering fees, that is, what money launderers charge their criminal clients, are estimated at 51%–5% of the laundered amount according to Lal (2003) and Reuter and Truman (2004). Thus, money laundering, including self-laundering, is estimated to be a $30 to $225 billion global “industry.” Moreover, the harm caused for instance by terrorism financing shows that money laundering is potentially even more significant than what volumes and laundering revenues would suggest.

Money laundering enforcement is particularly relevant for the United States. According to some lawmakers’ estimates, half of the globally laundered funds are transferred through US Banks (Schroeder 2001). Three known money laundering cases highlight the point. First, $7 billion of Russian funds were washed through the Bank of New York until 1999 (Reuter and Truman 2004). Second, Stephen Saccoccia alone laundered up to $550 million of drug money for both the Cali and the Medellin cartels until he was prosecuted in 1993 (Reuter and Truman 2004). Third, terrorists transferred $0.5 million for the 9/11 attack (National Commission on Terrorist Attacks 2004). The examples also show that the volumes involved in laundering proceedings of tax evasion and drug trafficking are enormous compared to terrorism financing.

---

13. Responding to the threat of money laundering, the United States has developed one of the strongest anti-money laundering regulation. The Banking Secrecy Act (1970), which in fact curbed banking secrecy to fight money laundering, was followed by a series of laws, each one of them further strengthening money laundering enforcement: The Money Laundering Control Act (1986), the Annunzio-Wylie Money Laundering Act (1992), the Money Laundering Suppression Act (1994), The Money Laundering and Financial Crimes Strategy Act (1998), and finally the USA Patriot Act (2001).
Equivalent Problem

The bank’s action set can be written in a more general form. Action $R_0$ denotes the bank’s filing decision if the observed signal is low (0). If the bank reports after observing the low signal, then $R_0 = 1$, or else $R_0 = 0$. Similarly, after observing the high signal, the bank either reports ($R_1 = 1$) or does not report ($R_1 = 0$). The bank incurs cost $c > 0$ by reporting. Thus, the action set of the bank has three elements: $(M, R_0, R_1) \in \{(0, 1)\}^3$.

However, this more general setup can be rewritten in the form presented before as $(M, T)$, where $T \in \{0, 1\}$ is a reporting threshold, instead of $(M, R_0, R_1)$. The bank reports all signals weakly higher than $T$ and does not report signals below it. Most importantly, this rewriting of the actions set excludes the $(R_0 = 1, R_1 = 0)$ action pair, while allowing for all other combinations. Thus, it has to be shown that the $(R_0 = 1, R_1 = 0)$ action pair can be excluded without loss of generality both from the first best and from the second best equilibria.

In the first best equilibrium, social welfare is maximized while the bank’s incentive considerations are abstracted away. Consequently, in the maximization and government investigation setting only the signal’s information content matters. Reporting only the low signal and not the high signal $(R_0 = 1, R_1 = 0)$ provides the same information as reporting only the high signal and not the low signal $(R_0 = 0, R_1 = 1)$. However, as money laundering is rare and the signals are sufficiently precise, the low signal is more likely. Thus, reporting the low signal and not reporting the high signal is more expensive, and therefore this reverse reporting cannot be part of a first best equilibrium. The reasoning is formalized in Lemma A1.

**Lemma A1.** The $(R_0 = 1, R_1 = 0)$ action pair does not arise in the first best equilibrium.

**Proof.** Follows from the discussion above. To see that the low signal is more likely,

$$\Pr(\sigma = 1) = 1 - \alpha - \delta + 2\alpha\delta < \alpha + \delta - 2\alpha\delta = \Pr(\sigma = 0).$$

It is equivalent with

$$1/2 < (1 - \alpha)\delta + \alpha(1 - \delta),$$

which directly follows from $0 < \alpha < 1/2 < \delta < 1$. \qed

In second best equilibria, the social welfare maximization is constrained by incentive considerations. The bank is motivated to report because, absent reporting, it might get fined. Thus, whenever the expected fine is higher than

---

14. According to FinCEN, completing a SAR filing takes between 1/2 and 3 h for a compliance officer.

15. Notice that constraining $T \in \{0, 1\}$ is without the loss of generality. Allowing for $T \in [0, 1]$ would produce exactly the same equilibrium reporting thresholds, that is, either zero or one.
the reporting cost, the bank reports. However, the expected fine depends on the
likelihood of money laundering. Thus, if the bank reports given the low signal
(when the probability of laundering is low), then it should also report under
the high signal (when the probability of laundering is high). Lemma A2 shows
formally that the \((R_0 = 1, R_1 = 0)\) pair can be excluded from the second best
equilibrium without loss of generality.

**Lemma A2.** In second best equilibria, the \((R_0 = 1, R_1 = 0)\) action pair does
not arise if in equilibrium \(M = 1\). Furthermore, if in equilibrium \(M = 0\), then
the \((R_0 = 1, R_1 = 0)\) action pair can be replaced with any other action pair
without loss of generality.

Lemmas A1 and A2 allow for rewriting the problem as formalized in
Corollary A3.

**Corollary A3.** The action set of the bank can be rewritten as \((M, T)\), where
\(T \in \{0, 1\}\) is a reporting threshold. The bank reports if \(\sigma \geq T\) and does not
report if \(\sigma < T\).

**Proof.** Follows from Lemmas A1 and A2 and noticing that except for
\((R_0 = 1, R_1 = 0)\) the new action set can replicate all actions of the original
action set. □

**Exogenous Fine Equilibria with \(F^* > F^{**}\)**

If the first best is not implementable, then the government prefers to implement
one of the Second Best equilibria. Corollary 6 shows that this is possible if the
fine is sufficiently low. (Note that Corollary 6 does not depend on whether the
first best is implementable or not.) If the second best is not implementable be-
cause the fine is too high, then Third Best equilibria are implemented. Lemma
A4 formalizes the argument.

**Lemma A4.** Sufficiently high fines that satisfy both \(F \geq F''\) and \(F > F'\)
implement the Third Best equilibria if the First Best equilibrium is not imple-
mentable. Then government investigation is set \(I_0 \geq I'_0(F)\). Furthermore, \(F'\) is
as defined in Lemma 8 and

\[
F'' \equiv \max \left\{ \frac{(\alpha + \delta - 2\alpha\delta)c}{\alpha\delta}, \frac{c + m}{\alpha} \right\}
\]

and \(I'_0(F) < 1\) for \(\forall F \geq F''\).

**Proof.** As First Best equilibrium is not implementable, the government
prefers to implement the second best, which is implementable by Corollary
6 with \(F \leq F'\). Otherwise, the government can only implement Third Best
equilibria. By IC1 and IC2 formalized in the proof of Lemma 8, Third Best
equilibria are implementable if
F(I_0) \geq \max \left\{ \frac{(\alpha + \delta - 2\alpha\delta)c}{\alpha\delta I_0}, \frac{c + m}{\alpha I_0} \right\}.

Thus, the Third Best is implementable with fines larger than \( F'' \), where

\[
F'' \equiv \max \left\{ \frac{(\alpha + \delta - 2\alpha\delta)c}{\alpha\delta}, \frac{c + m}{\alpha} \right\}.
\]

However, depending on the parameter values, this \( F'' \) can be both higher and lower than \( F' \), as straightforward examples can show. If \( F'' > F' \), then there is no pure strategy equilibrium with \( F' < F < F'' \), and Third Best equilibria prevail with \( F \geq F'' \). If \( F'' \leq F' \), then the Third Best is implemented with \( F > F' \).

Consequently, the resulting equilibria can be characterized.

**Proposition A5.** Low fines, \( F \leq F' \), implement Second Best equilibria and high fines, \( F'' < F \), implement the Third Best equilibria with \( I_0 \geq I'_0(F) \) if the First Best is not implementable.

**Proof.** Follows from Corollary 6 and Lemma A4.

The solution shows that increasing fines decrease social welfare. In this case, however, prosecution rates do not change because welfare decrease only through wasteful monitoring and reporting. Corollary A6 formalizes the result.

**Corollary A6.** The expected prosecution rate is constant in fines: \( \chi^{**} \), if the First Best equilibrium is not implementable.

**Proof.** Follows directly from Lemma 10 and Proposition A5 of the appendix.

Thus, even if the first best cannot take place, crying wolf arises as defined earlier. Excessively high fines still remain detrimental because they increase the social costs of operating the money laundering enforcement regime without improving the prosecution of money laundering. Nonetheless, in this case crying wolf does not decrease expected prosecution.

**Further Discussions**

**Motivation for exogenous fines.** Exogenous fine setting can be thought resulting from divided decision making. As in many law enforcement situations, punishments are not set by the law enforcement agencies in charge. For instance, legislators can set fines or expectations on fine levels—and there is no guarantee that the fine level set by them is optimal. First, they might have limited information on law enforcement issues. In particular, they might believe that stronger fines solve the problem. Second, they might have different objectives. For instance, some legislators would like to signal to voters being tough on crime. In any case, fine levels might not be set optimally. This explains why
fines might be set exogenously and how law enforcement agencies find the best action conditional on the fine level.

**Defensive Medicine and Information Overload.** Crying wolf may seem superficially similar to both defensive medicine and information overload, but it is in fact fundamentally different. First, crying wolf bears some likeness to the well-understood defensive medicine problem. Defensive medicine is also triggered by excessive fines, that is, threat of lawsuits. Furthermore, defensive medicine not only does reduce social welfare but also might hinder general medical goals. For instance, unnecessary X-rays might harm patients’ health. Yet, the crucial difference is that in crying wolf social welfare is decreased through information dilution. Excessive reports destroy the information available for law enforcement. However, defensive medical practices do not dilute information.

Second, crying wolf is different from information overload as described for instance in Garicano and Posner (2005). Information overload arises as a result of inefficient information processing. For example, in information overload intelligence data cannot be processed on time to start counter-terrorist measures. In crying wolf, the problem is not with processing information but rather with the coarse way of communicating it. Law enforcement immediately receives SARs. Instead, the question is rather how much to trust the bank’s judgment.

**SARs: Data Provision versus Identification.** The duality of information, discussed in general in the conclusion, is also present in SARs. SARs both identify suspects and provide otherwise unavailable raw data to law enforcement. The raw SAR database can be useful in locating wanted criminals, establishing links between different suspects, and for other searches. The distinction between data provision and identification explains why agencies interested in fighting money laundering (such as FinCEN) try to curb excessive reporting, whereas agencies with broader law enforcement goals (such as the FBI) prefer even more reports. Thus, database building is consistent with the FBI position that SAR filings are not harmful (Heller 2005). However, the decreasing number of prosecutions supports FinCEN’s point of view: the increasing number of reports is not useful for fighting money laundering specifically.

**Formal Extensions**

**Deterrence.** Though abstracting from deterrence is justified as it is argued below, it is interesting to see how deterrence could be incorporated into the basic model. The most notable effect of deterrence is that the First Best equilibrium is not necessarily characterized by higher expected prosecution rates. Under deterrence, first best expected prosecutions are affected by two conflicting forces. On the one hand, better information increases the likelihood of prosecuting money laundering. On the other hand, more efficient prosecution deters potential money laundering. Theoretically, either effect could dominate.

Nonetheless, the deterrence effect is likely to be weaker in practice because money laundering is inelastic, as discussed in Reuter and Truman (2004).
The economic reason for inelastic money laundering stems from the fact that money laundering is linked to the predicate crime. For instance, it is hard to imagine drug dealers stopping their illicit trade in response to money laundering prosecution alone.

Criminal deterrence can also be investigated by a simple extension of the model. Criminals anticipate government and bank equilibrium actions. Thus, when the money laundering enforcement regime is efficient, the likelihood of prosecuting money laundering is high, and fewer criminals will launder money. Let us suppose that criminals have a utility from successful (i.e., not prosecuted) money laundering ($U_M > 0$) and utility loss from prosecution ($U_P > 0$). A risk neutral criminal will launder money if the expected gains from laundering exceed the expected penalty. Formally,

$$(1 - \psi)U_M \geq \psi U_P,$$

where $\psi \in (0, 1)$ is the probability that money laundering is prosecuted. This probability can be expressed formally as

$$\psi = (1 - p_T)I_0^* + p_T I_1^*.$$  

The prosecution probability clearly depends on equilibrium monitoring, reporting, and investigation decisions. If the gain of money laundering is a random variable, then more efficient prosecution implies that only money launderers with high laundering utility ($U_M$) will engage in money laundering.

Deterrence can be investigated most easily by applying a simple linear reduced form of the laundering decision:

$$\alpha(\psi) = \alpha^* - b\psi,$$

where $\alpha^* > 0$, $b > 0$, and for the highest equilibrium $\psi$, $\alpha(\psi_{\text{max}}) > 0$. In equilibrium, a fixed point needs to be found also in terms of $\alpha(\psi)$.

Most importantly, deterrence does not change the model’s main finding that excessive fines result in harmful crying wolf.

Continuous Signal and Monitoring Model. The model can also be extended to include more general signals, bank monitoring, and reporting decisions. The most important result is that the continuous model does not change any of the model’s qualitative conclusions. Moreover, the continuous model highlights that adjusting reporting costs (i.e., using reporting fees) is generally necessary to implement the first best. Intuitively, fines will affect both the threshold ($T$) and the monitoring effort ($M$). Thus, only in borderline cases will it be possible to fine-tune the system solely by adjusting fines. Consequently, the government needs to charge reporting fees to increase the reporting threshold ($T$) to the optimal level.16

---

16. Theoretically, it is possible that the government needs to use reporting subsidies to decrease the reporting threshold. However, given the “defensive filing” problem, subsidies do not seem to be the policy relevant instruments.
In the continuous model, three assumptions are changed. First, it is assumed that the signal \((\sigma)\) is continuous. Moreover, it is assumed that it is distributed such that the posterior probability of money laundering \(\beta(\sigma)\) is uniform on \([0,1]\), which implies that, under rational updating, \(\alpha = 0.5\). Second, it is assumed that bank monitoring \((M)\) is continuous on \([0,1]\) with costs \(mM^2\), where \(m > 0\). Thus, the bank is not always informed or uninformed. Third, it is assumed that the reporting threshold \((T)\) is freely chosen by the bank on the \([0,1]\) interval.

The above continuous specification allows for rewriting the probabilities:

\[
\begin{align*}
p_T &= 1 - T, \\
q_{0T} &= \frac{T}{2}, \\
q_{1T} &= \frac{1 + T}{2}.
\end{align*}
\]

The first best welfare maximization can be written as

\[
W = M \left[ T \left( \frac{T^2}{2} I_0 \rho h - k T I_0^2 \right) + (1 - T) \left( \frac{1 + T}{2} I_1 \rho h - k T I_1^2 + c \right) \right] \\
+ (1 - M) \left[ \alpha I_0 \rho h - k I_0^2 \right] - mM^2 - \alpha h,
\]

which is a linear quadratic problem, whose interior solution\(^{17}\) is characterized by the first-order conditions

\[
\begin{align*}
(M) & \quad M = \frac{1}{2m} \left[ \frac{T^2}{2} I_0 \rho h - T k I_0^2 + \frac{1 - T^2}{2} I_1 \rho h - (1 - T)(k I_1^2 + c) \\
& \quad \quad \quad - \alpha I_0 \rho h + k I_0^2 \right], \\
(T) & \quad T = \frac{k(I_1 + I_0)}{\rho h}, \\
(I_0) & \quad I_0 = \frac{\rho h M T^2 + 2(1 - M)\alpha}{MT + 1 - M}, \quad (A1) \\
(I_1) & \quad I_1 = \frac{\rho h}{4k} (1 + T). \quad (A2)
\end{align*}
\]

Unfortunately, closed form solution is unattainable.

In the second best problem, it is investigated how the first best solution \((M^*, T^*, I_0^*, I_1^*)\) can be implemented. The bank’s profit maximization problem is given by

\[
\Pi = -M \left( \frac{T^2}{2} I_0 F + (1 - T)c \right) - (1 - M)\alpha I_0 F - mM^2,
\]

\(^{17}\) In this analysis, only interior solutions are explored and corner solutions are left to the reader.
which is, again, a linear quadratic problem. The interior solution is characterized by the first-order conditions

\[ M = \frac{1}{2m} \left[ \alpha I_0 F - \frac{T^2}{2} I_0 F - (1 - T)c \right], \quad (A3) \]

\[ T = \frac{c}{I_0 F}, \quad (A4) \]

where the first condition (A3) can be restated by substituting \( T \) in as

\[ M = \frac{1}{2m} \left[ \alpha I_0^* F + \frac{c^2}{2I_0^* F} - c \right]. \]

The first-order conditions allow an important observation. In general, fines cannot set both monitoring \((M)\) and reporting \((T)\) to the first best level. Thus, in the continuous model, it is generally necessary to use reporting fees (or subsidies) in addition to optimal fines.

Finally, the equilibria under exogenous fines are investigated. Under exogenously set fines, bank and government action can be jointly determined using the first-order conditions. There is no closed form solution to the problem, so the solution is simulated using csolve.m in matlab.

Figure 8 illustrates a solution with parameters \( \rho = 1, h = 9, k = 15, \alpha = 0.5, m = 1, \) and \( c = 0.1. \) Part (A) displays the information Laffer curve. Increasing fines first increase but later decrease the prosecution rate. The reason for the eventual decline in prosecution is that the government reacts less to reporting. Part (B) shows that how the expected number of reports increases in fines. The government reacts less to reporting because reports are becoming less informative as fines grow. Part (C) depicts that even though the bank monitors more as fines increase, it also reports lower signals. Part (D) shows clearly that as the fine grows, the bank lowers the reporting threshold. With very high fines, the bank reports almost all transactions thereby rendering its signal uninformative.

Parts (C) and (D) of Figure 8 also illustrate the dual and conflicting role of fines. On the one hand, higher fines increase monitoring as shown in Part (C). Increased monitoring improves the efficiency of the anti-money laundering regime. This monitoring effect is responsible for the initial positive response to increased fines. On the other hand, increased fines decrease the reporting threshold as shown in Part (D). This threshold effect might be positive if the reporting threshold is initially higher than the first best level. Even with intermediate fines, the monitoring and threshold effects might balance each other out. However, the negative consequences of the threshold effect eventually dominate. The lower threshold eventually destroys the information value of reports. This is how crying wolf renders reports uninformative.

Wasteful Fines. The model can be extended to allow for distorting and socially costly fines. In fact, some compliance costs might well be wasteful. Wasteful fines can be modeled by assuming that the government receives only \( \lambda \in (0, 1) \) fraction of the total fines.
Naturally, condition (1) does not change. The first best solution requires the same conditions. Nevertheless, in order to implement the first best solution in the second best problem, an additional condition must be satisfied:

\[
\frac{[\alpha (1 - \delta) \rho h]^2}{4k(\alpha + \delta - 2\alpha \delta)} + \frac{(\alpha \delta \rho h)^2}{4k(1 - \alpha - \delta + 2\alpha \delta)} > \frac{(\alpha \rho h)^2}{4k} + (\alpha + \delta - 2\alpha \delta)c + m + \lambda F^*. \tag{A5}
\]

The interpretation of the above condition is straightforward: with wasteful fines the social gains from prosecution must exceed not only monitoring and reporting costs but also the social waste of the lowest possible fine implementing the first best.

The model’s qualitative results on optimal fines, reporting fees, or comparative statics do not change under wasteful fines. However, there are two minor differences. First, the first best welfare cannot be implemented because fines always waste some utility. Second, under wasteful fines the government strictly prefers to impose the smallest possible fine.

Effort to Increase Signal Precision. A reasonable extension is to investigate the model if the bank can exert effort to increase the precision of the signal.
A Theory of “Crying Wolf”

The model’s main predictions do not change in this scenario: excessive fines lead to crying wolf, and reporting fees are needed. However, increased harm caused by money laundering might prescribe larger and not smaller fines if reporting fees are used as well. The intuition is that if money laundering is more harmful, more precise signals are needed. More precision is possible only if punishment for both false negatives (fines) and false positives (reporting fees) is increased. However, if reporting fees are not used, then increased harm still should be accompanied with lower fines so as not to trigger crying wolf.

Proofs

Proof of Lemma A2. Suppose that \((R_0 = 1, R_1 = 0)\). Let us assume further, without loss of generality, that \(I'_0\) and \(F'\) are the government’s best responses.

First, consider that the bank monitors, \(M = 1\), in equilibrium. Then, the expected value of fines conditional on nonreporting is \(I'_0 F' \beta_0\) with the low and \(I'_0 F' \beta_1\) with the high signal. In the second best, the bank reports only if the expected fines conditional on nonreporting are at least weakly higher than the reporting cost \(c\). Thus, reporting implies that the expected fine conditional on nonreporting is higher than the reporting cost:

\[ R_0 = 1 \implies I'_0 F' \beta_0 \geq c, \]
\[ R_1 = 0 \implies I'_0 F' \beta_1 \leq c, \]

which is impossible if both \(I'_0, F' > 0\) as \(\beta_0 < \beta_1\). Furthermore, if the bank monitors \((M = 1)\), then the government does not set \(I'_0 = 0\) or \(F' = 0\) in any second best equilibria. The reason is that if either \(I'_0 = 0\) or \(F' = 0\), then the bank is never fined. This implies, however, that the bank does not monitor (and sets \(M = 0\)) because monitoring reduces private profits by \(m > 0\), but it does not save on fines.

Second, consider equilibrium bank monitoring \(M = 0\). In this case, the bank is never informed and it never reports. Thus, equilibrium \((R_0 = 1, R_1 = 0)\) can be replaced with any other reporting actions. The government’s best response, social welfare, and observed bank actions do not change. □

Proof of Proposition 1. Notice, first, that in the first best fines do not play any role. Thus, as the government prefers to make the bank as well-off as possible and it sets the fine to zero, \(F = 0\).

In order to explore the possible equilibria, start with the bank action set. There are four possible bank action pairs:

\((M = 0, T = 0), \ (M = 0, T = 1), \ (M = 1, T = 0), \ (M = 1, T = 1).\)

First, if \(M = 0\), then the bank never reports. Setting \(T\) is therefore inconsequential, and it does not affect welfare or best response. Thus, bank action pairs \((M = 0, T \in \{0, 1\})\) can be treated together. The social welfare function then is written as

\[ W(I_0, I_1, F = 0, M = 0, T \in \{0, 1\}) = \alpha I_0 \rho h - k I_0^2 - \alpha h, \]  

(A6)
which yields the following government investigation best responses:

\[ I_0 = \frac{\alpha \rho h}{2k} \equiv I^{**}, \quad I_1 \in [0,1]. \]

Substituting back to the welfare function yields

\[ W^{**} = I^{**} \rho h - k(I^{**})^2 - \alpha h \]
\[ = \frac{(\alpha \rho h)^2}{4k} - \alpha h. \]

Second, consider the \((M = 1, T = 0)\) action pair. Then the social welfare function can be written as

\[ W = q_1 T I_1 \rho h - k I_1^2 - c - m - \alpha h, \]

which is maximized by setting

\[ I_1 = \frac{q_1 T \rho h}{2k} = \frac{\alpha \rho h}{2k} \equiv I^{**} \]

and is not affected by \(I_0\), as the bank always reports. Thus, the social welfare

\[ W^{***} = I^{**} \rho h - k(I^{**})^2 - c - m - \alpha h \]
\[ = W^{**} - c - m. \]

Thus, the \((M = 1, T = 0)\) action pair cannot be part of the First Best equilibrium.

Under bank action pair \((M = 1, T = 1)\), the social welfare function can be written as

\[ W(I_0, I_1, F = 0, M = 1, T = 1) \]
\[ = (\alpha + \delta - 2\alpha \delta)(\beta_0 I_0 \rho h - k l_0^2) \]
\[ + (1 - \alpha - \delta + 2\alpha \delta)(\beta_1 I_1 \rho h - k I_1^2 - c) - \alpha h, \]

which yields the following investigation best responses:

\[ I_0^* = \frac{\alpha (1 - \delta) \rho h}{2k(\alpha + \delta - 2\alpha \delta)}, \]
\[ I_1^* = \frac{\alpha \delta \rho h}{2k(1 - \alpha - \delta + 2\alpha \delta)}. \]

Substituting back to the welfare function yields

\[ W^* = \frac{[\alpha (1 - \delta) \rho h]^2}{4k(\alpha + \delta - 2\alpha \delta)} + \frac{(\alpha \delta \rho h)^2}{4k(1 - \alpha - \delta + 2\alpha \delta)} - (\alpha + \delta - 2\alpha \delta)c - m - \alpha h. \]

Finally, the equilibrium with positive bank investigation provides higher welfare as

\[ W^* - W^{**} = \frac{[\alpha (1 - \delta) \rho h]^2}{4k(\alpha + \delta - 2\alpha \delta)} + \frac{(\alpha \delta \rho h)^2}{4k(1 - \alpha - \delta + 2\alpha \delta)} \]
\[ - \frac{(\alpha \rho h)^2}{4k} - (\alpha + \delta - 2\alpha \delta)c - m > 0. \]
The expression is positive by assumption set in equation (1).

Finally, note that all possible government best response investigations are on the unit interval. First, all \((I^*, I_0^*, I_1^*)\) are nonnegative. Second, by \(\delta > (1-\delta)\),

\[
I^* = \frac{\alpha \rho h}{2k} < \frac{\alpha \delta \rho h}{2k(1-\alpha - \delta + 2\alpha \delta)} = I_1^*.
\]

Third, again by \(\delta > (1-\delta)\),

\[
I_0^* = \frac{\alpha (1-\delta) \rho h}{2k(\alpha + \delta - 2\alpha \delta)} < \frac{\alpha \rho h}{2k} = I^*.
\]

Finally, by the assumption stated in equation (2), \(I_1^* < 1\). Thus, all three investigation probabilities are on the unit interval. Furthermore, they can be ranked as \(I_0^* < I^* < I_1^*\). □

Proof of Proposition 5. There are three possible best response pairs by the proof of Proposition 1. In terms of actions and social welfare,

First Best \((M = 1, T = 1)\) \((I_0 = I_0^*, I_1 = I_1^*)\) \(W^*\)

Second Best \((M = 0, T \in \{0, 1\})\) \((I_0 = I^*, I_1 \in [0, 1])\) \(W^{**}\)

best response pair 3 \((M = 1, T = 0)\) \((I_0 \in [0, 1], I_1 = I^{**})\) \(W^{***}\)

The welfare implied by the three best response pairs is derived in Proposition 1:

\[W^{***} < W^{**} < W^*\]. \hspace{1cm} (A8)

Thus, the government would like to first implement (if it is possible) the First Best equilibrium. By Lemma 3, the first best is implementable if

\[m \leq \frac{(1-\alpha)(2\delta - 1)}{(1-\delta)c}.
\]

As the government prefers to increase banking profits, as long as social welfare is not affected, it sets fines as low as possible. The lowest possible fine that still implements the First Best is \(F^*\), as demonstrated by Lemma 2. As the government can commit to investigation action, the First Best equilibrium is uniquely implemented with \(F = F^*\) fines.

If the First Best is not implementable, the government prefers to implement one of the Second Best equilibria, all of which require the very same incentive compatibility constraints:

\[
\text{IC}_1 \quad \Pi(M = 1, T = 1, I_0 = I^{**}, I_1 \in [0, 1]) \\
\leq \Pi(M = 0, T \in \{0, 1\}, I_0 = I^{**}, I_1 \in [0, 1]),
\]

\[
\text{IC}_2 \quad \Pi(M = 1, T = 0, I_0 = I^{**}, I_1 \in [0, 1]) \\
\leq \Pi(M = 0, T \in \{0, 1\}, I_0 = I^{**}, I_1 \in [0, 1]).
\]
Starting with IC$_1$:

$$-(1 - p_1)q_{01}I^{**}F - p_1c - m \leq -\alpha I^{**}F,$$

$$F \leq \frac{p_1c + m}{(\alpha - (1 - p_1)q_{01})I^{**}} = 2k(1 - \alpha - \delta + 2\alpha \delta)c + m \over \alpha \delta \rho h.$$  

Then IC$_2$:

$$-c - m \leq -\alpha I^{**}F,$$

$$F \leq \frac{c + m}{\alpha I^{**}} = 2k(\alpha + \delta - 2\alpha \delta)(c + m) \over \alpha^2 \rho h.$$  

Thus, the Second Best equilibria are implementable with fines such that

$$F \leq \min \left\{ 2k(1 - \alpha - \delta + 2\alpha \delta)c + m, 2k(\alpha + \delta - 2\alpha \delta)(c + m) \right\} \over \alpha \delta \rho h.$$

The government again uses the smallest fine necessary, which in this case is zero, $F = 0$. Government commitment along with the satisfied IC constraints guarantee that one of the Second Best equilibria will prevail in equilibrium. Finally, the government never implements best response pair 3. It is welfare dominated by Second Best equilibria, which are always implementable. □

Proof of Corollary 6. Follows from the proof of Proposition 5 and observing that $F' < F^*$ as

$$2k(1 - \alpha - \delta + 2\alpha \delta)c + m \over \alpha \delta \rho h < 2k(\alpha + \delta - 2\alpha \delta)(1 - \alpha - \delta + 2\alpha \delta)c + m \over \alpha^2 \delta (1 - \delta) \rho h,$$

$$\alpha(1 - \delta) < (\alpha + \delta - 2\alpha \delta) = \alpha(1 - \delta) + \delta(1 - \alpha),$$

which implies that with $F \leq F'$ the government cannot implement the First Best equilibrium; thus, it prefers to implement Second Best equilibria. □

Proof of Lemma 8. The proof is lengthy and divided into four parts. First, by Lemma 2, if fines are higher than $F^{**}$, then the First Best equilibrium is not implementable.

Second, notice that if inequality (8) holds and fines are higher than $F^{**}$, then the Second Best equilibria cannot prevail either. The necessary fine level for the Second Best equilibria is $F'$ derived in equation (9) which is lower than $F^*$ by the proof of Corollary 6. Furthermore, inequality (8) implies that $F^* \leq F^{**}$, which means, by induction, that $F' < F^* \leq F^{**}$. Thus, with $F \geq F^{**}$, Second Best equilibria are not implementable.

Third, with fines $F \geq F^{**}$, only the Third Best equilibria can be implemented. By proof of Proposition 5, the following incentive compatibility
constraints are necessary:

\[ \text{IC}_1 \quad \Pi(M = 1, T = 0, I_0 \in [0, 1], I_1 = I^{**}) \leq \Pi(M = 1, T = 1, I_0 \in [0, 1], I_1 = I^{**}), \]

\[ \text{IC}_2 \quad \Pi(M = 0, T \in \{0, 1\}, I_0 \in [0, 1], I_1 = I^{**}) \leq \Pi(M = 1, T = 1, I_0 \in [0, 1], I_1 = I^{**}). \]

Starting with IC$_1$:

\[-(1 - p_1)q_{01}I_0 F - p_1 c - m \leq -c - m, \]

\[ \frac{(1 - p_1)c}{(\alpha - (1 - p_1)q_{01})I_0} = \frac{(\alpha + \delta - 2\alpha\delta)c}{\alpha\delta I_0} \leq F. \]

Then IC$_2$:

\[ \lambda g - \alpha I_0 F \leq \lambda g - c - m, \]

\[ \frac{c + m}{\alpha I_0} \leq F. \]

Thus, Third Best equilibria are implementable with fines such that\(^{18}\)

\[ F(I_0) \geq \max \left\{ \frac{(\alpha + \delta - 2\alpha\delta)c}{\alpha\delta I_0}, \frac{c + m}{\alpha I_0} \right\}. \]

\[ F(I_0) \text{ can be smaller than } F^{**} \text{ with some } I_0 \in (0, 1]. \] First consider

\[ \frac{(\alpha + \delta - 2\alpha\delta)c}{\alpha\delta I_0} \leq \frac{c}{q_{01}I_0^*} = F^{**}, \]

\[ \frac{1 - \delta}{\delta} I_0^* \leq I_0 \]

and notice that both \( I_0^* \) and \( \frac{1 - \delta}{\delta} \) are smaller than one, by Proposition 1 and the \( 1/2 < \delta < 1 \) assumption, respectively. Second,

\[ \frac{c + m}{\alpha I_0} \leq \frac{c}{q_{01}I_0^*} = F^{**}, \]

\[ \frac{q_{01}}{\alpha} \left( I_0^* + \frac{m}{c} \right) \leq I_0 \]

because by inequality (8)

\[ \frac{1 - \delta}{\alpha + \delta - 2\alpha\delta} \left( I_0^* + \frac{m}{c} \right) \leq \frac{1 - \delta}{\alpha + \delta - 2\alpha\delta} \left( I_0^* + \frac{(1 - \alpha)(2\delta - 1)}{(1 - \delta)} \right) < 1 \]

and the inequality follows from \( I_0^* < 1 \). In sum, if inequality (8) holds, the third best is implemented with no reporting investigation such that

\[^{18}\text{Notice that the state can freely set } I_0 \text{ because there is no reporting in equilibrium.}\]
where \( I_0'(F) \leq 1 \) with \( F \geq F^{**} \).

Finally, it is shown that Third Best equilibria cannot be implemented with fines \( F < F^* \). To see this, consider the necessary fines for the third best:

\[
F(I_0) \geq \max \left\{ \frac{(\alpha + \delta - 2\alpha\delta)c + m}{\alpha\delta I_0}, \frac{c + m}{\alpha I_0} \right\} = \frac{c + m}{\alpha}.
\]

Consequently, it suffices to show that

\[
c + m \geq \frac{(1 - \alpha - \delta + 2\alpha\delta)c + m}{\alpha\delta I_0^*} = F^*,
\]

which is equivalent to

\[
c \geq \frac{(1 - \delta I_0^*)}{\delta I_0^* - (1 - \alpha - \delta + 2\alpha\delta)}m.
\]

Rearranging is possible because

\[
\delta I_0^* - (1 - \alpha - \delta + 2\alpha\delta) > 0 \tag{A10}
\]

to see that equation (A10) holds:

\[
\delta I_0^* > (1 - \alpha - \delta + 2\alpha\delta), \quad \frac{\alpha\delta(1 - \delta)\rho h}{(\alpha + \delta - 2\alpha\delta)(1 - \alpha - \delta + 2\alpha\delta)} > k.
\]

This latter statement follows from the assumption stated in equation (2):

\[
k \frac{(1 - \delta)}{(\alpha + \delta - 2\alpha\delta)} > \frac{\alpha\delta(1 - \delta)\rho h}{(\alpha + \delta - 2\alpha\delta)(1 - \alpha - \delta + 2\alpha\delta)}
\]

and observing that

\[
\frac{(1 - \delta)}{(\alpha + \delta - 2\alpha\delta)}k > k,
\]

which trivially follows from \( 1 - \alpha > \alpha \).

Now, turning back to equation (A9)—note that by equation (8),

\[
c \geq \frac{(1 - \delta)}{(1 - \alpha)(2\delta - 1)}m.
\]

So, it suffices to show that equation (8) implies equation (A9). To show this, first notice that the right-hand side of equation (A9) is decreasing in \( I_0^* \):

\[
\frac{\partial}{\partial I_0^*} \frac{(1 - \delta I_0^*)}{\delta I_0^* - (1 - \alpha - \delta + 2\alpha\delta)} = \frac{-\delta^2(\alpha + \delta - 2\alpha\delta)}{[\delta I_0^* - (1 - \alpha - \delta + 2\alpha\delta)]^2} < 0.
\]

Consequently, it is enough to show that equation (8) implies equation (A9) for \( I_0^* = 1 \) because it is implied then for all the other \( I_0^* \). To see this, consider

\[
\frac{(1 - \delta)}{\delta - (1 - \alpha - \delta + 2\alpha\delta)} = \frac{(1 - \delta)}{(1 - \alpha)(2\delta - 1)} < 1.
\]
Thus, with $I^* = 1$, equations (8) and (A9) are equivalent, and with any other $I^*_0$, equation (8) implies equation (A9), which finishes the proof as $I^*_0 < 1$ by the proof of Proposition 1.

Proof of Lemma 10. In the Second Best and Third Best equilibria, the expected prosecution rate is trivially

$$\chi^{**} = \alpha I^{**} = \frac{\alpha^2 \rho h}{2k}.$$  

In the First Best, from equation (A7)

$$\chi^* = (1 - p_1)q_0 I^*_0 + p_1 q_{11} I^*_1$$

$$= \alpha (1 - \delta) I^*_0 + \alpha \delta I^*_1$$

$$= \left( \frac{(1 - \delta)^2}{\alpha + \delta - 2\alpha \delta} + \frac{\delta^2}{1 - \alpha - \delta + 2\alpha \delta} \right) \frac{\alpha^2 \rho h}{2k}.$$  

To see that $\chi^* > \chi^{**}$, consider the following. First, the statement is equivalent to

$$1 < \frac{(1 - \delta)^2}{\alpha + \delta - 2\alpha \delta} + \frac{\delta^2}{1 - \alpha - \delta + 2\alpha \delta}. \quad (A11)$$

Notice further that, with $\alpha = 1/2$, the inequality can be written as

$$0 < 1/2 + 2\delta (1 - \delta) \quad (A12)$$

and it always holds when $\alpha = 1/2$.

Now, differentiate the right-hand side of equation (A11) with respect to $\alpha$:

$$\frac{\partial}{\partial \alpha} \left( \frac{(1 - \delta)^2}{\alpha + \delta - 2\alpha \delta} + \frac{\delta^2}{1 - \alpha - \delta + 2\alpha \delta} \right)$$

$$= (1 - 2\delta) \left[ \frac{\delta}{p_1} - \frac{1 - \delta}{1 - p_1} \right] \left[ \frac{\delta}{p_1} + \frac{1 - \delta}{1 - p_1} \right] < 0 \quad (A13)$$

because

$$p_1 = 1 - \alpha - \delta + 2\alpha \delta,$$

$$(1 - 2\delta) < 0,$$

$$\left[ \frac{\delta}{p_1} + \frac{1 - \delta}{1 - p_1} \right] > 0.$$  

The first two statements are obvious. For the third, consider that it is equivalent to

$$\delta + p_1 > 0,$$

which can be written as

$$\delta + 1 - \alpha - \delta + 2\alpha \delta > 0.$$
Simplifying yields

\[ 2\delta(1-\alpha) > (1-\alpha), \]

which always holds because \( \delta > \frac{1}{2} \).

Thus, by the negativity of derivative (A13), inequality (A11) holds with \( \forall \alpha \in (0, 1/2) \). \( \square \)

References


