Screening Ethics when Honest Agents Keep Their Word

Ingela Alger* and **Régis Renault****

April 2004

Abstract

Using the canonical principal-agent setting with adverse selection, we study the implications of honesty when it requires pre-commitment. Within a two-period hidden information problem, an agent learns his match with the assigned task in period 2 and, if honest, reveals it to the principal if he has committed to it. The principal may offer a menu of contracts to screen ethics. Both honest and dishonest agents are willing to misrepresent their ethics. Equilibrium ethics screening occurs if both honesty and a good match are sufficiently likely: the principal leaves a smaller rent to an honest while screening matches for a dishonest by inducing a message reversal by the dishonest. Otherwise, if dishonesty is likely, the principal offers the standard second-best contract, while if both dishonesty and a good match are unlikely, she offers the first-best contract, implying that no match screening occurs for a dishonest.

Keywords: ethics, honesty, loyalty, adverse selection, screening.
JEL classifications: D82

*Economics Department, Boston College, 140 Commonwealth Avenue, Chestnut Hill MA 02467. E-mail: ingela.alger@bc.edu

**THEMA, Université de Cergy-Pontoise, 33 Bld du Port, 95011 Cergy-Pontoise Cedex, France. E-mail: regis.renault@eco.u-cergy.fr

Acknowledgement: An earlier version of this paper was entitled "Honest Agents and Equilibrium Lies." We are grateful to seminar participants at MIT, Université de Caen, University of St Andrews, University of Virginia, and ESEM 99 for useful comments.

1 Introduction

For the past thirty years, a large body of literature has developed on the theme that individuals use their private information in an opportunistic manner. However, evidence suggests that some individuals behave honestly even if they thereby forgo material benefits. For instance, using data from a firm that kept monitoring at a high level while making the employees believe that it had been relaxed, Nagin et al. (2002) found that only some employees chose to increase the extent of shirking.[1] Some authors have introduced such an heterogeneity in ethics in various contexts. In Jaffee and Russell (1976), banks may ration credit if borrowers differ in their willingness to default on loans. Assuming that some taxpayers report their income truthfully regardless of economic incentives, Erard and Feinstein (1994) are able to rule out unrealistic equilibria. Analyzing various economic settings, Tirole (1992), Kofman and Lawarrée (1996), and Alger and Ma (2003) find that it is suboptimal to deter collusion if potential dishonesty is sufficiently unlikely.

Among these papers, only Alger and Ma allow for screening the agent's ethics, by which we mean that in equilibrium the mechanisms implemented for an honest and a dishonest agent may differ. They find that the principal chooses to resort to such screening only when the probability that the agent is potentially dishonest is sufficiently small. In this paper and our companion work, Alger and Renault (2003), we address this question in the canonical principal-agent model with adverse selection and look at the implications of different specifications of an honest behavior. Here we focus on a specification of honesty that is in line with that considered in Alger and Ma so that our results may be contrasted with theirs to yield some insight as to the robustness of their findings in contexts other than the specific health insurance setting they consider. We also explicitly spell out the underlying psychological motives for the type of honest behaviour that is assumed here.

We consider a model where a principal contracts with an agent, whose match with the principal may be either good or bad. Moreover, the agent may be either opportunistic or honest. The principal observes neither the agent's ethics, nor his match. A contract specifies

[1] See the references in Alger and Ma (2003) and Alger and Renault (2003) for further examples.

a decision (e.g., output) and a transfer (e.g., a wage). A first take at defining honesty within this model would be to consider that it involves revealing any private information at no cost. In our companion paper we consider such an "unconditional honesty" as a benchmark; we show that the optimal contract then always involves ethics screening. The principal specifies the same decision for an honest and an opportunistic agent, while leaving a rent only to the latter. However, such a definition of honesty does not receive much support from modern psychology. In 1928 Hartshorne and May laid out what is still the basic tenet for much of this literature, namely, that moral behavior cannot be viewed as emanating from "an inner entity operating independently of the situations in which the individuals are placed" (p.385). Instead, the general perception is that behavior is conditional on various factors; in particular, studies have shown that such concerns as financial needs, the fear of getting caught, perceived equity, and loyalty all affect the propensity to behave honestly (see, e.g., Spicer and Becker, 1980, and Terris and Jones, 1982), all of which would translate differently into a formal model. In our companion paper, we focus on honesty as being conditional on the perceived equity of the contract. In this paper, we instead look at loyalty as a trigger of honest behavior.

The following two examples illustrate situations where a sense of loyalty may lead an agent to follow the prescriptions of some explicit or implicit contract. In subcontracting situations, such as with plumbers and carpenters, there frequently is *ex ante* uncertainty pertaining to the cost of the assignment. Typically the subcontractor would learn the relevant information while completing the job, whereas the customer would remain uninformed. A prior discussion of these uncertainties with the house owner might prevent some subcontractors from inflating the cost. In the workplace, an employee may discover how well he is matched with a given task only in the course of performing it: either because of unforeseen contingencies, or because he has never performed this task before. This may for instance result in different times needed to achieve a given output. In this case an honest employee would reveal this information if he has committed to doing so by signing a pre-specified contract with the employer.

In their study of commitment in the workplace, Meyer and Allen (1991) use answers to surveys to identify three types of commitment: "Affective commitment refers to the employee's emotional attachment to, identification with, and involvement in the organization. (...) Continuance commitment refers to an awareness of the costs associated with leaving the organization. (...) Finally, normative commitment reflects a feeling of obligation to continue employment." (p.67) In their 1997 book, they devote a chapter to understanding how organizational commitment develops. Existing research indicates that individuals vary in their propensity to become committed, but that "relations between demographic variables and affective commitment are neither strong nor consistent" (p.43). These findings suggest that commitment may be important for some individuals only and that there seems to be no clear relation between commitment and observable differences such as gender or age. They further report evidence pointing to a link between an employee's sense of commitment and his honesty (or lack of willingness to shirk); numerous studies have shown that an employee's degree of commitment to the firm is positively correlated with (*i*) his/her likelihood of "voluntary absence" (absenteeism due to reasons that were under the employee's own control); (*ii*) his job performance.

We use these insights to formalize honesty in a two-period principal-agent setting. The agent discovers his match in the second period. However, he knows his ethics in the first period. An honest agent would reveal his match truthfully provided that a contract specifying an allocation contingent on his match has been signed in the first period. In the first period the principal may offer a menu of contracts in order to screen ethics. Because the agent has not yet developed a sense of commitment in the first period, he is willing to misrepresent his ethics, should that give him a higher expected surplus.

The principal would ideally choose to leave no rent to an honest agent, while leaving a rent to an opportunistic agent in order to screen matches. But since an honest agent is willing to lie about his ethics, inducing truthful revelation of the match by an opportunistic agent requires leaving as large a rent for an honest as for an opportunistic agent. The only way for the principal to leave a smaller rent to an honest agent is to allow for misrepresentation of the match by an opportunistic agent. This means that the revelation principle does not apply. We show that the only relevant equilibrium message structures involving misrepresentation are either to always claim that the match is bad, or to systematically lie about the match. With the former, the principal forgoes any benefit from match screening for an opportunistic agent, but may leave no rent to an honest. The message structure involving a systematic lie enables the principal to screen the match for an opportunistic (because lies are revealing in this case), at the cost of leaving some rent to an honest. However, because a dishonest obtains a rent by misrepresenting matches, which an honest is not willing to do, an honest agent's rent is smaller than that of a dishonest. Moreover, the match screening conditional on the agent being opportunistic is third-best in order to limit the honest agent's rent.

Among the three candidates for an equilibrium contract, only the revealing-lies contract involves actual ethics screening. The intuition is as follows. First, if the principal wants to induce truthful match revelation, she would simply offer the standard second-best contract: since the rent then is independent of ethics, it is as if the agent were opportunistic with certainty. Second, since in the optimal non-revealing lies contract the opportunistic agent always claims that the match is bad, it is as if the dishonest agent represented a bad match with certainty: it is then optimal to specify first-best allocations, independent of ethics.

Interestingly, although the revealing-lies contract enables the principal to screen both ethics and matches, it is likely to be dominated by some other contract. Indeed, it is only optimal if honesty is sufficiently likely and if the agent has a good match with the job with a high enough probability. In particular, it is dominated by the standard second-best contract if the the agent is more likely to be opportunistic than honest. As explained above, the revealing-lies contract enables the principal to leave a smaller rent to the honest by letting the opportunistic misrepresent matches in equilibrium, and by exploiting the fact that the honest is not willing to misrepresent the match should he choose the contract meant for the opportunistic agent. Clearly, if the probability that the agent is opportunistic is above one half, the principal would like to do the reverse, namely, give a smaller rent to the opportunistic. However, this is impossible since the opportunistic agent is unrestrained in his willingness to misrepresent the match. When the agent is more likely to be honest, the revealing-lies contract is preferred to the first-best contract only if a good match is likely enough, since only then does the benefit from "third-best" match screening for the opportunistic outweigh the cost of leaving a rent to the honest.

If the match is likely to be bad, so that the principal does not resort to ethics screening, the standard second-best contract clearly dominates the first-best contract if the probability of an opportunistic agent is large enough while this is reversed if the probability of an opportunistic agent is low.

In Alger and Ma (2003), an insurer contracts with a patient and a physician. With some probability the physician is honest: he is not willing to submit a fraudulent claim about the patient's health state, even if that would increase their joint surplus. An insurance contract is signed before the uncertainty about the patient's health state is resolved. Thus, the information and contracting structures are identical to ours. Like us, Alger and Ma find that the equilibrium contract is the standard second-best one if the probability of an opportunistic physician is sufficiently large. Apart from that the results are quite different. First, in their setting the principal may not induce the opportunistic physician to systematically lie; this is due to an exogenous restriction on the contract. Second, the contract involving a non-revealing lie screens ethics, both in terms of the decision (the amount of treatment) and the transfers. Furthermore, the first-best contract is never optimal for a positive probability of dishonesty; this is a result of the insurance motive, which calls for a smaller difference between the allocations associated with an honest and an opportunistic physician. The risk aversion also renders the analysis more complex; in fact, they were not able to fully characterize the optimal contract.[2] We do not face such a problem in the standard setting with quasi-linear preferences.

[2]They showed that in equilibrium the insurer would offer the contract involving a non-revealing lie for a probability of dishonesty close to zero, and the standard second-best contract for a probability of dishonesty close to one. However, they could not rule out the possibility that there might be more than one switching between the former and the latter contract as that probability increases from zero to one.

In Alger and Renault (2003), as a benchmark we consider unconditional honesty. The optimal contract leaves no rent to an honest agent, independent of the probability of honesty, so that the standard second-best approach is not robust.[3] This is true whether or not an honest agent is willing to misrepresent his ethics. Indeed, the principal is able to leave no rent to an honest agent by letting the opportunistic agent lie about both his ethics and his match when his match is bad. An honest agent may receive a rent only if his behavior is conditional on the contract being fair, which is the main focus of that paper. Here we find that an *ex ante* opportunistic behavior regarding ethics may result in a rent for an honest agent, even though honesty is not conditional on fairness. The reason is that here the principal is somewhat limited in ways to fool the honest; in particular, she may not induce the opportunistic to misrepresent his ethics conditional on his match realization. We further find that honesty being conditional on fairness limits the principal's ability to screen ethics more than loyalty does: indeed, the revealing-lies contract in this paper would not be deemed fair. Thus, the robustness of the standard second-best approach is obtained, and the first-best contract is optimal if honesty is sufficiently likely.

The next section introduces the formal model. In Section 3, we characterize the optimal contract, and we conclude in Section 4. Proofs are in appendix.

2 The Model

We model an honest behavior triggered by a commitment to reveal some private information. To this end we consider a two-period framework where a contract is signed in period 1 prior to private information being revealed to the agent. In many situations, there may also be some private information in period 1, however we abstract from it for the sake of simplicity.

Although our model applies to many situations, we specify it to suit the employer-employee relationship. An employer (the principal) hires an individual (the agent) to

[3] Deneckere and Severinov (2001) and Forges and Koessler (2003) develop a general framework allowing for a larger class of mechanisms; when applied to the setting in Alger and Renault (2003) they predict the same equilibrium allocations.

perform a task with output x. This output gives the principal surplus $\pi(x)$, where π is a strictly increasing and concave function, with $\pi(0) = 0$ and $\pi'(0) = +\infty$. The principal pays the agent a transfer t, so that her net surplus is $\Pi(x, t) = \pi(x) - t$.

There is uncertainty as to how well the agent's abilities match the requirements of the task he is assigned. Let $v(x, \theta)$ denote the cost for the agent of producing x units, where the parameter θ represents the match.[4] The match may be either good ($\theta = \underline{\theta}$) or bad ($\theta = \bar{\theta}$), with the probability of a bad match being α. The cost function v is strictly increasing and strictly convex in x, with $v(0, \theta) = 0$ and $v'(0, \theta) = 0$ (the prime indicates a partial derivative with respect to x). Finally, when the match is good, the production cost incurred by the agent is smaller than when it is bad, for any output level: $\forall x > 0,\ v'(x, \bar{\theta}) > v'(x, \underline{\theta})$. This is the standard Spence-Mirrlees condition.[5] The employee's net surplus is the transfer net of the production cost $U(x, t, \theta) = t - v(x, \theta)$. The agent's reservation utility is normalized to zero. For further reference, the first-best values of x, $\bar{x}^*$ if $\theta = \bar{\theta}$ and $\underline{x}^*$ if $\theta = \underline{\theta}$, are defined by $\pi'(\bar{x}^*) = v'(\bar{x}^*, \bar{\theta})$ and $\pi'(\underline{x}^*) = v'(\underline{x}^*, \underline{\theta})$, respectively. Given the properties of π and v, they are uniquely defined, and $\bar{x}^* < \underline{x}^*$. We will call an output with its corresponding transfer an allocation $y = (x, t)$.

In the standard setting, the principal would discriminate between a good and a bad match using a mechanism, to be at work in the second period:

Definition 1 (Mechanism) *A mechanism m, defines a space $\mathcal{M}_2$ of messages μ_2 and a mapping $y_m : \mathcal{M}_2 \to \mathbb{R}^+ \times \mathbb{R}^+$, where $\mathbb{R}^+ \times \mathbb{R}^+$ is the set of feasible allocations. A mechanism is said to be direct if the message space is the set of matches, $\mathcal{M}_2 = \{\underline{\theta}, \bar{\theta}\}$; otherwise, it is indirect.*

[4]The results would not be substantially affected if we allowed for θ to directly affect the principal's surplus π (see further the discussion following Proposition 4); however, the notation and some of the arguments used in the proofs would be more cumbersome.

[5]The assumptions on the functional forms essentially ensure unique interior solutions. Other applications may require these regularity conditions to be slightly altered. For instance, if the principal is a seller signing a contract with a buyer, x being the quantity supplied, then π would be a decreasing and v an increasing function, and t would be negative. The parameter θ would determine the buyer's willingness to pay. A buyer may feel committed to reveal a high willingness to pay for an artist's work upon having a discussion with the artist.

The principal faces an additional uncertainty, regarding the agent's ethics, denoted k: with probability $\gamma \in (0, 1)$, the agent is dishonest ($k = d$), and with the complementary probability, he is honest ($k = h$).[6] The agent knows his ethics during the whole game, and it is his private information. A dishonest agent always maximizes his surplus, like in any standard principal-agent setting. By contrast, an honest agent may forego material benefits to truthfully announce his type θ (at date 2). In order for him to feel committed to truthful match revelation in period 2, it must be that all allowed messages specify a match realization. If a mechanism specifies messages that are unrelated to the true values of θ, telling the truth is a fuzzy notion and we might as well assume that an honest agent would not be reluctant to announce such messages, so that the principal could not gain anything by using them.[7] We further assume that an honest agent does not feel committed to revealing his match truthfully if there are several messages specifying the same match realization.[8] From this discussion, the contract offered in the first period may be described as a menu of direct mechanisms:

Definition 2 (Contract) *A contract C defines a space $\mathcal{M}_1$ of messages μ_1 that may be sent by the agent at date 1, and a mapping $c : \mathcal{M}_1 \to M_D$ where M_D is the set of direct mechanisms m (see definition 1).*

Formally, honesty is equivalent to the following restriction on the message space at date 2:

[6] All probability distributions are common knowledge.

[7] This is related to the idea that honest behavior may be linked to a fear of being found out lying. In many cases, one can argue that the value of θ is verifiable information, albeit at a very high cost; a potentially honest agent may underestimate that cost. But if the honest agent sends the message "blue" for instance, there is no information to be verified.

[8] There could be several messages specifying a given realization of θ along with some other item. As we argue in Alger and Renault (2003), using such a rich message structure would enable the principal to implement the same allocation as if honesty was unconditional. This is true even if an honest agent feels no restriction as to his announcement of the additional item: for instance if the second item is ethics, the result holds even if he is willing to lie about ethics (these results build on the work of Deneckere and Severinov, 2001, and Forges and Koessler, 2003). Our goal here is to analyze honesty as being conditional on commitment in a way that constrains the principal away from the solution achieved with unconditional honesty. By restricting the set of relevant second period mechanisms, we ensure that the principal may only obtain commitment by an honest agent at some cost.

Assumption 1 (Ethics I) *(Messages at date 2)*

(i) A dishonest may announce any $\theta \in \{\underline{\theta}, \bar{\theta}\}$.

(ii) An honest with match $\theta \in \{\underline{\theta}, \bar{\theta}\}$ *must announce* θ *if a contract was signed at date 1; otherwise, he may announce any* $\theta \in \{\underline{\theta}, \bar{\theta}\}$.

Since no contract has been signed prior to date 1 an honest agent does not feel committed to revealing his ethics. Formally, we have:

Assumption 2 (Ethics II) *Whether the agent is honest or dishonest, he may announce any* $\mu_1 \in \mathcal{M}_1$.

For further use below, let $\hat{\theta}^*(\theta, m)$ denote the agent's best response in mechanism m if his match is θ, and $\mu_1^*(k)$ the best response at date 1. As a tie-breaking rule, we assume that whenever an agent is indifferent between revealing the truth and lying, he reveals the truth.

As a benchmark we describe the optimal contracts if ethics were known. First, if the principal knew that the agent is honest, she would implement the first-best decisions and extract the whole surplus by offering a contract with one message leading to the first-best mechanism:

Definition 3 (First-best mechanism) *The first-best mechanism, denoted* m^*, *defines* $\mathcal{M}_2 = \{\underline{\theta}, \bar{\theta}\}$, *and the mapping:* $(\bar{x}^*, \bar{t}^*)$ *if* $\mu_2 = \bar{\theta}$, *and* $(\underline{x}^*, \underline{t}^*)$ *if* $\mu_2 = \underline{\theta}$, *where* $\bar{t}^* = v(\bar{x}^*, \bar{\theta})$ *and* $\underline{t}^* = v(\underline{x}^*, \underline{\theta})$.

If the agent is known to be dishonest, the revelation principle applies and the mechanism that maximizes the principal's expected surplus may be found by imposing individual rationality and incentive compatibility constraints. This yields the standard second-best mechanism:

Definition 4 (Standard second-best mechanism) *The standard second-best mechanism, denoted* m^s, *defines* $\mathcal{M}_2 = \{\underline{\theta}, \bar{\theta}\}$, *and the mapping:* $(\bar{x}^s, \bar{t}^s)$ *if* $\mu_2 = \bar{\theta}$, *and* $(\underline{x}^s, \underline{t}^s)$ *if* $\mu_2 = \underline{\theta}$, *where*

$$\bar{t}^s = v(\bar{x}^s, \bar{\theta}) \qquad \underline{t}^s = v(\underline{x}^*, \underline{\theta}) + [v(\bar{x}^s, \bar{\theta}) - v(\bar{x}^s, \underline{\theta})]$$

$$v'(\bar{x}^s, \bar{\theta}) = \pi'(\bar{x}^s, \bar{\theta}) - \frac{1-\alpha}{\alpha}[v'(\bar{x}^s, \bar{\theta}) - v'(\bar{x}^s, \underline{\theta})] \qquad \underline{x}^s = \underline{x}^*.$$

In equilibrium, the agent receives a rent if the match is good, and the allocation is sub-optimal if the match is bad.

If ethics is the agent's private information, and if the first-period menu offered by the principal comprises the first-best and the standard second-best mechanism, the agent always selects the wrong mechanism from the principal's viewpoint. We next turn to determining the optimal contract.

3 Analysis

3.1 Preliminaries

The standard approach to the present asymmetric information problem would rely on the revelation principle, whereby for any contract there exists a direct incentive compatible contract that implements the same allocation in every state of nature. Here it is straightforward to prove that this principle does not hold. Consider the following contract: $(m_h, m_d) = (m^*, m^*)$. In equilibrium, the honest agent reveals θ truthfully and gets no rent. The dishonest agent always announces $\bar{\theta}$; he gets no rent if $\theta = \bar{\theta}$, and the rent $v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})$ if $\theta = \underline{\theta}$. According to the revelation principle, there should exist a menu of direct mechanisms (m'_h, m'_d) yielding the same allocations as the original ones in every state of nature, and inducing the agent to tell the truth. Therefore, m'_d must specify the allocation $(\bar{x}^*, \bar{t}^*)$ whether the agent announces $\underline{\theta}$ or $\bar{\theta}$. As a result, for the honest agent to reveal h truthfully, m'_h must give him an expected rent $(1-\alpha)[v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})] > 0$, which yields a contradiction with the requirement that the allocation should be the same in every state of nature. Hence the revelation principle does not hold in our framework.[9]

[9] This is similar, although not formally equivalent, to the result in Green and Laffont (1986) that in a one-stage information revelation problem, when an agent's message space is restricted, the revelation principle may fail to apply. See Deneckere and Severinov (2001), Forges and Koessler (2003), and Alger and Renault (2003) for further discussions of this issue in one-stage models. Alger and Ma (2003) have the same informational and contractual structure as we do; as a result this observation is true in their setting as well.

Given that, the agent's equilibrium behavior is *a priori* not predictable. At first sight, the task to determine the optimal contract thus seems daunting. The following observation, which was also made by Alger and Ma (2003), facilitates the determination of the optimal contract.

Lemma 1 *Without loss of generality, attention may be restricted to contracts C that define the message space $\mathcal{M}_1 = \{h, d\}$ and that induce truth-telling at date 1, i.e, $\mu_1^*(h, C) = h$ and $\mu_1^*(d, C) = d$.*

Lemma 1 implies that we may impose two date 1 incentive compatibility constraints ensuring that an honest agent announces h while a dishonest agent announces d. We may now use a lighter notation: a mechanism m_k, where $k \in \{h, d\}$, specifies the allocation $(\underline{x}_k, \underline{t}_k)$ if the date 2 message is $\underline{\theta}$, and $(\bar{x}_k, \bar{t}_k)$ if the date 2 message is $\bar{\theta}$. We will say that ethics screening occurs if $m_d \neq m_h$.

Given that the equilibrium behavior of the agent is not known *a priori*, we have to determine the contracts that are optimal conditional on assumed behavior, before being able to determine the equilibrium contract. We start by analyzing mechanisms that do induce truth-telling at date 2 by the dishonest agent, before turning to mechanisms inducing lies.

3.2 Incentive-compatible contracts

Here we assume that in equilibrium a dishonest agent reveals θ at date 2: $\forall \theta \in \{\underline{\theta}, \bar{\theta}\}$ we have $\hat{\theta}^*(\theta, d, m_d) = \theta$. Thus the principal chooses the pair of mechanisms $\{m_d, m_h\} = \{\{\bar{x}_d, \bar{t}_d, \underline{x}_d, \underline{t}_d\}, \{\bar{x}_h, \bar{t}_h, \underline{x}_h, \underline{t}_h\}\}$ so as to maximize:

$$\gamma[\alpha(\pi(\bar{x}_d) - \bar{t}_d) + (1 - \alpha)(\pi(\underline{x}_d) - \underline{t}_d)]$$

$$+(1 - \gamma)[\alpha(\pi(\bar{x}_h, \bar{\theta}) - \bar{t}_h) + (1 - \alpha)(\pi(\underline{x}_h, \underline{\theta}) - \underline{t}_h)].$$

For the agent to reveal his ethics at date 1, the following incentive compatibility constraints must hold:

$$(1) \qquad \alpha[\bar{t}_h - v(\bar{x}_h, \bar{\theta})] + (1 - \alpha)[\underline{t}_h - v(\underline{x}_h, \underline{\theta})] \geq$$

$$\alpha[\bar{t}_d - v(\bar{x}_d, \bar{\theta})] + (1-\alpha)[\underline{t}_d - v(\underline{x}_d, \underline{\theta})]$$

(2)
$$\alpha[\bar{t}_d - v(\bar{x}_d, \bar{\theta})] + (1-\alpha)[\underline{t}_d - v(\underline{x}_d, \underline{\theta})] \geq$$

$$\alpha\, max\{\bar{t}_h - v(\bar{x}_h, \bar{\theta}), \underline{t}_h - v(\underline{x}_h, \bar{\theta})\} + (1-\alpha) max\{\underline{t}_h - v(\underline{x}_h, \underline{\theta}), \bar{t}_h - v(\bar{x}_h, \underline{\theta})\}$$

An honest agent reveals θ truthfully irrespectively of the mechanism chosen at date 1. Therefore, the incentive constraint for the honest type (1) is straightforward. In contrast, we do not know the dishonest agent's behavior if he were to claim to be honest at date 1: therefore, the right-hand side of constraint (2) is not based on truth-telling. However, by assumption, the dishonest agent truthfully reveals θ in equilibrium, yielding the expression on the left-hand side of the constraint. Second, for the dishonest agent to reveal his type θ in equilibrium, two incentive compatibility constraints must hold, but as is standard only the following is binding:

(3)
$$\underline{t}_d - v(\underline{x}_d, \underline{\theta}) \geq \bar{t}_d - v(\bar{x}_d, \underline{\theta})$$

Finally, the participation constraints are:

(4)
$$\bar{t}_k - v(\bar{x}_k, \bar{\theta}) \geq 0 \quad k = h, d$$

(5)
$$\underline{t}_k - v(\underline{x}_k, \underline{\theta}) \geq 0 \quad k = h, d.$$

The principal maximizes his expected utility subject to the constraints (1)-(5), yielding:

Proposition 1 *The optimal contract inducing the dishonest to reveal his true θ is to offer the standard second-best mechanism independent of the announced ethics.*

The intuition is as follows. The principal must give the dishonest agent a rent to make him reveal θ at date 2. But since the honest agent behaves opportunistically at date 1, she must give him the same rent. As a result, the optimal contract is as if the agent were dishonest with certainty. Proposition 1 indicates that the only way to leave a smaller rent to the honest than to the dishonest agent is to let the dishonest agent manipulate the information at date 2.

3.3 Contracts inducing lies

From Lemma 2 we may without loss of generality impose date 1 incentive compatibility constraints, which ensure truthful ethics revelation. Which type of second period incentive compatibility constraints should be imposed, depends on the assumed equilibrium and out-of-equilibrium behavior of the dishonest agent; the exact formulation of all constraints also depends on this. Our first goal is to restrict the set of such equilibrium and out-of-equilibrium behavior. This is achieved by the following lemma.

Lemma 2 *If m_d does not induce the dishonest agent to reveal the truth at date 2, then*

(i) in equilibrium, the dishonest agent either sends the message $\hat{\theta} = \bar{\theta}$ independently of his type, or always sends a false message so that $\hat{\theta} = \bar{\theta}$ when he is of type $\underline{\theta}$ and $\hat{\theta} = \underline{\theta}$ when he is of type $\bar{\theta}$.

(ii) if the dishonest agent were to pick m_h, he would send the message $\hat{\theta} = \bar{\theta}$ independently of his type.

Regarding the out-of-equilibrium messages, it would clearly not be optimal to design m_h so as to induce the dishonest to reveal θ truthfully, as that would simply bring us back to the incentive-compatible case. It is also intuitive that the principal should not induce the dishonest with type $\bar{\theta}$ to announce $\underline{\theta}$ in m_h; then there would be no benefit to screening matches for honest agents, since an agent with a good match should be paid as much as an agent with a bad match for a given output. This argument does not apply to a dishonest agent's equilibrium messages: if a dishonest agent of type $\bar{\theta}$ announces $\underline{\theta}$, the principal may still screen matches by letting him announce $\bar{\theta}$ if his type is $\underline{\theta}$.

We can now proceed to characterizing the optimal contract, conditional on some assumed equilibrium behavior. As implied by Lemma 2, we need to consider only two cases: the revealing lies case (when he always lies in equilibrium, the principal can infer θ), and the non-revealing lies case (the dishonest agent always claims $\theta = \bar{\theta}$ in equilibrium).

In the revealing lies case the principal chooses the pair of mechanisms $\{m_d, m_h\} = \{\{\bar{x}_d, \bar{t}_d, \underline{x}_d, \underline{t}_d\}, \{\bar{x}_h, \bar{t}_h, \underline{x}_h, \underline{t}_h\}\}$, so as to maximize:

$$\gamma[\alpha(\pi(\underline{x}_d) - \underline{t}_d) + (1-\alpha)(\pi(\bar{x}_d) - \bar{t}_d)] \tag{6}$$

$$+(1-\gamma)[\alpha(\pi(\bar{x}_h) - \bar{t}_h) + (1-\alpha)(\pi(\underline{x}_h) - \underline{t}_h)].$$

In the appendix we verify that the only binding constraints are the incentive compatibility constraints ensuring that the agent reveals whether he is honest or dishonest:

$$\alpha[\bar{t}_h - v(\bar{x}_h, \bar{\theta})] + (1-\alpha)[\underline{t}_h - v(\underline{x}_h, \underline{\theta})] \geq \tag{7}$$

$$\alpha\ max\{\bar{t}_d - v(\bar{x}_d, \bar{\theta}), 0\} + (1-\alpha)\ max\{\underline{t}_d - v(\underline{x}_d, \underline{\theta}), 0\}$$

$$\alpha[\underline{t}_d - v(\underline{x}_d, \bar{\theta})] + (1-\alpha)[\bar{t}_d - v(\bar{x}_d, \underline{\theta})] \geq \bar{t}_h - \alpha v(\bar{x}_h, \bar{\theta}) - (1-\alpha)v(\bar{x}_h, \underline{\theta}), \tag{8}$$

and the participation constraints corresponding to a bad match, which is given by (4) for an honest and by

$$\underline{t}_d - v(\underline{x}_d, \bar{\theta}) \geq 0 \tag{9}$$

for a dishonest. The following proposition describes the optimal revealing lies contract.

Proposition 2 *(Revealing lies) For $\gamma < 1/2$, the optimal contract inducing the dishonest to always lie in equilibrium is such that $\bar{x}_d^R = \underline{x}_h^R = \underline{x}^*$,*

$$v'(\underline{x}_d^R, \bar{\theta}) = \pi'(\underline{x}_d^R) - \frac{1-\gamma}{\gamma}\frac{1-\alpha}{\alpha}[v'(\underline{x}_d^R, \bar{\theta}) - v'(\underline{x}_d^R, \underline{\theta})]$$

$$v'(\bar{x}_h^R, \bar{\theta}) = \pi'(\bar{x}_h^R) - \frac{\gamma}{1-\gamma}\frac{1-\alpha}{\alpha}[v'(\bar{x}_h^R, \bar{\theta}) - v'(\bar{x}_h^R, \underline{\theta})]$$

and

$$\underline{t}_d^R = v(\underline{x}_d^R, \bar{\theta}) \qquad \bar{t}_d^R = v(\bar{x}_d^R, \underline{\theta}) + [v(\bar{x}_h^R, \bar{\theta}) - v(\bar{x}_h^R, \underline{\theta})]$$

$$\bar{t}_h^R = v(\bar{x}_h^R, \bar{\theta}) \qquad \underline{t}_h^R = v(\underline{x}_h^R, \underline{\theta}) + [v(\underline{x}_d^R, \bar{\theta}) - v(\underline{x}_d^R, \underline{\theta})].$$

Both the honest and the dishonest obtain a rent, however it is lower for the honest thanks to the message reversal. Here the rent-efficiency trade-off is implied by the ethics incentive compatibility constraints being binding. If a dishonest claimed to be honest he would earn a rent by pretending that the match is bad when it is good, while if an honest agent claimed to be dishonest he would earn a rent by telling the truth when the match is good. Hence the rent of an honest depends positively on $\underline{x}_d$, whereas that of a dishonest depends positively on $\bar{x}_h$. This in turn implies that the quantities $\bar{x}_h$ and $\underline{x}_d$ are distorted downwards, and depend on the probability that the agent is dishonest. Since for $\gamma < 1/2$ it is more likely that the agent is honest, the downwards distortion is more pronounced for $\underline{x}_d$ than for $\bar{x}_h$; moreover, we have $\underline{x}_d < \bar{x}^s < \bar{x}_h$, $\bar{x}_h$ is increasing and $\underline{x}_d$ decreasing in γ, and both tend to $\bar{x}^s$ as γ tends to one half. Thus, the honest agent's rent rises as the probability that the agent is dishonest becomes larger; in fact, the mechanism intended for the honest agent tends toward the first-best mechanism as that probability tends to zero.

Finally, the proposition only specifies an optimal revealing lies contract for values of γ less than $\frac{1}{2}$. In this case the optimal allocation obtained by taking into account only the ethics incentive compatibility constraints satisfies the strict inequalities ensuring that the dishonest always lies; those are strict because of our tie-breaking rule that when indifferent the agent tells the truth. When $\gamma > \frac{1}{2}$ these inequalities would not hold for the allocation specified in the proposition. There is actually no optimal revealing lies contract in this case because the set of revealing lies contract is not a closed subset of $\mathbb{R}^8$. However, the surplus that the principal can achieve with a revealing lies contract is strictly less than with the standard second-best contract if $\gamma > \frac{1}{2}$. To see this, note that the set of revealing lies contracts would be larger without the tie-breaking rule; without this rule it can be verified that the optimal revealing lies contract implements the standard second-best allocations. The intuition is as follows. As we pointed out above, for γ below one half the principal chooses a lower quantity for a dishonest with a bad match than for an honest with a bad match ($\underline{x}_d < \bar{x}^s < \bar{x}_h$), because she is more likely to leave a rent to an honest than to a dishonest. By the same token, for γ above one half, she would like to leave a smaller rent to a dishonest agent than to an honest, which would require $\bar{x}_h < \bar{x}^s < \underline{x}_d$. This however would violate the second period constraint ensuring that a dishonest with a good match chooses $\bar{x}_d$. The latter constraint is therefore binding, and along with the first period ethics incentive constraints, it induces the principal to implement the standard second-best allocations.

In the non-revealing lies case the principal chooses the pair of mechanisms $\{m_d, m_h\} = \{\{\bar{x}_d, \bar{t}_d, \underline{x}_d, \underline{t}_d\}, \{\bar{x}_h, \bar{t}_h, \underline{x}_h, \underline{t}_h\}\}$, so as to maximize:

$$\gamma[\alpha\pi(\bar{x}_d) + (1-\alpha)\pi(\bar{x}_d) - \bar{t}_d] \tag{10}$$

$$+(1-\gamma)[\alpha(\pi(\bar{x}_h) - \bar{t}_h) + (1-\alpha)(\pi(\underline{x}_h) - \underline{t}_h)],$$

To characterize the optimal contract in this case it is enough to take into account the participation constraints (4) for $k = h, d$, and (5) for $k = h$.

Proposition 3 *(Non-revealing lies) An optimal contract inducing the dishonest to always claim that the match is bad is to offer the first-best mechanism independent of the announced ethics.*

The non-revealing lies contract implements the first-best allocations if the agent is honest, whereas the allocation is independent of the match if the agent is dishonest.[10]

3.4 The equilibrium contract

Having determined the contracts that are optimal conditional on the three relevant message structures, we can characterize the unconditional equilibrium. The standard second-best and revealing-lies contracts together define the largest surplus that the principal can achieve while screening matches for the dishonest: the former dominates for $\gamma \geq \frac{1}{2}$, while the latter is preferred otherwise. Hence to derive the equilibrium contract we need to compare this surplus with that obtained with the first-best contract. Figure 1, where the two relevant

[10]The allocation corresponding to a good match for the dishonest is in fact indeterminate, since it is never chosen in equilibrium.

surpluses (first-best and match screening) are drawn as a function of γ, depicts the three situations that may arise. We show in the proof of Proposition 4 that the surplus curves are as shown in the picture (Π^S denotes the surplus with the standard second-best contract, Π^R the surplus derived from the revealing lies contract, and Π^* the surplus associated to the first-best contract).

In panel (a) the first-best surplus curve always lies beneath the match screening surplus curve, so that the equilibrium contract is whichever match screening contract is relevant. As we will see in the proposition below, this occurs when the probability of a bad match is small. Intuitively this makes sense, since resorting to the first-best contract is costly only when the match is good. By contrast, when this probability is large, offering the first-best contract is always optimal when dishonesty is sufficiently unlikely, as shown in panels (b) and (c). Then the equilibrium contract depends on whether the first-best surplus crosses the second-best surplus above or below $\gamma = \frac{1}{2}$. In the former case, the revealing lies contract is never optimal (panel c), whereas in the latter case it is optimal for intermediate values of γ (panel b). In Alger and Renault (2003), where we are also led to compare the surpluses derived from offering the first-best and second-best contracts, we show that the threshold value for γ above which the second-best dominates the first-best contract, γ^*, is strictly increasing in α, the probability of a bad match.[11] We also find that it is bounded above by

$$\hat{\gamma} \equiv \frac{v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})}{[\pi(\underline{x}^*) - v(\underline{x}^*, \underline{\theta})] - [\pi(\bar{x}^*) - v(\bar{x}^*, \bar{\theta})]} < 1. \tag{11}$$

These results are summarized in the following proposition, and are visualized in Figure 2.[12]

Proposition 4 *There exists a unique $\alpha_1 \in (0, 1)$ and a unique $\alpha_2 \in (\alpha_1, 1)$ such that:*

1. If $\alpha \leq \alpha_1$: the principal offers the incentive-compatible contract if $\gamma \geq 1/2$, and the revealing-lies contract otherwise.

[11]Note that the difference in the principal's surplus from offering the incentive compatible and the non-revealing lies contracts is clearly monotonic in γ; in their setup, where the non-revealing lies contract was more complex, Alger and Ma (2003) were not able to establish such a monotonicity result.

[12]In the figure we have assumed that γ^{**} is a monotonic function of α, and that both γ^* and γ^{**} are linear in α.

2. If $\alpha \in [\alpha_1, \alpha_2)$, there exists a unique $\gamma^{**}(\alpha) \in (0, 1/2)$ such that the principal offers the incentive-compatible contract if $\gamma \geq 1/2$, the revealing-lies contract if $\gamma \in [\gamma^{**}, 1/2)$, and the non-revealing lies contract if $\gamma < \gamma^{**}$.

3. If $\alpha \geq \alpha_2$, there exists a unique $\gamma^{*}(\alpha) \in (1/2, 1)$ such that the principal offers the incentive-compatible contract if $\gamma \geq \gamma^{*}$, and the non-revealing lies contract otherwise.

In our setup the standard principal-agent model, where the agent is dishonest with certainty, is robust to the introduction of honest agents: for every value of the probability α of a bad match, the principal offers the standard second-best contract independent of ethics when dishonesty is sufficiently likely. This robustness is strongest for small values of α, when screening matches for the dishonest is highly valuable. If $\alpha < \alpha_1$, the benefits of screening matches for the dishonest are so large that the honest receives a positive rent for any value of the probability γ that the agent is dishonest: the principal offers the standard second-best contract if $\gamma \geq 1/2$, and the revealing-lies contract otherwise. For intermediary values of α, the benefits from screening matches for the dishonest are smaller; therefore, when γ is sufficiently small the principal forgoes them altogether and leaves no rent to the honest, by offering the first-best contract. However, ethics screening then still occurs for intermediary values of γ, and the standard second-best contract is offered whenever dishonesty is more likely than honesty. As α increases beyond α_2, though, the benefits from leaving no rent to an honest outweigh the benefits from screening matches for a dishonest even if dishonesty is more likely than honesty ($\gamma^{*} \geq 1/2$): then, ethics screening does not occur at all, since the principal switches from the first-best to the standard second-best contract as the probability of dishonesty increases.

Surpluses are affected in intuitive ways. Both the principal and the dishonest agent gain from the presence of honest agents if only if the probability of honesty rises above the threshold at which the principal drops the standard second-best contract. Whereas the surplus of the principal increases in a continuous manner (see Figure 1), it involves some discontinuities for the opportunistic agent whenever the principal switches to the first-best contract. By contrast, an honest is penalized by an increase in the likelihood of honesty. In the second-best contract he is treated as well as a dishonest agent, but otherwise he receives a smaller or no rent; if he receives a rent, which is the case in a revealing lies contract, it is decreasing in the likelihood of honesty.

Among the three contracts that may be optimal, only the revealing lies contract involves explicit ethics screening. It enables the principal to leave a smaller rent to the honest by letting the opportunistic misrepresent matches in equilibrium, and by exploiting the honest agent's distaste for misrepresenting the match should he choose the contract meant for the opportunistic agent. Clearly, if $\gamma > 1/2$, the principal would like to do the reverse, namely, give a smaller rent to the opportunistic. However, this is impossible since the opportunistic agent is unrestrained in his willingness to misrepresent the match. When the agent is more likely to be honest, the revealing-lies contract is preferred to the first-best contract only if a good match is likely enough, since only then does the benefit from "third-best" match screening for the opportunistic outweigh the cost of leaving a rent to the honest. As a result, two conditions must be met for the principal to resort to ethics screening by using the revealing-lies contract. First, honesty must be more probable than dishonesty. Second, the good match must be sufficiently likely.

Our results indicate that ruling out ethics screening altogether as in most of the previous literature implies a loss of generality. The revealing-lies contract allows the principal to screen among matches for dishonest agents while leaving a smaller rent to the honest than with the second-best contract. Nevertheless, this contract is a bit unsettling. In a workplace context, if an honest agent were to choose the mechanism meant for the dishonest, he would be asked to commit to producing more with a bad than with a good match. It is questionable whether an honest agent would feel compelled to commit to such an absurd scheme. But maybe more importantly, the ability of the principal to commit to this type of mechanism strongly depends upon her expectation that no honest agent would ever select it (after all, both the honest and the dishonest are indifferent between announcing their ethics truthfully or not); if she assigned a positive probability to such an event, she would have an incentive to renegotiate and offer the standard second-best mechanism instead.

This would be accepted by an honest agent since his surplus is larger than what he would obtain by telling the truth in the original mechanism. Our results show that if the option of the revealing lies contract were removed, the principal would be hurt only if a bad match is sufficiently unlikely. Then the principal would offer the standard second-best contract if the agent is likely opportunistic, and the first-best contract otherwise. In Alger and Renault (2003), this prediction emerges if an honest behavior is conditional on the fairness of the contract, although we do not impose any restrictions on the mechanisms the principal may use.

The above conclusions regarding ethics screening contrasts with those in Alger and Ma (2003), where ethics screening is optimal whenever honesty is sufficiently likely. In their model an insurer (the principal) contracts with a patient and a physician in the first period. A contract specifies the amount of treatment (the decision in our model) and transfers from the patient to the insurer as well as from the insurer to the physician. In the second period, the physician submits a claim to the insurer about the patient's health state; if the physician is opportunistic, he may collude with the patient against the insurer by submitting a false claim. The fact that there are three economic actors instead of two is inconsequential (since *ex post* the pair physician-patient chooses the claim that maximizes their joint surplus, one may lump them together into an entity called "agent"). Instead, there are two more fundamental differences that concur to explain why they obtain ethics screening whenever honesty is sufficiently likely whereas we find that a good match must also be sufficiently likely for that to occur.

In their setup an exogenous restriction on the contract effectively rules out the possibility of using a revealing-lies contract.[13] As a result, it is clearly the non-revealing lies contract which is optimal whenever γ is sufficiently small. By contrast with our model, the non-revealing lies contract involves ethics screening; furthermore, it never specifies the first-best decision, even for an honest physician. This is due to the insurance motive: the risk aversion

[13]The amount of treatment is assumed to be nil whenever the physician claims that the patient is healthy. Hence, if the physician and patient prefer to claim that the patient is ill when in fact he is healthy, they also prefer that claim when the patient is ill.

of the patient makes it optimal to reduce the difference between the allocations associated with an honest and an opportunistic physician.

A third difference is that, using our notation, in their model the principal's surplus π is directly affected by the match θ. Since in their model the insurance market is competitive, the principal's objective is to maximize the patient's expected utility. As a result, even if the physician always claims that the patient is ill, the objective function still mirrors the fact that the patient is healthy with some probability. We analyzed that case in a previous version of this paper (Alger and Renault, 2002). Qualitatively, the results are the same as here, except for the decision x specified for an opportunistic agent in the non-revealing lies contract: this decision would then be an increasing function of the probability that the match is good, and it would be equal to $\bar{x}^*$ (as here) only if that probability is zero. However, and this clearly points out the role played by risk aversion in Alger and Ma, the principal would still offer the first-best contract to an honest agent.

4 Concluding Remarks

As we saw in the introduction, modern psychology has taught us that it would not be appropriate to view honesty as an unrestrained desire to reveal hidden information. Rather, it should be dependent upon conditions under which the individual is led to choose whether to give up the benefits of an opportunistic behavior. Here we have focused on the role of loyalty in inducing honesty. *Ex ante* opportunistic behavior regarding ethics revelation enables an honest agent to sometimes garner a rent which would be denied to him if honesty was unconditional. It also implies that the principal must let the dishonest misrepresent matches if she wants to leave a smaller rent to an honest than to an opportunistic agent. As in Alger and Ma (2003) who use an information structure similar to ours, we find that the principal may choose to screen ethics. However, in contrast with their results, we find that ethics screening is not used if the match between the agent and the task is likely to be bad and, even when ethics screening is optimal for some parameter values, it may be dominated by offering the first-best contract independent of ethics if the probability of honesty is high.

The results of the standard model, where the agent is opportunistic with certainty, appear to be remarkably robust to the introduction of honest agents. In Alger and Renault (2003) we found that a sense of equity on the part of the agent had the same implication. In fact the outcome under both types of honesty would be very similar. The only difference hinges on the ability of the principal to use the revealing lies contract. It remains to be seen whether this robustness would be true in a model with a larger set of matches, and/or if honesty is triggered by something else than a sense of loyalty toward the principal or a concern for fairness. Also, we have exogenously assumed the nature of an honest behavior and the conditions under which it would emerge. Research on this topic should evolve towards endogenizing these aspects as being the result of dynamic social interactions, derived from basic underlying preferences.

Appendix

Proof of Lemma 1:

For the purpose of this proof, let y denote an allocation (x, t). Consider some contract which associates a direct mechanism to each message μ_1, through the function m. Now, consider another contract, in which the message space at date 1 is $\{h, d\}$. Let $\hat{k}$ denote the message sent at date 1 in this contract. To each $\hat{k}$, the contract associates a direct mechanism through the function $\tilde{m}$. Each mechanism $\tilde{m}(\hat{k})$ in turn defines an allocation $\tilde{y}$ to each message $\hat{\theta} \in \{\underline{\theta}, \bar{\theta}\}$ sent by the agent at date 2, through the function $\tilde{y}_{\tilde{m}}$. Let this function be such that:

$$\tilde{y}_{\tilde{m}(\hat{k})}(\hat{\theta}) = y_{m(\mu_1^*(\hat{k}))}(\hat{\theta}) \quad \forall \hat{k},\ \forall \hat{\theta}.$$

This implies that:

$$\hat{\theta}^*(\theta, \hat{k}, \tilde{m}(\hat{k})) = \hat{\theta}^*(\theta, \hat{k}, m(\mu_1^*(\hat{k}))) \quad \forall \hat{k},\ \forall \theta.$$

Then, we can show that truth-telling at date 1 is an equilibrium of the new mechanism, by applying the standard technique. *Q.E.D.*

Proof of Proposition 1:

Add constraints (1) and (2) to get:

$$\alpha[\bar{t}_h - v(\bar{x}_h, \bar{\theta})] + (1-\alpha)[\underline{t}_h - v(\underline{x}_h, \underline{\theta})] \geq$$

$$\alpha\, max\{\bar{t}_h - v(\bar{x}_h, \bar{\theta}), \underline{t}_h - v(\underline{x}_h, \bar{\theta})\} + (1-\alpha) max\{\underline{t}_h - v(\underline{x}_h, \underline{\theta}), \bar{t}_h - v(\bar{x}_h, \underline{\theta})\}$$

This implies that the following constraints must hold:

(12) $$\bar{t}_h - v(\bar{x}_h, \bar{\theta}) \geq \underline{t}_h - v(\underline{x}_h, \bar{\theta})$$

(13) $$\underline{t}_h - v(\underline{x}_h, \underline{\theta}) \geq \bar{t}_h - v(\bar{x}_h, \underline{\theta})$$

We can add these two constraints to the program, since they are implied by constraints (1) and (2). Let us now omit constraints (1) and (2), and find the solution to the relaxed program: maximize the expected utility subject to the constraints (3)-(13). By inspection, we note that we can split this program into the following two programs: for $k = d$ and $k = h$, respectively, find $m(k) = \{\bar{x}_k, \bar{t}_k, \underline{x}_k, \underline{t}_k\}$ so as to maximize $\alpha(\pi(\bar{x}_k, \bar{\theta}) - \bar{t}_k) + (1 - \alpha)(\pi(\underline{x}_k, \underline{\theta}) - \underline{t}_k)$, subject to the constraint (3) and the relevant part of constraints (4) and (5) for $k = d$, and subject to the constraints (12), (13) and the relevant part of constraints (4) and (5) for $k = h$. These programs are identical in structure, and therefore have the same solution, so that $m(h) = m(d)$. The omitted constraints (1) and (2) are satisfied. The solution is then trivial, since the maximization problem is well-known from standard principal-agent models. *Q.E.D.*

The following result is used in several subsequent proofs.

Claim 1 *In an optimal contract where the dishonest agent does not reveal θ truthfully, the mechanism for the honest is such that $\underline{x}_h \geq \underline{x}^*$ and $\bar{x}^h \leq \bar{x}^*$.*

Proof: First we prove that $\underline{x}_h \geq \underline{x}^*$. Suppose it is not the case. If the principal increases $\underline{x}_h$, this increases social surplus; thus, if the transfer $\underline{t}_h$ is increased so as to keep the honest agent's surplus unchanged, the principal's surplus increases as long as the dishonest still reveals ethics truthfully. We now check that this is the case. If the dishonest pretends he is honest, his surplus would be the same as in the original contract if he announces $\bar{\theta}$, and would be at most that of the original contract if he announces $\underline{\theta}$. Hence the corresponding expected surplus is lower than what he got by lying about ethics in the original contract. A symmetric argument can be used to show that $\bar{x}_h \leq \bar{x}^*$. *Q.E.D.*

Proof of Lemma 2

Possible equilibrium messages are:

E1: $\hat{\theta}^*(\underline{\theta}, d, m(d)) = \hat{\theta}^*(\bar{\theta}, d, m(d)) = \bar{\theta}$

E2: $\hat{\theta}^*(\underline{\theta}, d, m(d)) = \hat{\theta}^*(\bar{\theta}, d, m(d)) = \underline{\theta}$

E3: $\hat{\theta}^*(\underline{\theta}, d, m(d)) = \bar{\theta}$ and $\hat{\theta}^*(\bar{\theta}, d, m(d)) = \underline{\theta}$,

and possible out-of-equilibrium messages are:

O1: $\hat{\theta}^*(\underline{\theta}, d, m(h)) = \hat{\theta}^*(\bar{\theta}, d, m(h)) = \bar{\theta}$

O2: $\hat{\theta}^*(\underline{\theta}, d, m(h)) = \hat{\theta}^*(\bar{\theta}, d, m(h)) = \underline{\theta}$

O3: $\hat{\theta}^*(\underline{\theta}, d, m(h)) = \bar{\theta}$ and $\hat{\theta}^*(\bar{\theta}, d, m(h)) = \underline{\theta}$

O4: $\hat{\theta}^*(\underline{\theta}, d, m(h)) = \underline{\theta}$ and $\hat{\theta}^*(\bar{\theta}, d, m(h)) = \bar{\theta}$.

Eliminating O4: O4 requires standard incentive compatibility constraints in $m(h)$. Then, as in the proof to Proposition 1, the optimal contract is to offer the standard second-best mechanism independent of ethics. Implementing the same allocation while having the dishonest lie in equilibrium would require that he lies if $\theta = \underline{\theta}$ (and also if $\theta = \bar{\theta}$); but this contradicts our tie-breaking rule, since in the second-best contract the dishonest is indifferent between lying and telling the truth when $\theta = \underline{\theta}$ (see also our discussion following Lemma 2).

Eliminating O2: O2 requires $\underline{t}_h \geq v(\underline{x}_h, \bar{\theta})$, implying that both the honest and the dishonest get an expected rent which is at least $(1-\alpha)[v(\underline{x}_h, \bar{\theta}) - v(\underline{x}_h, \underline{\theta})]$. Since $\underline{x}_h \geq \underline{x}^* > \bar{x}^s$ (from Claim 1 above), this rent is strictly larger than with the standard second-best contract. The principal would therefore be better off with the standard second-best contract.

Eliminating O3: Assume **O3**. For $\hat{\theta}^*(\bar{\theta}, d, m(h)) = \underline{\theta}$ the following must hold:

$$\underline{t}_h - \bar{t}_h > v(\underline{x}_h, \bar{\theta}) - v(\bar{x}_h, \bar{\theta}). \tag{14}$$

Together with Claim 1, this implies $\underline{t}_h - \bar{t}_h > v(\underline{x}_h, \underline{\theta}) - v(\bar{x}_h, \underline{\theta})$, which implies that $\hat{\theta}^*(\underline{\theta}, d, m(h)) = \underline{\theta}$.

Eliminating E2: Consider a contract $(m(d), m(h))$ such that the equilibrium messages for the dishonest are $\hat{\theta}^*(\underline{\theta}, d, m(d)) = \hat{\theta}^*(\bar{\theta}, d, m(d)) = \underline{\theta}$. Since the dishonest does not send the message $\bar{\theta}$, one may without loss of generality assume that $\bar{t}_d = v(\bar{x}_d, \bar{\theta})$ and $\bar{x}_d < \bar{x}^*$.

For the dishonest of type $\bar{\theta}$ to announce $\underline{\theta}$, it must be that he gets a strictly positive rent $\underline{t}_d - v(\underline{x}_d, \bar{\theta}) > 0$ (the inequality is strict because of the tie-breaking rule). As a result, the honest also gets a strictly positive rent.

Now consider the alternative contract $(m(d)', m(h)')$ such that the dishonest's quantities are switched $\bar{x}'_d = \underline{x}_d$ and $\underline{x}'_d = \bar{x}_d$, while the honest's quantities remain at $\bar{x}'_h = \bar{x}_h$ and $\underline{x}'_h = \underline{x}_h$. Furthermore, set the transfers such that the surplus is zero for any agent announcing his true θ. A sufficient condition for the principal's surplus to be strictly larger with the alternative contract is that the dishonest announces d in the first period and $\bar{\theta}$ in the second, and the honest reveals the truth. Then the quantities are the same in all states of nature, and the agent's rent is smaller.

The ethics incentive constraint for the honest is satisfied. It remains to be proved that the dishonest's behavior is also unaffected. First, since $\underline{x}_h > \bar{x}_h$ he would still announce $\underline{\theta}$ in mechanism $m(h)'$. A sufficient condition for the dishonest's ethics incentive constraint to be satisfied is that the dishonest's surplus if he is type $\underline{\theta}$ is greater in mechanism $m(d)'$ than in mechanism $m(h)'$. This is true if $\bar{x}'_d \geq \bar{x}_h$, which we now prove. We first show that $\bar{x}'_d \geq \bar{x}^*$. Suppose that $\bar{x}'_d < \bar{x}^*$. Then the principal could increase $\bar{x}'_d$ while increasing $\bar{t}'_d$ so as to keep $\bar{t}'_d - v(\bar{x}'_d, \bar{\theta}) \equiv \bar{U}$ constant. This does not affect the honest's incentive constraint. Moreover, the dishonest's behavior is not affected. The principal's surplus may be written as $\pi(\bar{x}'_d) - v(\bar{x}'_d) - \bar{U}$; which is strictly increasing in $\bar{x}'_d$ for $\bar{x}'_d < \bar{x}^*$. Together with Claim 1, this implies $\bar{x}'_d \geq \bar{x}_h$. *Q.E.D.*

Proof of Proposition 2:

We ignore the constraints ensuring the assumed second-period equilibrium and out-of-equilibrium message structures (i.e., **E3** and **O1**); we will verify that they are satisfied by the candidate solution.

We now prove that the ethics incentive constraints are binding. First, note that if the honest pretends he is dishonest, he receives a strictly positive rent if his type is $\underline{\theta}$, since the allocation $(\underline{x}_d, \underline{t}_d)$ satisfies the individual rationality constraint for type $\bar{\theta}$. The ethics incentive constraint for the honest then implies that at least one of the honest agent's individual rationality constraints is not binding. Thus if the ethics incentive constraint for the honest were slack, the principal could increase her surplus by decreasing one of the transfers for the honest agent. A similar argument shows that the ethics incentive constraint for the dishonest must be binding as well.

We now show that the individual rationality constraints for type $\bar{\theta}$, (4) for $k = h$ and (9), are binding. Suppose first that it is not binding for the honest agent. Then the principal should decrease $\bar{t}_h$ and increase $\underline{t}_h$ so as to keep the expected transfer unchanged; this enables the principal to decrease one transfer for the dishonest, while keeping the dishonest's incentive constraint on ethics satisfied (since the dishonest would always announce $\bar{\theta}$ if he were to announce h in the first period).

To prove that the individual rationality constraint for the dishonest of type $\bar{\theta}$ is binding, one preliminary step is required: we prove that if the honest announces d and then is of type $\bar{\theta}$, he would not participate, that is, $\bar{t}_d - v(\bar{x}_d, \bar{\theta}) < 0$. To prove that, we first show that $\bar{x}_d > \bar{x}_h$. From Claim 1 a sufficient condition for this to be true is that $\bar{x}_d \geq \underline{x}^*$. Suppose to the contrary that $\bar{x}_d < \underline{x}^*$; then the principal could increase $\bar{x}_d$ as well as the transfer $\bar{t}_d$ so as to keep the surplus of the dishonest agent of type $\underline{\theta}$ unchanged. This increases the principal's surplus while relaxing the ethics incentive constraint for the honest.

We can now prove that $\bar{t}_d - v(\bar{x}_d, \bar{\theta}) < 0$. Suppose it was not true. Then the dishonest's surplus if he is type $\underline{\theta}$ is $\bar{t}_d - v(\bar{x}_d, \underline{\theta}) \geq v(\bar{x}_d, \bar{\theta}) - v(\bar{x}_d, \underline{\theta}) > v(\bar{x}_h, \bar{\theta}) - v(\bar{x}_h, \underline{\theta}) = \bar{t}_h - v(\bar{x}_h, \underline{\theta})$. Since the individual rationality constraint for type $\bar{\theta}$ in the contract for the honest is binding, this implies that the dishonest's ethics incentive constraint is slack, yielding a contradiction.

Now, since $\bar{t}_d - v(\bar{x}_d, \bar{\theta}) < 0$, the honest participates only if his type is $\underline{\theta}$. Then the argument used to show that the participation constraint for the honest of type $\bar{\theta}$ is binding can easily be adapted to show that the participation constraint for the dishonest of type $\bar{\theta}$ is also binding.

The transfers are therefore as in the proposition; replacing them in the objective function and taking first-order conditions yields the quantities in the proposition.

Now we verify that the assumed message structures obtain. Given the contract in the proposition, clearly a dishonest with type $\bar{\theta}$ would behave as assumed, whereas a dishonest with type $\underline{\theta}$ will have the assumed behavior if and only if $\bar{x}_h > \underline{x}_d$, which is true when $\gamma < \frac{1}{2}$. *Q.E.D.*

Proof of Proposition 3:

To begin, note that without loss of generality, we can set $\underline{t}_d - v(\underline{x}_d, \underline{\theta}) = 0$.

• **We prove that $\bar{x}_d \leq \bar{x}_h$.** From Claim 1, a sufficient condition is that $\bar{x}_d \geq \bar{x}^*$, which we now prove. Assume that $\bar{x}_d < \bar{x}^*$. Then one could increase $\bar{x}_d$ while increasing $\bar{t}_d$ so as to keep $\bar{t}_d - v(\bar{x}_d, \bar{\theta}) \equiv \hat{U}$ unchanged. The ethics incentive constraint for the honest is unaffected, while that of the dishonest is slackened. The surplus of the principal may be written $\pi(\bar{x}_d) - v(\bar{x}_d, \bar{\theta}) - \hat{U}$, which has increased.

• We can use the same argument as in the proof of Proposition 2 to show that the individual rationality constraint for the honest of type $\bar{\theta}$ (4) for $k = h$ is binding.

• **We show that the participation constraint for the dishonest of type $\bar{\theta}$ (4) for $k = d$ is binding.** Suppose that it was not binding so that $\bar{t}_d - v(\bar{x}_d, \bar{\theta}) > 0$. Then the ethics incentive constraint for the dishonest is not binding. To see this, note that the surplus of the dishonest if he announces d and is of type $\underline{\theta}$is $\bar{t}_d - v(\bar{x}_d, \underline{\theta}) > v(\bar{x}_d, \bar{\theta}) - v(\bar{x}_d, \underline{\theta}) \geq v(\bar{x}_h, \bar{\theta}) - v(\bar{x}_h, \underline{\theta}) = \bar{t}_h - v(\bar{x}_h, \underline{\theta})$, where the last term represents his surplus if he announces h in the first period. If he is of type $\bar{\theta}$he also gets a strictly larger rent by announcing d than by announcing h. But then the principal could decrease $\bar{t}_d$.

• **We show that the participation constraint for the honest of type $\underline{\theta}$ (5) for $k = h$ is binding.** Suppose that it was not binding. Together with the previous results, this would imply that the ethics incentive constraint for the honest is not binding. But then the principal may increase her surplus by decreasing $\underline{t}_h$.

• The transfers having been determined, the quantities in the proposition are obtained from the first-order conditions. *Q.E.D.*

Proof of Proposition 4:

Let Π^S, Π^*, and Π^R denote the principal's expected surplus given the optimal incentive-compatible, non-revealing lies, and revealing lies contracts, respectively. Further let the indices h and d indicate the principal's expected surplus conditional on the agent being honest and dishonest, respectively. Letting $S(x,\theta) = \pi(x) - v(x,\theta)$ denote total surplus at x given that the agent's type is θ, we have:

$$\Pi^S = \alpha S(\bar{x}^s, \bar{\theta}) + (1-\alpha) S(\underline{x}^*, \underline{\theta}) - (1-\alpha)[v(\bar{x}^s, \bar{\theta}) - v(\bar{x}^s, \underline{\theta})]$$

$$\begin{aligned} \Pi^* &= \gamma \left[\alpha S(\bar{x}^*, \bar{\theta}) + (1-\alpha) S(\bar{x}^*, \underline{\theta}) - (1-\alpha)[v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})]\right] \\ &\quad + (1-\gamma)[\alpha S(\bar{x}^*, \bar{\theta}) + (1-\alpha) S(\underline{x}^*, \underline{\theta})] \end{aligned}$$

$$\begin{aligned} \Pi^R &= \gamma \left[\alpha S(\underline{x}_d^R, \bar{\theta}) + (1-\alpha) S(\underline{x}^*, \underline{\theta}) - (1-\alpha)[v(\bar{x}_h^R, \bar{\theta}) - v(\bar{x}_h^R, \bar{\theta})]\right] \\ &\quad + (1-\gamma) \left[\alpha S(\bar{x}_h^R, \bar{\theta}) + (1-\alpha) S(\underline{x}^*, \underline{\theta}) - (1-\alpha)[v(\underline{x}_d^R, \bar{\theta}) - v(\underline{x}_d^R, \bar{\theta})]\right]. \end{aligned}$$

First, note that Π^S is independent of γ. Second, Π^* is linear in γ; the slope is equal to

$$\left[\alpha S(\bar{x}^*, \bar{\theta}) + (1-\alpha) S(\bar{x}^*, \underline{\theta}) - (1-\alpha)[v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})]\right] - [\alpha S(\bar{x}^*, \bar{\theta}) + (1-\alpha) S(\underline{x}^*, \underline{\theta})],$$

which is strictly negative. Finally consider Π^R. By the envelope theorem:

$$\begin{aligned} \frac{\partial \Pi^R}{\partial \gamma} &= (1-\alpha)\left[[v(\underline{x}_d^R, \bar{\theta}) - v(\underline{x}_d^R, \underline{\theta})] - [v(\bar{x}_h^R, \bar{\theta}) - v(\bar{x}_h^R, \underline{\theta})]\right] \\ &\quad + \alpha[S(\underline{x}_d^R, \bar{\theta}) - S(\bar{x}_h^R, \bar{\theta})]. \end{aligned}$$

Here we assume that $\gamma < \frac{1}{2}$, since otherwise the optimal revealing lies contract is not obtainable in equilibrium. Thus, $\underline{x}_d^R < \bar{x}_h^R < \bar{x}^*$; moreover, $\underline{x}_d^R$ and $\bar{x}_h^R$ tend to $\bar{x}^s$ as γ tends to $\frac{1}{2}$. Therefore the above expression is strictly negative for $\gamma < \frac{1}{2}$ and it tends to zero as γ tends to $\frac{1}{2}$. Furthermore, $\underline{x}_d^R$ is increasing in γ while $\bar{x}_h^R$ is decreasing in γ. Therefore $S(\underline{x}_d^R, \bar{\theta}) - S(\bar{x}_h^R, \bar{\theta})$ is strictly increasing in γ, and the term multiplying $1-\alpha$ is also increasing in γ. Therefore Π^R is a strictly decreasing and strictly convex function of γ.

A second step is to note a few relations between the expressions Π^S, Π^* and Π^R, and their implications. First, Π^R tends to Π^S as γ tends to $\frac{1}{2}$. Together with the above, this implies that $\Pi^R > \Pi^S$ for $\gamma < \frac{1}{2}$. Second, Π^N is equal to the first-best principal's surplus for $\gamma = 0$ and falls below the second-best expected surplus Π^T as γ tends to 1 (because the second-best mechanism is optimal for $\gamma = 1$). Together with the above, this means that for every α, there is a unique $\gamma^* \in (0, 1)$ such that $\Pi^* = \Pi^T$, above which $\Pi^* < \Pi^T$ and below which $\Pi^* > \Pi^T$. Finally, note that $\lim_{\gamma\to 0} \Pi^N = \lim_{\gamma\to 0} \Pi^R$.

We now show that the three cases of the proposition may arise. Since Π^* is decreasing and linear in γ, whereas Π^R is decreasing and convex in γ, a necessary and sufficient condition for the non-revealing lies contract to be optimal for some values of γ is that Π^* be less steep than Π^R for γ close to 0. Since we only need to compare $\frac{\partial \Pi^*}{\partial \gamma}$ and $\frac{\partial \Pi^R}{\partial \gamma}$ for γ close to 0, we compare $\frac{\partial \Pi^*}{\partial \gamma}$ to:

$$\lim_{\gamma\to 0} \frac{\partial \Pi^R}{\partial \gamma} = -(1-\alpha)[v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})] - \alpha S(\bar{x}^*, \bar{\theta}).$$

Π^* is less steep than Π^R for γ close to 0 if $\frac{\partial \Pi^*}{\partial \gamma} > \lim_{\gamma\to 0} \frac{\partial \Pi^R}{\partial \gamma}$, i.e., if:

$$\begin{aligned} &\alpha S(\bar{x}^*, \bar{\theta}) + (1-\alpha) S(\bar{x}^*, \underline{\theta}) + (1-\alpha)[v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})] \\ &-(1-\alpha) S(\underline{x}^*, \underline{\theta}) - (1-\alpha)[v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})] > 0 \end{aligned}$$

It is easy to verify that this holds for α close to 1, whereas it is violated for α close to 0. We now check whether the left-hand side is monotonic in α, so that there exists a unique threshold value $\bar{\alpha}$ such that the left-hand side is equal to zero. The inequality may be written:

$$S(\bar{x}^*, \bar{\theta}) - (1-\alpha) S(\underline{x}^*, \underline{\theta}) + (1-\alpha)[v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})] > 0.$$

The derivative of the left hand side with respect to α is positive if

$$S(\underline{x}^*, \underline{\theta}) - [v(\bar{x}^*, \bar{\theta}) - v(\bar{x}^*, \underline{\theta})] > 0,$$

or, equivalently, if:

$$S(\underline{x}^*, \underline{\theta}) + \pi(\bar{x}^*) - v(\bar{x}^*, \bar{\theta}) > \pi(\bar{x}^*) - v(\bar{x}^*, \underline{\theta}).$$

This is true since the right-hand side is smaller than $S(\underline{x}^*, \underline{\theta})$, by the definition of $\underline{x}^*$. *Q.E.D.*

References

Alger, I. and A. C.-t. Ma (2003), "Moral Hazard, Insurance, and Some Collusion," *Journal of Economic Behavior and Organization*, 50:225-247.

Alger, I. and R. Renault (2002), "Honest Agents and Equilibrium Lies," mimeo, Boston College and Université de Cergy-Pontoise.

Alger, I. and R. Renault (2003), "Screening Ethics when Honest Agents Care about Fairness," Boston College WP 489.

Deneckere, R. and S. Severinov (2001), "Mechanism Design and Communication Costs," mimeo University of Wisconsin.

Erard, B. and J.S. Feinstein (1994), "Honesty and Evasion in the Tax Compliance Game," *Rand Journal of Economics*, 25:1-19.

Forges, F. and F. Koessler (2003), "Communication Equilibria with Partially Verifiable Types," mimeo, Université de Cergy-Pontoise.

Green, J.R. and J.-J. Laffont (1986), "Partially Verifiable Information and Mechanism Deisgn," *Review of Economic Studies*, 53:447-456.

Hartshorne, H. and M. May (1928), *Studies in the Nature of Character*, New York: Macmillan..

Jaffee, D.M. and T. Russell (1976), "Imperfect Information, Uncertainty, and Credit Rationing," *Quarterly Journal of Economics*, 90:651-666.

Kofman, F. and J. Lawarrée (1996), "On the Optimality of Allowing Collusion," *Journal of Public Economics*, 61:383-407.

Meyer, J.P. and N.J. Allen (1991), "A Three-Component Conceptualization of Organizational Commitment," *Human Resource Management Review*, 1:61-89.

Meyer, J.P. and N.J. Allen (1997), *Commitment in the Workplace*, Thousand Oaks CA: SAGE Publications..

Nagin, D., J. Rebitzer, S. Sanders and L. Taylor (2002), "Monitoring, Motivation and Management: The Determinants of Opportunistic Behavior in a Field Experiment" *American Economic Review*, 92:850-73.

Picard, P. (1996), "Auditing Claims in the Insurance Market with Fraud: The Credibility Issue," *Journal of Public Economics*, 63:27-56.

Spicer, M.W. and L.A. Becker (1980), "Fiscal Inequity and Tax Evasion: An Experimental Approach," *National Taxation Journal*, 33:171-175.

Terris, W. and J. Jones (1982), "Psychological Factors Related to Employees' Theft in the Convenience Store Industry," *Psychological Reports*, 51:1219-1238.

Tirole, J. (1992), "Collusion and the Theory of Organizations," in Laffont, J.-J., ed., *Advances in Economic Theory: Proceedings of the Sixth World Congress of the Econo-*

metric Society, Cambridge: Cambridge University Press.

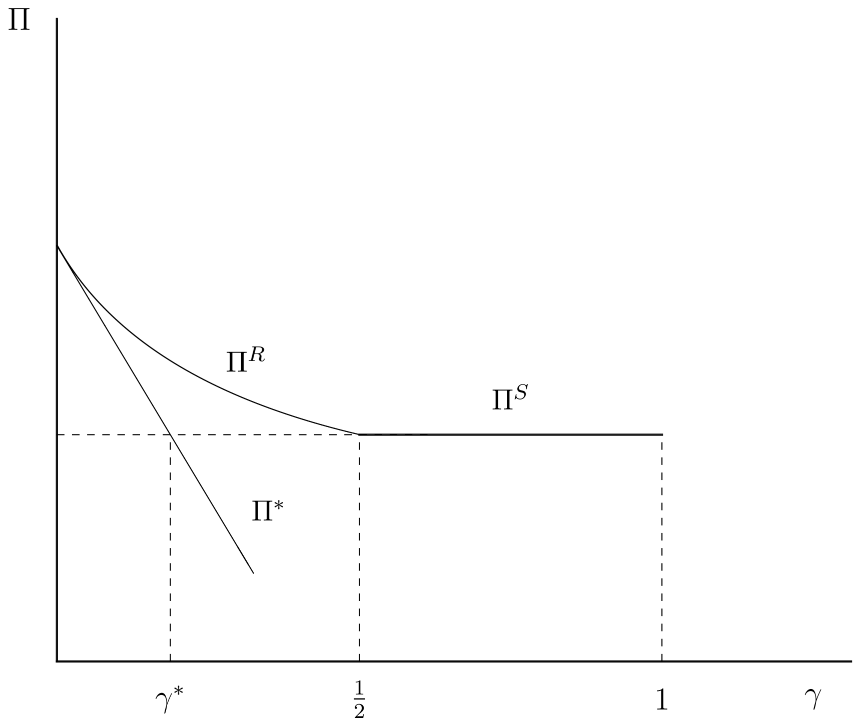

Figure 1 a. Case 1 in Proposition 4.

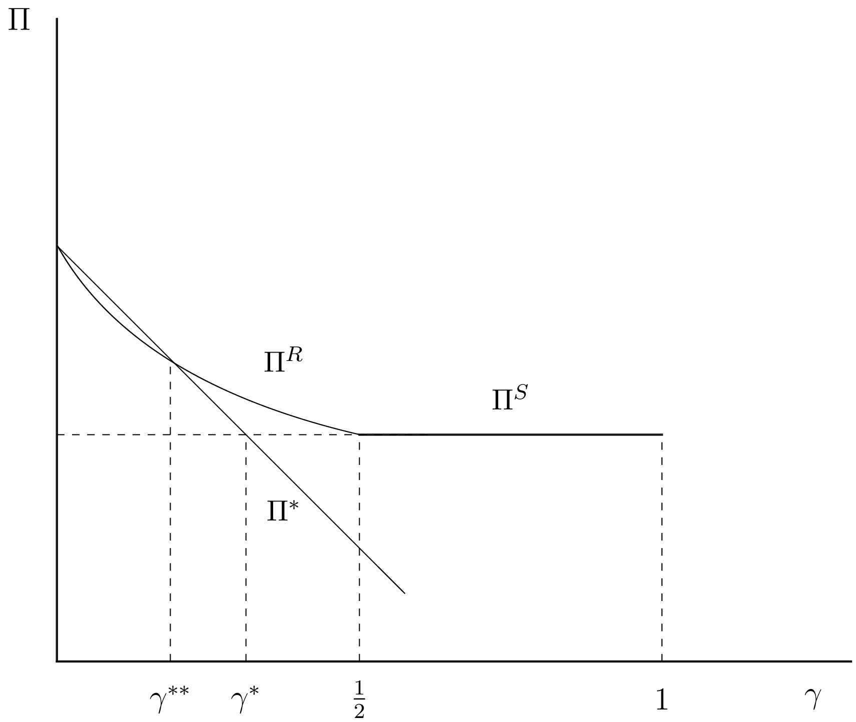

Figure 1 b. Case 2 (i) in Proposition 4.

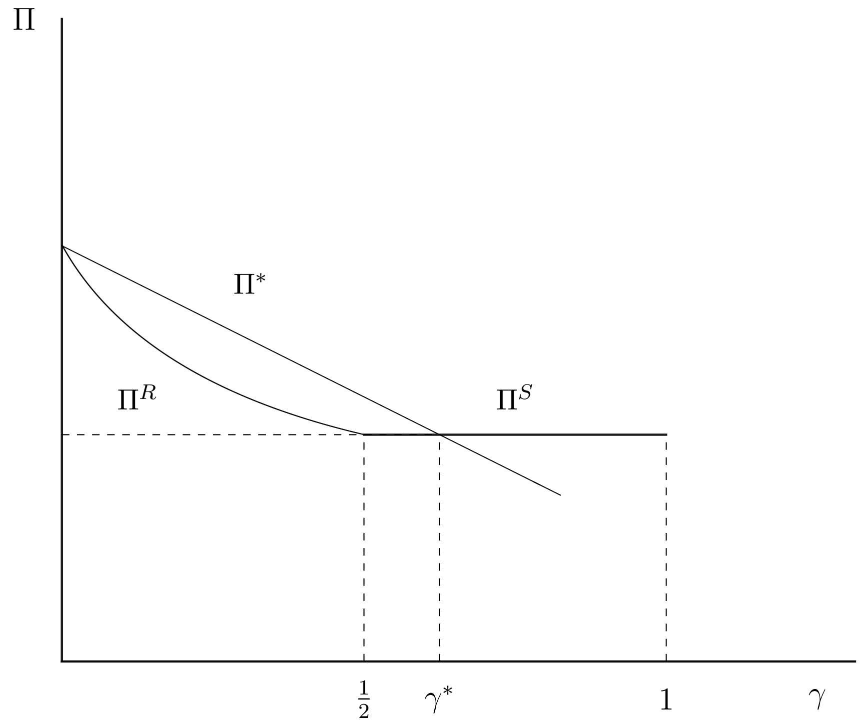

Figure 1 c. Case 2 (*ii*) in Proposition 4.

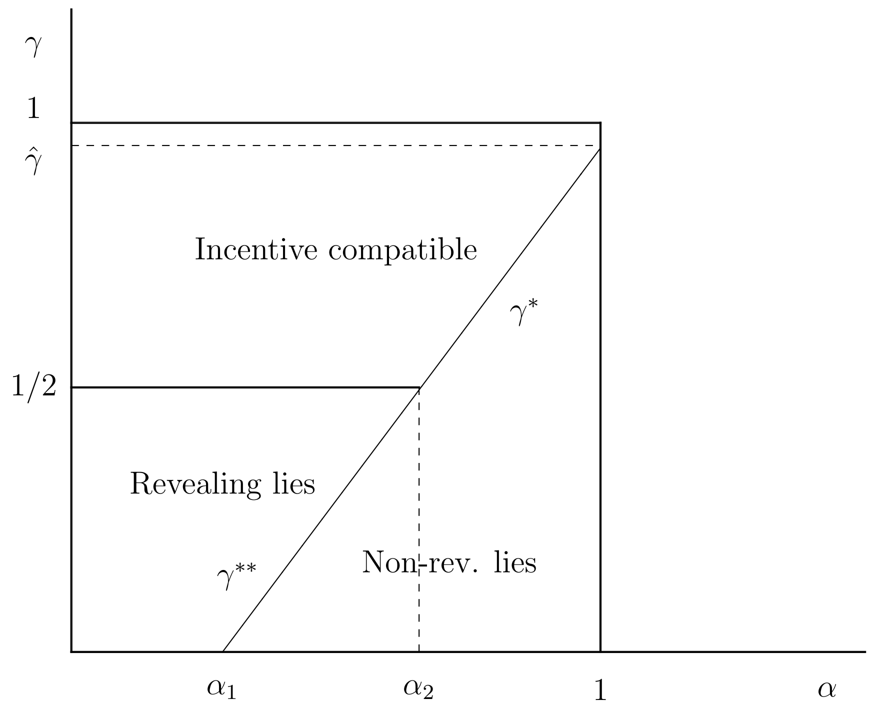

Figure 2. Visualization of Proposition 4.