Prosocial Disclosure and Contracts*

Mohammad Lashkarbolookie†

Latest Version

November 9, 2025

Abstract

This paper studies how regulations mandating the disclosure of prosocial outcomes (e.g., ESG performance) affect prosocial contracts (e.g., sustainability-linked loans). I develop a multitasking principal–agent model with limited liability and private agent types. The agent exerts costly effort on two tasks: one yielding an unverifiable outcome and another producing an outcome that can be verifiably disclosed at a cost. The agent's private type captures intrinsic utility over the outcomes. While mandatory disclosure can provide information that enables new contracts, it may affect the efficiency of contracts that would otherwise arise under a voluntary regime. Two mechanisms drive this effect. First, the principal may offer stronger incentives under voluntary disclosure to induce the agent's participation. Second, voluntary disclosure can serve a screening role, allowing the principal to identify intrinsically motivated agents. When the voluntary regime results in non-disclosure (full disclosure), mandating disclosure enhances (reduces) contracting efficiency. In cases where voluntary disclosure is partial, mandatory disclosure reduces welfare when prosocial and financial objectives are strongly complementary, but can improve welfare when principal free-riding weakens incentives for motivated agents.

KEYWORDS: Non-Financial Disclosure, Sustainability-linked Contracts, Disclosure Regulation, Multi-tasking Moral Hazard, Adverse Selection

JEL CLASSIFICATION: D82, D86, L50, Q50

*I am grateful to Viral Acharya, Gorkem Celik, Anastasios Dosis, François Le Grand, Andras Niedermayer, Natacha Raffin, Regis Renault, Catherine Roux, Lutz Sager, Wilfried Sand-Zantman, Petros Sekeris, and Péter Vida for their valuable comments and suggestions, which significantly improved this paper. I also thank participants at the 3rd AICC Conference 2025, the THEMA and ESSEC internal seminars, as well as the organizers and committees of the ASSET 2025, EWMES 2025, and MES 2026 conferences for including this paper in their programs.

†ESSEC Business School and THEMA (PhD Candidate), Email: mohammad.lashkarbolookie@essec.edu

1 Introduction

In recent decades, growing emphasis has been placed on the non-financial consequences of firms' activities such as environmental impact, labor practices, and broader social responsibility. As consumers and investors increasingly incorporate these factors into their decisions, firms have adopted various voluntary disclosure mechanisms to report their prosocial performance.[1] To strengthen transparency, many governments have also introduced mandatory disclosure regulations.[2]

Prosocial disclosures play a critical role in facilitating *prosocial contracts*; contractual arrangements that explicitly tie rewards to an agent's prosocial outcomes. Notable examples include sustainability-linked loans (SLLs) and sustainability-linked bonds (SLBs)[3], which adjust financial terms based on firms' sustainability performance indicators, as well as ESG-based executive pay for managers that tie their compensation to the ESG performance of the firm. The focus of this paper is on the disclosure of such contractible information about the prosocial performance of firms and the contracts that explicitly incentivize prosocial effort by providing financial incentives.

The central question of this paper is how mandatory non-financial disclosure regulations affect prosocial contracting; whether such regulations can promote and strengthen prosocial contracting and improve welfare relative to a voluntary disclosure regime.

To address this question, I develop a theoretical framework for prosocial disclosure and contracting in an environment characterized by multiple hidden actions and private agent types. Specifically, I study a multitask principal–agent model with limited liability, in which the principal's payoff from a delegated project depends on two stochastic outcomes whose realization probabilities are determined by the agent's efforts. The outcomes differ in verifiability: one—the prosocial outcome—can be verifiably disclosed at

[1] Examples include Corporate Social Responsibility (CSR) reports, Environmental, Social, and Governance (ESG) metrics, and frameworks such as the Global Reporting Initiative (GRI), the Sustainability Accounting Standards Board (SASB), and the Task Force on Climate-related Financial Disclosures (TCFD).

[2] Notable regulations include the Corporate Sustainability Reporting Directive (CSRD) and the Sustainable Finance Disclosure Regulation (SFDR) in the European Union, and mandatory TCFD reporting in the United Kingdom.

[3] According to Bloomberg, total sustainable debt issuance reached 1,740 billion dollars in 2024, representing a 12 percent increase compared to 1,547 billion dollars in 2023. https://www.bloomberg.com/professional/insights/sustainable-finance

a cost, while the other remains unverifiable. The agent may also derive intrinsic utility from exerting effort, generating a baseline level of prosocial activity even in the absence of contractual incentives. When disclosure occurs, the principal can condition rewards on the verifiable outcome, thereby strengthening incentives for that task. I characterize equilibrium contracts under both voluntary and mandatory disclosure regimes and compare their efficiency and welfare implications.

There are several empirical observations that motivate the question and methodology of this paper. First, prosocial contracting with strong incentives does not necessarily emerge even when prosocial outcomes are verifiably disclosed.[4] The emergence and design of prosocial contracts are sensitive to agent characteristics, particularly their intrinsic prosocial commitment in the absence of external incentives.[5] These observations cast doubt on the effectiveness of mandatory non-financial disclosure regulations in promoting prosocial contracting and motivate a theoretical analysis that explicitly accounts for the intrinsic and possibly hidden prosocial motivation of agents. Second, the broader effects of prosocial contracts beyond their immediate contractual outcomes remain mixed, both in terms of financial performance[6] and non-contractible prosocial efforts.[7] These findings point to the relevance of a multitasking framework for studying prosocial contracting.

My model—a multitasking moral hazard framework with private agent types—captures key features of such environments: reliance on verifiable outcomes for incentivizing effort toward prosocial goals, sensitivity of contracts to agents' intrinsic values which might

[4]For example, Auzepy, Bannier, and Martin (2023) find that sustainability-linked loans (SLLs) often rely only partially on performance indicators that generate credible sustainability incentives.

[5]Loumioti and Serafeim (2022) show that SLLs with rewards or penalties tied to ESG performance are more commonly issued to low-ESG-risk borrowers, whereas high-ESG-risk borrowers are less likely to receive loans with explicit ESG-related incentives. Similarly, Kim et al. (2021) demonstrate that while high-transparency firms use SLLs to advance genuine ESG goals, low-transparency firms employ SLLs to signal responsibility while actually worsening ESG performance post-issuance.

[6]For instance, Gladilina et al. (2024) show that the adaptation of ESG contracting can enhance ESG performance and foster profit growth and competitiveness, suggesting positive spillovers on financial outcomes. In contrast, Cohen et al. (2023) find that linking executive pay to ESG performance improves ESG outcomes but has no significant effect on financial performance.

[7]For example, Basu et al. (2022) show that banks with higher ESG ratings issue fewer mortgages in low-income areas, indicating adverse effects on other prosocial efforts. Similarly, Zhang (2022) and Yang et al. (2020) provide evidence that environmental regulations can unintentionally encourage firms to prioritize optics over substance, diverting effort away from effective prosocial action toward more symbolic measures.

be unobservable to stakeholders, and the potential for prosocial contracts to influence broader non-contractible performance.

In a multi-tasking environment with limited liability, the intrinsic utility of the agent over the outcomes shapes the principal's optimal contract design in two ways. First, the agent's baseline effort creates a *free-riding* effect: the principal can benefit from the agent's effort in autarky without sharing her gain from this efforts with the agent. Second, tying rewards to the verifiable outcome can have *cross-task* effects: it may either encourage or crowd out effort on the unverifiable task, depending on whether the tasks are complements or substitutes for the agent. Overall, the intrinsic utility of the agent can make prosocial contracting either more or less profitable for the principal, depending on the balance between the free-riding and cross-task effects.

I first consider the case of observable types. When the principal knows the agent's intrinsic utility over the outcomes, she can induce disclosure of the verifiable outcome by offering a contingent contract that rewards the agent when the desirable outcome is realized. For the agent to participate in prosocial disclosure and contracting, his gain from the contract must at least cover the disclosure cost. The principal is willing to provide such a contract as long as it remains profitable for her. Importantly, when the disclosure cost increases to the point where the principal's optimal contract no longer induces participation, she may offer a contract with a higher incentive term, raising agent's share of the surplus by reducing her own share, to secure the agent's disclosure. Thus, with observable types, prosocial disclosure and contracting arises whenever the principal can induce agent's participation while having a positive incremental profit.

In a setting with observable types, when a voluntary disclosure regime fails to generate disclosure and contracting, a mandatory disclosure regulation can enable contracting by compelling the agent to provide verifiable information. Such a contract increases the agent's effort on the verifiable task and thereby generates social surplus, but it reduces net welfare since the surplus created is insufficient to offset the disclosure cost. By contrast, when disclosure and contracting arise voluntarily, mandating disclosure can lower both contract efficiency and net welfare. In this case, because disclosure occurs irrespective of contractual arrangements, the principal no longer needs to raise the agent's payoff to

secure participation and may offer a contract with a lower incentive term and efficiency. Thus, whenever voluntary disclosure already supports contracting, a disclosure mandate can harm contract efficiency and welfare by relaxing the participation constraint of the agent.

I next turn to the case of private agent types. I consider two types of agents: a good type, who derives intrinsic utility from the realization of the outcomes and exerts positive effort on both tasks in autarky, and a bad type, who has no intrinsic utility over the outcomes. As discussed above, the intrinsic motivation of the good type can make prosocial contracting with him more or less profitable for the principal relative to contracting with the bad type, depending on the strength of the free-riding and cross-task effects. Accordingly, I examine two forms of adverse selection. In the first, prosocial contracting with the good type is more profitable for the principal, particularly when the tasks exhibit strong complementarity for both parties. In the second, contracting with the bad type is more profitable, due to the free-riding effect and possibly a negative cross-task effect that crowds out the good type's autarky effort on the task with unverifiable outcome.

A key feature of the model is that, in both forms of adverse selection described above, the good-type agent exerts higher effort on the task with the verifiable outcome, for any given incentive offered by the principal. As a result, three classes of equilibria can arise: (i) full disclosure, where both types disclose and contract; (ii) partial disclosure, where only the good type discloses and contracts; and (iii) non-disclosure, where prosocial disclosure and contracting break down entirely.

Under both forms of adverse selection, when the voluntary regime results in full disclosure, mandatory disclosure can reduce contracting efficiency and welfare. Similar to the case of observable types, under a disclosure mandate the principal no longer needs to increase the agent's share of the surplus to induce participation, and may therefore offer lower contract terms, leading to a welfare loss. By contrast, the effect of mandatory disclosure when the voluntary regime results in partial or non-disclosure equilibria differs across the two specifications of agent types.

When prosocial and financial objective exhibit strong complementarity for both parties and contracting with the good-type is more profitable than with the bad-type, vol-

untary disclosure of the verifiable outcome can have a socially desirable screening role. In this case, the principal may optimally exclude the bad-type agent from prosocial contracting in order to offer stronger incentives to the good-type agent. In this setting, when a partial disclosure equilibrium emerges, mandating disclosure of the verifiable outcome has two effects: first, it enables contracting with the bad-type agent, and second, it negatively affects the contract received by the good-type agent by eliminating the screening role of voluntary disclosure. In such an environment, mandatory disclosure can lower contracting efficiency compared to the voluntary regime, particularly when the negative effect on the contract offered to the good-type agent dominates the welfare gain of enabled contract with the bad-type. This constitutes one of the main results of this paper: in environments with strong complementarity between prosocial and financial goals for the principal and the agent, a mandatory disclosure regulation can reduce contracting efficiency and welfare by eliminating the screening role of disclosure.

Next, I consider environments in which prosocial contracting is more profitable for the principal with the bad-type agent, which occurs when strong free-riding or negative cross-task effects limit the incentives the principal is willing to offer the intrinsically motivated agent. Under this form of adverse selection, when the voluntary regime leads to partial or non-disclosure, mandatory disclosure can enhance contracting efficiency and welfare, particularly when free-riding severely constrains the financial incentives the principal is willing to provide to the good-type agent. In this case, mandating disclosure not only enables contracting with the bad-type agent but can also improve the contract offered to the good-type agent in a pooling equilibrium. This constitutes another main result of the paper: in the presence of strong free-riding in contracting with the good-type agent, mandatory disclosure can yield higher contracting efficiency and welfare compared to the voluntary regime.

This paper develops a theoretical framework to study how prosocial disclosure and contracting interact with mandatory disclosure regulation in a multitasking principal–agent setting with private agent types. The analysis identifies three main effects of mandatory disclosure on prosocial contracting: (i) it provides information that enables prosocial contracts, thereby improving welfare; (ii) it can weaken the contracts that would otherwise

arise under a voluntary regime by relaxing agents' participation constraints, reducing contracting efficiency; and (iii) it can eliminate the screening role of voluntary disclosure when the principal excludes non-committed agent types from contracting. This, in turn, may reduce welfare in the presence of strong task complementarity or increase welfare when strong free-riding effects dominate.

From a policy standpoint, the analysis underscores that mandatory disclosure is not a universally welfare-enhancing instrument. Its effectiveness depends critically on the prevailing disclosure environment and the underlying heterogeneity of agents. In environments where some agents already engage in prosocial disclosure and contracting, imposing disclosure requirements may yield ambiguous effects: it can diminish contracting efficiency in sectors characterized by strong complementarity between prosocial and financial objectives, yet enhance prosocial contracting and welfare in environments where pronounced free-riding weakens voluntary incentives.

1.1 Related Works

While this paper is, to the best of my knowledge, the first to model the contractual implications of prosocial disclosure and analyze the impact of mandatory disclosure regulation on prosocial contracting, it builds upon and contributes to a broader literature examining the effects of regulations mandating the disclosure of prosocial performance. For a comprehensive review of such studies, see Moharram et al. (2024).

A closely related study is Aghamolla and An (2023), which examines the impact of mandatory disclosure regulation in the context of a firm manager's interaction with shareholders. In their model, a manager—who seeks to maximize shareholder surplus through investment decisions—receives two private signals: one regarding the profitability of a project and the other its ESG quality. The manager then decides whether to disclose these signals to shareholders who hold heterogeneous ESG preferences. They show that while mandatory disclosure improves the firm's ESG outcomes, it may reduce overall welfare.

Another closely related paper is Goldstein et al. (2022), which analyzes the effect of ESG disclosure on information aggregation in financial markets. They develop a rational expectations equilibrium model in which traditional and green investors are informed

about both financial and ESG risks, but differ in their preferences over these dimensions. The paper shows that improving the quality of ESG information can reduce the informativeness of market prices regarding financial payoffs and thereby raise firms' cost of capital. In this setting, a mandatory disclosure regulation that enhances ESG information quality can act as a Pigouvian tax, promoting green investment at the expense of financial efficiency.

The results in this paper complement the findings of these two studies by showing that, in a setting characterized by hidden actions and private agent types, mandatory disclosure of verifiable information can not only reduce overall welfare, but also diminish prosocial performance itself.

There are also other theoretical studies that highlight the potential adverse effects of mandatory disclosure regulations on the overall quality of public information. For instance, Bond and Zeng (2022) analyze the emergence of non-disclosure equilibria in settings where senders are uncertain about the preferences of their audience. They show that a regulation mandating minimum ESG disclosure levels can backfire by encouraging firms to standardize their disclosures or reduce them to the mandated minimum, thereby weakening the informativeness of the disclosed data. Also, Weksler and Zik (2023) study disclosure in markets for ratings, where issuers are initially endowed with homogeneous soft information about their values before deciding whether to pay for formal ratings. Their results show that mandating the disclosure of ratings can reduce the overall information available to the public.

This paper also contributes to the growing literature on the optimal design of sustainability-linked contracts. Bonham and Riggs-Cragun (2025) study the role of contractual and regulatory incentives in shaping financial and ESG activities in a multitasking environment. They show that ESG contracting can enhance green innovation and improve performance measurement quality, while also amplifying risk in green firms and reducing risk in brown firms. Barbalau and Zeni (2022) develop a theory of optimal security design for green investment in the presence of greenwashing, characterizing conditions under which outcome-based securities (such as sustainability-linked loans) or project-based contracts (such as green bonds) are optimal. This paper complements these studies by in-

troducing agents' intrinsic values, modeled as private types, into the analysis of prosocial contracting.

Moreover, this study relates to the literature on socially responsible investment and its real and financial impacts, including works such as Pedersen, Fitzgibbons, and Pomorski (2021), Albuquerque, Koskinen, and Zhang (2019), Heinkel, Kraus, and Zechner (2001), and Oehmke and Opp (2024). It contributes to the broader literature on prosocial incentives and their influence on organizational behavior. Notably, studies like Akerlof and Kranton (2005), Bénabou and Tirole (2010), Besley and Ghatak (2017), and Bénabou and Tirole (2006) which investigate the interplay between social and monetary motivations in economic interactions.

The remainder of the paper is organized as follows. Section 2 introduces the model. Section 3 characterizes the equilibrium contracts under both mandatory and voluntary disclosure regimes. Section 4 analyzes welfare outcomes of prosocial disclosure and contracting and examines the impact of disclosure regulation. Section 5 concludes with a discussion of the policy implications.

2 The Model

There is a project that generates benefits for a principal (e.g., an investor or lender), and the tasks necessary to implement it are delegated to an agent (e.g., a manager or borrower). The principal is unable to perform these tasks directly, making delegation essential. Both the principal and the agent are risk-neutral and do not discount future pay-offs.

The Principal. The value of the project to the principal depends on two independent stochastic outcomes: a potentially verifiable outcome, $\mathcal{V}$, and a non-verifiable outcome, $\mathcal{N}$. Each outcome is binary[8] as depicted in Figure 1. The probability of each state depends on the effort exerted by the agent. Let $e = (e_v, e_n)$ denote the vector of the agent's efforts along the two dimensions, where $(e_v, e_n) \in [0, 1]^2$. For simplicity, I assume that

[8]The assumption of binary outcomes is made purely for simplicity. As long as both the principal and the agent are risk-neutral, extending the model to a continuous outcome space does not qualitatively affect the results.

the probability of success for each outcome equals the agent's effort in the corresponding dimension, i.e. $\Pr(\mathcal{V}=1)=e_v$ and $\Pr(\mathcal{N}=1)=e_n$[9].

I denote the value of the project to the principal by Y_{vn} when both outcomes are successfully realized, by Y_v when only the verifiable outcome is successful, and by Y_n when only the non-verifiable outcome is successful, normalizing it to zero when neither outcome is realized. The expected value of the project for the principal, as a function of the agent's effort, is then given by:

$$Y(e_v,e_n)\equiv e_v e_h Y_{vn}+e_v(1-e_n)Y_v+(1-e_v)e_n Y_n$$

Figure 1: Outcomes and Tasks

The verifiable outcome is a socially valuable consequence of the agent's activities—such as reduced carbon emissions, improved labor practices, or greater gender equality. I interpret the unverifiable outcome as a cash-flow-related consequence that cannot be observed or verified. For instance, in the case of ESG-based executive compensation, the unverifiable outcome may correspond to the firm's long-term cash flows that can not be observed in the short run. In the context of a sustainability-linked loan, it can be interpreted as the expected repayment of an unsecured loan.[10] Crucially, this outcome is assumed to be stochastic and non-contractible.[11]

The randomness of outcomes may stem from the agent's limited control over the final consequences of their actions or from the inherently stochastic nature of the environment.

[9] Introducing a concave mapping from effort to probabilities would not qualitatively affect the results.

[10] Alternatively, the unverifiable outcome may capture prosocial effects that are not reflected in measurable indicators such as ESG scores.

[11] The model abstracts from observable and contractible financial outcomes. In this sense, one may think of the principal as already contracting on verifiable financial metrics, while considering whether to add a prosocial clause to the existing contract, taking into account its potential effect on non-contractible financial performance, or prosocial objectives that are not included in standard metrics.

For example, a firm's long-term cash flow may be influenced by factors beyond its control, such as market competition or demand shocks. Similarly, a firm's prosocial outcomes, such as environmental performance, can be affected by external factors like climate risks, energy prices, or technological shocks. As a result, while effort increases the likelihood of favourable outcomes, it does not fully determine the level of final performance.

The Agent. The effort exerted by the agent is costly. This cost may represent either non-pecuniary disutility or pecuniary expenses associated with performing the tasks. Let $C(e)$ denote the agent's cost of exerting the effort vector e. To simplify the analysis, I adopt the following functional form for the cost of effort[12]:

$$C(e,h) = \frac{1}{2}e^2 + seh + \frac{1}{2}h^2.$$

Parameter s captures the interaction between the two tasks in the agent's cost of effort. I assume that $C(e)$ is convex and that the tasks are cost-substitutes, i.e., $s \in [0,1)$. This implies that exerting effort on one task increases the marginal cost of effort on the other. Cost-substitutability may arise from the agent's limited time and attention, or from the inherent difficulty of simultaneously achieving higher prosocial and financial performance.

I assume that both the agent's effort and the associated cost are not verifiable. The model thus focuses on activities subject to moral hazard, abstracting from perfectly contractible actions. Moreover, the agent is subject to limited liability: the agent's gain from any contract with the principal must remain non-negative in all states of nature, regardless of the realized outcome. This implies that, while the principal must delegate the task, monetary transfers from the agent to the principal, whether to capture project value or as penalties for poor performance, are not feasible. This setting captures relationships such as investor–manager or lender–borrower, where the agent's unobservable effort influences the principal's payoff, and the stochastic nature of outcomes prevents verification of the agent's true actions.

[12]The assumption of a quadratic cost of effort simplifies the characterization of equilibria under adverse selection but is not essential for the results. The same findings obtain under a broad class of cost functions. In Section 3.2, I characterize the sufficient conditions on a general cost function that yield the same results as those derived under the quadratic specification.

A key feature of the model is that the agent may derive intrinsic[13] benefit from the realization of the outcomes. In the case of a prosocial outcome, this benefit may stem from a warm-glow effect or reputational gains when the desired outcome is achieved. For a cash-flow-related outcome, the agent may derive non-pecuniary satisfaction (for example from the psychological reward of leading a successful project), or gain creditworthiness from project success.

Importantly, I assume heterogeneity among agents who differ in their intrinsic interest in the outcomes. For simplicity, suppose the agent can be one of two types, $i \in \{g, b\}$, with the *good-type* agent occurring with probability $m^g = \lambda$ and the *bad-type* agent occurring with probability $m^b = 1 - \lambda$. I assume that the good-type agent intrinsically benefit from the realization of the outcomes, while the bad-type agent does not have such a benefit. I denote the good-type agent's intrinsic benefit by B_{vn} when both outcomes are realized successfully, by B_v when only the verifiable outcome is successful, and by B_n when only the non-verifiable outcome is successful, normalizing it to zero when neither outcome is realized. The expected intrinsic benefit of the agents from the project writes:

$$B(e_v, e_n) \equiv B^g(e_v, e_n) = e_v e_h B_{vn} + e_v(1 - e_n)B_v + (1 - e_v)e_n B_n,$$

$$B^b(e_v, e_n) = 0.$$

Note that while the tasks are cost-substitutes for the agent, they may be complements in the benefit function of the good-type agent if $B_{12}(e) > 0$. Let $U^i(e_v, e_n) \equiv B^i(e) - C(e)$ denote an agent's net utility from exerting effort. For the good-type agent, the cross-partial derivative $U^g_{12}(e) = B^g_{12}(e) - s$ can therefore be either positive or negative, depending on the relative strength of cost substitutability and benefit complementarity.

Information. I consider prosocial disclosure as a mechanism that enables contracting on future outcomes through a verifiable metric. Specifically, the agent can commit to disclosing verifiable information about the prosocial outcome $\mathcal{V}$, while the outcome $\mathcal{N}$ and the agent's effort levels are assumed to be non-verifiable. A central assumption is the

[13] I use the term *intrinsic* to refer to motivations that are distinct from explicit contractual incentives that provide monetary rewards for prosocial performance. In this context, reputational gains are also considered intrinsic motivations.

existence of a reliable and certifiable metric for $\mathcal{V}$, potentially verified by an independent third party. Disclosure of this outcome thus constitutes hard, credible information that can support contractual enforcement. Key performance indicators in sustainability-linked loans or ESG ratings are examples of such metrics.

The disclosure is costly to the agent. The cost may reflect the agent's effort to monitor the prosocial consequences of their activities or the payment required to a third-party agency that certifies the disclosed information. I assume the cost of disclosure, denoted as $f \geq 0$, to be identical for all types of agents and independent of the effort level and the realization of the outcome.

Moreover, I assume the principal may not be able to identify the agent's type ex ante. Since types differ only in the utility they derive from outcome realizations, it is reasonable to assume that this distinction is not observable. For example, in ESG lending, a lender may be unable to assess a borrower's true commitment to financial or prosocial goals. Likewise, investors often cannot determine whether managers are genuinely aligned with their objectives or merely signalling responsibility to appeal to stakeholders.

While each agent's type is private information, I assume that both the principal's and agents' objective functions, as well as the prior distribution of agent types, are common knowledge.

The game. I model the voluntary disclosure of the verifiable outcome and the associated contracting as a multi-stage game with the following sequence:

- **Stage 1:** Nature chooses the agent's types.
- **Stage 2:** The principal offers contracts contingent on agent's disclosure decision.
- **Stage 3:** The agent chooses one contract, if any, committing to a disclosure decision, and decides on the level of efforts.
- **Stage 4:** The outcomes are realized, and the terms of the contracts are executed.

In the first stage of the game, agent's type is chosen randomly by nature. In the second stage, the principal offers contracts that specify transfers based on the agent's disclosure

decision. The contract in case of disclosure commitment can condition transfers to the realization of the disclosed outcome. In the third stage, the agent selects a contract, deciding to disclose or not, and then chooses effort levels for the tasks. In the final stage, outcomes are realized, information is disclosed, and transfers are made according to the terms of the chosen contract.

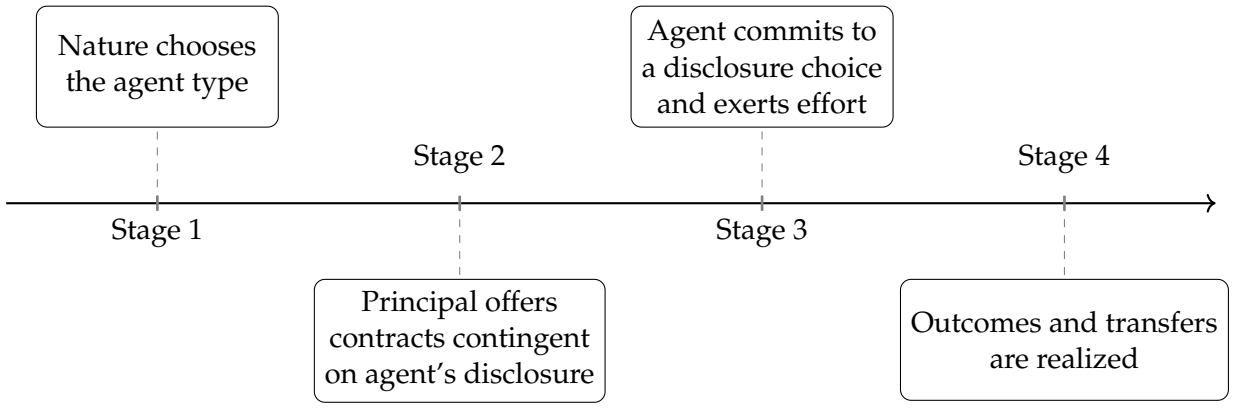


Figure 2: Timing of the Voluntary Disclosure Game

Under a mandatory disclosure regulation, the timing of the game remains the same, but agents are assumed to commit to disclosure with certainty before the game begins. Consequently, with a mandatory disclosure regulation, contracts are no longer contingent on the agent's disclosure decision.

I characterize the perfect Bayesian equilibrium of the game. An equilibrium consists of: a set of disclosure-contingent contracts X^i offered by the principal and accepted by the agent; the agent's disclosure decision and effort levels, e^i; and a consistent belief system for the principal.

3 Equilibrium Characterization

In this section, I analyze the equilibrium of the game, shaped by three informational frictions. First, the principal faces a moral hazard problem due to the agent's unobservable efforts and the asymmetry in outcome verifiability. Second, private information about the agent's type further restricts the design of optimal contracts. Third, the cost of dis-

closure influences the agent's incentive to engage in prosocial contracting. To clarify the equilibrium structure, I begin by analyzing the case with observable types in Section 3.1, isolating the moral hazard problem. I then characterize the equilibrium with private types in Section 3.2.

3.1 Multi-task Moral Hazard

Consider the case in which the principal can observe the agent's type. In this setting, where the agent's effort and cost are not verifiable, information about the verifiable outcome $\mathcal{V}$ is the only contractible element the principal can use to incentivize effort. Let's begin by analyzing the scenario in which this information is not available. I refer to this case as *incentive autarky* and denote the agent's effort choice in autarky by $\tilde{e}^i$.

Since the bad-type agent derives no benefit from the realization of outcomes, he exerts no effort in autarky. In contrast, the good-type agent chooses an effort level that maximizes his utility:

$$\tilde{e}^g = (\tilde{e}^g_v, \tilde{e}^g_n) = \text{argmax}_e\{U^g(e)\}.$$

I assume that the objective function of the good-type agent is concave. Moreover, I assume that the intrinsically motivated agent exerts an interior level of effort on the prosocial task in autarky[14]. Formally:

Assumption 1.

a) $U_{12}(e) \in (-1, 1)$[15].

b) $\tilde{e}^g_v \in (0, 1)$.

When the verifiable outcome $\mathcal{V}$ can be disclosed, the principal may offer a contract that rewards the agent upon its realization. In this setting, since the verifiable outcome is the only contractible information, the principal's optimal contract under non-disclosure is null[16]. Let $X = (w, t)$ denote the contract offered by the principal contingent on the

[14] Assuming $\tilde{e}^g_v > 0$ simplifies the characterization of equilibrium contracts under private types, but it is not necessary for obtaining the main results. See Section 3.2.

[15] $U_{12}(e) = U_{12} = B(1,1) - B(1,0) - B(0,1) - s$

[16] Let $X(\delta) = (w, t; \delta)$ denote a contract offered by the principal, where $\delta \in \{0, 1\}$ denotes the agent's disclosure decision regarding the verifiable outcome $\mathcal{V}$ ($\delta = 0$ for non-disclosure and $\delta = 1$ for a disclosure commitment). Then $X(\delta = 0) = (0, 0)$, and $X = X(\delta = 1)$

agent's commitment to disclosure of the verifiable outcome, where t is an unconditional transfer and w is a payment conditional on $\mathcal{V} = 1$. Limited liability requires $w, t \geq 0$.

It is straightforward that an unconditional transfer t does not affect the effort choice of the agent. The incremental profit of the principal relative to autarky from offering a contract (w, t) to the agent i can be written as:

$$\pi^i(w, t) \equiv Y(e(w)) - Y(\tilde{e}^i) - we^i_v(w) - t.$$

The principal then solves the following problem under a voluntary disclosure game:

$$\max_{(w,t)} \{\pi^i(w, t)\} \tag{1}$$

$$\text{s.t } e^i(w) = \text{argmax}_{e'}\{t + we'_v + U^i(e')\}, \text{ (Incentive Constraint)} \tag{2}$$

$$t + we^i(w) + U^i(e^i(w)) \geq U^i(\tilde{e}^i) + f, \text{ (Participation Constraint)} \tag{3}$$

The principal's optimal contract therefore maximizes her expected payoff given the agent's individually optimal effort choices, while also providing sufficient incentives for the agent to engage in prosocial disclosure and contracting.

Let us begin with the characterization of the equilibria ignoring the participation constraint (assuming $f = 0$). The principal can increase her expected value of the project, $Y(e(w))$, by offering a conditional transfer w. However, by doing so, the principal essentially shares the value of the project with the agent. This means that, while a higher incentive term w increases the project value, the principal's share of this value declines as the conditional payment $we^i(w)$ rises. Hence, the optimal outcome-contingent payment chosen by the principal must equal the marginal increase in the project value through incentivizing higher effort and the marginal increase in the cost of contracting, i.e. profit sharing, with the agent. Such a trade-off between incentivizing effort and sharing surplus is a hallmark of moral hazard settings under limited liability constraints.

In a multitasking framework, the marginal increase in project value from incentivizing higher effort on the task with verifiable outcome depends on the agent's effort choice on the other task. Consider the bad-type agent's choice $e^b(w)$. Since the bad type exerts no effort in autarky and the tasks are cost substitutes, offering a transfer contingent on the outcome $\mathcal{V}$ induces positive effort on that task, $e^b_v(w) > 0$, but has no effect on the

task with unverifiable outcome, $e_n^b(w) = 0$. Thus, motivating the bad type to exert effort on the task with verifiable outcome entails no cross-task effort. The principal's optimal outcome-contingent transfer to the bad type, denoted $\hat{w}_0^b$, reflects a trade-off between the gain from inducing higher effort on the task with verifiable outcome and the loss from profit-sharing.

Principal's optimal incentive term for the bad-type agent is positive, $\hat{w}_0^b > 0$, as long as the principal benefits from the realization of the prosocial outcome, regardless of the realization of the non-verifiable outcome. If the principal values the prosocial outcome only if the non-verifiable outcome is successful, the principal gains no profit from contracting with the bad-type agent[17].

Now consider the good-type agent's effort under a positive incentive transfer $\omega > 0$. Since the good type derives utility from the realization of outcomes and exerts some effort in autarky, the principal faces a multitasking problem when contracting with this agent. An incentive term $w > 0$ increases the marginal gain of effort on the task with verifiable outcome and hence induces higher effort on that task, $e_v^g(w) > \tilde{e}_v^g$. The effect on the task with unverifiable outcome, however, depends on the cross-partial of the good-type's utility function, U_{12}^g. If the benefit complementarity between tasks, B_{12}, is sufficiently strong, so that the tasks are net utility complements for the agent ($U_{12}^g > 0$), then raising effort on the task with verifiable outcome also increases effort on the unverifiable task, $e_v^g(w) > \tilde{e}_n^g$. Conversely, if the benefit complementarity of the two tasks, B_{12}, is not large enough to offset cost substitutability ($U_{12}^g < 0$), then greater effort on the task with verifiable outcome can crowd out effort on the other task, $e_n^g(w) \leq \tilde{e}_n^g$. This interaction between effort incentives across tasks is a distinctive feature of moral hazard in multitasking environments, as in Holmstrom and Milgrom (1991).

Thus, offering an incentive transfer w to the good-type agent affects the principal's profit through three channels. First, it increases project value by inducing greater effort on the task with verifiable outcome. Second, it may either increase or decrease project value through changes in effort on the task with unverifiable outcome, depending on the degree of utility complementarity between tasks for the agent. Third, it reduces the

[17] If $Y_v = 0$, then $\hat{w}_0^b = 0$.

principal's profit by requiring a larger transfer to the agent, thereby sharing more of the surplus. The trade-off among these effects determines the principal's optimal incentive level, denoted $\hat{w}_0^g$, which may be higher or lower than the optimal incentive for the bad-type agent. The following lemma formalizes this result.

Lemma 1. *Assume $f = 0$ and agent types are observable to the principal. In equilibrium:*

a) The principal offers a contract with a positive incentive term to the bad-type agent $\hat{w}_0^b > 0$ if and only if $Y_v > 0$.

b) The principal offers a contract with a positive incentive term to the good-type agent, $\hat{w}_0^g > 0$, if and only if

$$\left(Y_1(\tilde{e}^g) - \tilde{e}_v^g\right) + U_{12}^g\left(Y_2(\tilde{e}^g) + \tilde{e}_v^g U_{12}^g\right) > 0.$$

c) The principal offers a higher incentive term to the good-type agent $\hat{w}_0^g > \hat{w}_0^b$, if and only if the utility complementarity of the tasks either for the principal or the good-type agent, Y_{12} or U_{12}^g, is positive and sufficiently high.

The good-type agent's intrinsic utility over the outcomes affect the profit of prosocial contracting for the principal through three channels:

First, the *direct complementarity effect*: If the tasks are complements for the principal, the effort that the good-type agent may exert on the task with unverifiable outcome increases the marginal profit of inducing effort on the task with verifiable outcome. In particular, for any incentive term w, if $e_n^g(w) > 0$, then the principal derives higher profit from inducing effort on the task with verifiable outcome compared to when $e_n^g(w) = 0$.

Second, the *cross-task effect*: incentivizing effort on one task can distort effort on the other. This effect is positive only if the tasks are utility complement for the good-type agent. If $U_{12}^g > 0$, increasing effort on the verifiable task also boosts effort on the unverifiable one, hence providing higher profit for the principal.

Third, the *free-riding effect*: unlike the bad-type agent, the good-type agent exerts some effort on the verifiable task even in absence of contractual incentives. As a result, when the principal offers a transfer contingent on the verifiable outcome, she must compensate not only for the marginal increase in effort beyond the autarky level, but also for the effort

the agent would have anyway supplied. This reduces the profit of contracting with the good-type agent for the principal. In particular, if the principal and the agent derive the same value from the realization of the outcomes, i.e. $Y(e_v, e_n) = B^g(e_v, e_n)$, the principal has no incentive to motivate higher effort from the good-type agent and does not gain from contracting with him.

Importantly, both the free-riding effect and the negative cross-task effect can be strong enough to eliminate the principal's incentive for prosocial contracting with the good-type agent. The principal's optimal incentive transfer for the good type can exceed that for the bad type, $\hat{w}_0^g > \hat{w}_0^b$, only if the free-riding effect is outweighed by either a sufficiently strong positive cross-task effect, i.e., a large U_{12}^g, or a sufficiently strong complementarity effect, i.e., a large Y_{12}.

Let us now consider the case where disclosure is voluntary and $f > 0$. Let $V^i(w)$ denote the incremental indirect utility of the agent from any outcome contingent transfer w relative to autarky:

$$V^i(w) = \text{Max}_{e^i}\{\omega e_v + U^i(e) - U^i(\tilde{e}^i)\}.$$

For sufficiently low f such that the participation constraint is non-binding, i.e., $V^i(\hat{w}^i) \geq f$, the incentive term $\hat{w}^i$ supports an equilibrium. As the cost of disclosure rises above $V^i(\hat{w}_0^i)$, the principal must increase the agent's net gain from contracting to satisfy the participation constraint. The following proposition characterizes the principal's optimal contract design for varying levels of disclosure cost.

Proposition 1. *Assume agent types are observable to the principal. There exist thresholds $\underline{f}^i$ and $\bar{f}^i$, where $0 \leq \underline{f}^i \leq \bar{f}^i$, such that:*

a) $\forall f \in [0, \underline{f}]$, *the principal offers a contract with $\hat{w}_0^i$ and $t = 0$.*

b) $\forall f \in [\underline{f}, \bar{f}]$, *the principal offers a contracts with $\hat{w}^i(f)$ increasing in f and $t = 0$.*

c) $\forall f > \bar{f}$, *the principal offers no contract.*

As noted above, the agent's limited liability implies that any incentive to increase effort must involve the principal sharing part of her surplus with the agent. By setting the incentive term w^i above $\hat{w}^i$, the principal can increase the agent's share of the project's

value enough to compensate for the disclosure cost, albeit at the expense of a reduced profit for herself. In particular, disclosure can be ensured by choosing $\hat{w}^i(f)$ such that $V^i(\hat{w}^i(f)) = f$, provided this still yields a higher profit than in autarky. Let $\bar{w}^i$ denote the outcome-contingent transfer for which the principal's incremental profit from disclosure and contracting is zero, i.e., $\pi^i(\bar{w}^i) = 0$. The principal therefore offers a disclosure-contingent contract only if $f \leq \bar{f}^i = V^i(\bar{w}^i)$.

Furthermore, proposition 1 establishes that the principal never offers a lump-sum transfer to cover the excess disclosure cost. This means that, for the principal, satisfying the agent's participation constraint by setting $w^i > \hat{w}^i$ is always more profitable than doing so with an unconditional transfer t. The intuition lies in the fact that the total surplus from prosocial contracting, the sum of the principal's and agent's payoff, increases with incentive terms above the principal's optimal contract. In other words, in the range $[\hat{w}^i, \bar{w}^i]$, raising the incentive term above $\hat{w}^i_0$ reduces the principal's share of the surplus but increases the overall surplus available to be shared. Consequently, using a higher incentive term to ensure disclosure is more profitable for the principal than relying on an unconditional transfer.[18]

3.2 Adverse Selection

In this section, I characterize the equilibrium contracts under private agent types. First, let us consider two characteristics of the moral hazard problem that shapes principal's contract design under private types.

Building on the previous section's analysis, an agent strictly prefers a higher outcome-contingent transfer. An increase in the incentive term w raises effort on the task with verifiable outcome $e_v(w)$ and simultaneously increases the payment the agent receives for any level of such effort. Consequently, $V^i(w)$ is strictly increasing in w, with its growth proportional to the agent's effort on the verifiable task, $e^i_v(w)$.

Lemma 2. *$V^i(\omega)$ is strictly increasing and convex, and*

$$\frac{dV^i(w)}{dw} = e^i_v(w).$$

[18] See Section 4 and Lemma 7 for a detailed welfare analysis of prosocial contracting under observable types.

Lemma 2 suggests that, whenever the optimal contracts for the two agent types differ, the principal faces an adverse selection problem. Moreover, it shows that an agent's gain from a contract with incentive term w is proportional to the effort exerted on the task with verifiable outcome, $e_v^i(w)$. Recall that, in autarky, the bad-type agent exerts no effort on either task, whereas the good-type agent exerts a positive level of effort on the task with verifiable outcome. An implication of this assumption is that, the good type exerts strictly more effort on the incentivized task than the bad type, for any outcome transfer w. Consequently, the good type obtains strictly greater utility from any contract[19]. The following lemma formalizes this observation.

Lemma 3. *For any incentive term $w > 0$, the good-type agent exerts more effort on the task with verifiable outcome $e_v^g(w) > e_v^b(w)$, and has a higher gain from contracting $V^g(w) > V^b(w)$.*

These observations, together with Lemma 1, demonstrate that the principal's contracting problem can take multiple forms under private agent types. Specifically, there are cases in which the principal derives greater profit from prosocial contracting with the good type, thereby giving the bad type an incentive to mimic the good type. Conversely, there are cases in which prosocial contracting with the bad type is more profitable for the principal, in which case the good type has an incentive to mimic the bad type. Importantly, in all scenarios, the good type obtains strictly higher utility from any prosocial contract[20].

The next two sections analyze these two scenarios under private agent types and characterize the resulting equilibrium contracts across different levels of disclosure cost.

[19] While this characteristic of the model is a result of Assumption 1 which imposes $\tilde{e}_v^g > 0$, it only simplifies the number of cases to be discussed. If the good-type agent's choice of effort in autarky is a corner solution with $\tilde{e}_v^g = 0$, he might exert a lower effort level $e_v(w)$ compared to the bad-type, for some values of w. The main results regarding the effect of a disclosure mandate would not be qualitatively different in such cases.

[20] The assumption of a quadratic cost function simplifies the characterization of the equilibrium cases that follow. Under a quadratic cost, two cases arise: (i) Adverse selection type I, where for any w, the principal's profit is higher when the good-type agent accepts the contract than when the bad-type agent does; and (ii) Adverse selection type II, where the opposite holds. These two configurations can emerge under a wide range of functional forms for the effort cost. However, with a general cost function, there may exist regions of w for which contracting with the good-type agent is more profitable, and other regions where the bad-type is more profitable. Thus, while the quadratic specification is not necessary for the analysis or the results, it facilitates a cleaner characterization of equilibria.

3.2.1 Adverse Selection Type I; When contracting with the good-type is more profitable

As shown in Lemma 1, strong task complementarity—either for the principal or for the good-type agent—can make prosocial contracting with the good type more profitable than with the bad-type for the principal. In this section, I consider cases where the principal prefers a higher incentive term for the good-type agent, $\hat{w}^g > \hat{w}^b$. In this environment, the principal gains a higher profit from any incentive term when offered to the good-type agent. This implies that when disclosure cost becomes binding, the highest incentive term that the principal is willing to offer the good-type agent to induce his participation is also higher than that of the bad-type agent, i.e. $\bar{w}^g > \bar{w}^b$.

Note that this environment—where motivating the intrinsically motivated agent is more profitable than incentivizing the agent with no intrinsic interest in the outcome—can arise only in a multi-task setting. In the absence of multi-tasking, the agent's intrinsic motivation generates a free-riding effect that makes contracting with him less profitable than contracting with the bad type. The key driver of this environment is the complementarity between prosocial and financial objectives, either from the perspective of the principal or the agent, within a multi-task framework[21].

In the setting with private types, and following the Revelation Principle, the principal's problem can be written as follows[22]:

$$\max_{(w,t)} \sum_i m^i \pi(e^i(w))$$

$$\text{s.t.} \qquad e^i = \text{argmax}_e \left\{ w^i e_v + U^i(e) \right\} \qquad \text{(Incentive Constraint)},$$

$$t^i + V^i(w^i) \geq t^j + V^i(w^j), \quad j \neq i \qquad \text{(Incentive Compatibility Constraint)}$$

[21] There are empirical evidences for complementarity of prosocial and monetary objectives for socially responsible agents. For instance, (Fehrler and Kosfeld, 2014) show that agents are willing to exert higher efforts when they can choose a job with a prosocial mission aligning to their preferences. Also, (Cassar, 2019) find that agents exert higher effort on a mission when its success helps a prosocial cause.

[22] I impose that the non-disclosure contract for each type is null, i.e., $X^i(\delta = 0) = (0, 0)$. In this environment, there could exist equilibria in which the principal offers an unconditional transfer to discourage the bad-type agent from disclosing, $X^b(\delta = 0) = (0, t)$, where $t > 0$. I rule out this possibility for two reasons. First, an agreement that rewards an agent for concealing information may be illegal or unenforceable. Second, a contract that pays for inaction could create perverse incentives, attracting participation from outsiders solely seeking such transfers.

$$t^i + V^i(w^i) - f \geq 0 \quad \text{(Participation Constraint)}$$

Let us begin with the case where disclosure is costless, $f = 0$, so that both agent types disclose the verifiable outcome. In general, the principal has two possible strategies for dealing with adverse selection problem. The first is pooling, in which a single contract with incentive term w^p is offered to both types. The second is separating, in which the principal designs two distinct incentive-compatible contracts, each tailored so that an agent prefers the contract intended for his own type. In particular, the principal might be able to use the unconditional lump sum transfer t to design incentive compatible contracts.

Let $\check{X}^g = (\check{w}^g, \check{t}^g)$ and $\check{X}^b = (\check{w}^b, \check{t}^b)$ denote any pair of contracts intended for the good-type and bad-type agents. The following lemma characterizes the necessary and sufficient condition for such contracts to be offered in equilibrium.

Lemma 4.

a) If $\hat{w}^g > \hat{w}^b$, type revelation is feasible by contracts $\check{X}^i$ such that $\check{w}^g > \check{w}^b$ and $\check{t}^b > \check{t}^g$.

b) Separating contracts $\check{X}^g$ and $\check{X}^b$ yield higher profit than the best pooling contract, if and only if $\check{w}^g > \bar{w}^b$.

Recall from Lemma 3 that, for any incentive term w, the good-type agent exerts more effort on the task with verifiable outcome and thus has a higher gain from any incentive term. This implies that, when the principal is willing to offer a higher incentive term to the good-type agent, $\check{w}^g > \check{w}^b$, separating the types through offering a menu of incentive compatible contracts is feasible. In other words, Lemma 4 ensures that the necessary *single-crossing* condition for the existence of a separating equilibrium is satisfied for this type of adverse selection.

Moreover, Lemma 4 states that separating contracts are more profitable than pooling only when the principal is willing to offer an incentive term above $\bar{w}^b$ to the good type agent. The incentive compatibility constraints imply that, to induce type revelation, the principal must provide an *information rent* to the bad-type agent to make him at least indifferent between the contract designed for him and the one intended for the

good type. This means that separating agent types is costly for the principal. Recall from Lemma 2 that, for $w \in [0, \bar{w}^i]$, raising an agent's payoff is more profitably achieved through outcome-contingent transfers than lump-sum transfers. Hence, when the incentive term of the contract intended for the good type is low such that $\check{w}^g < \bar{w}^b$, the principal strictly prefers pooling at $w^p = \check{w}^g$ over offering a contract with a lower incentive term $\check{w}^b < \check{w}^g$ and an information rent to the bad-type agent. Put differently, separating contracts can only dominate pooling when the principal is willing to grant the good type a contract that yields negative profit if accepted by the bad type.

The following lemma characterizes the equilibrium under adverse selection of type I with no disclosure cost.

Lemma 5. *Assume that $\hat{w}_0^g > \hat{w}_0^b$, and $f = 0$.*

1. *If $\hat{w}_0^g \leq \bar{w}^b$, in equilibrium, the principal offers a pooling contract $\hat{w}_0^p(\lambda) \in [\hat{w}_0^b, \hat{w}_0^g]$, increasing in λ.*

2. *If $\hat{w}_0^g > \bar{w}^b$, there exist a threshold $\lambda_0^s \in [0, 1]$ such that:*

   a) $\forall \lambda \leq \lambda_0^s$, in equilibrium, the principal offers a pooling contract $\hat{w}_0^p(\lambda) \in [\hat{w}_0^b, \bar{w}^b]$, strictly increasing in λ.

   b) $\forall \lambda > \lambda_0^s$, in equilibrium, the principal offers $\check{X}^g(\lambda) = (\check{w}^g(\lambda), 0)$ to the good-type agent and $\check{X}^b(\lambda) = (\check{w}^b(\lambda), \check{t}^b)$ to the bad-type agent where $\check{w}^b = \bar{w}^b$, and $\check{w}^g \in [\bar{w}^b, \hat{w}_0^g]$ and $\check{t}^b(\lambda) > 0$ are increasing in λ.

Figure 3 illustrates the two types of equilibrium contracts that emerge in this setting. When $\hat{w}_0^g \leq \bar{w}^b$, the principal does not find separating contracts profitable and thus offers a pooling contract with contingent transfer $\hat{w}_0^p(\lambda)$ to both agent types. As shown in panel (a) of Figure 3, as the prior probability of the good-type agent increases from zero to one, $\hat{w}_0^p(\lambda)$ rises from $\hat{w}_0^b$ to $\hat{w}_0^g$.

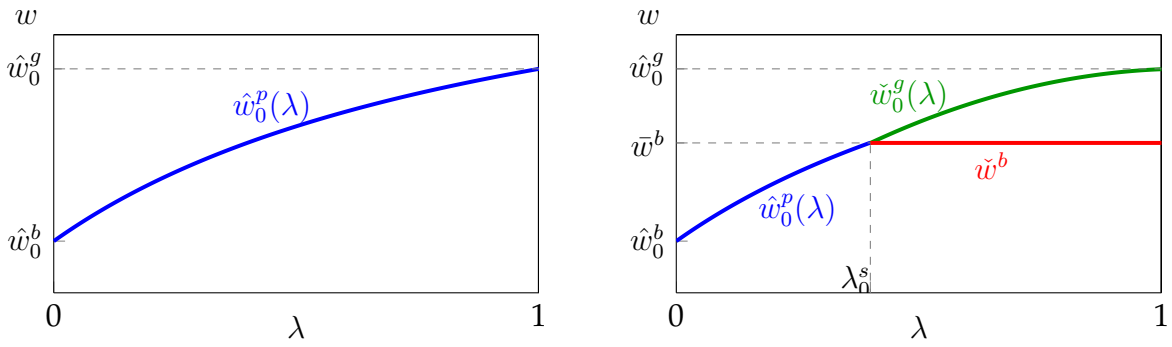


Figure 3: Equilibrium Contract Under Adverse Selection Type I and $f = 0$

When $\hat{w}_0^g > \bar{w}^b$, separating contracts become more profitable than pooling if the principal is willing to offer the good-type agent a contingent transfer exceeding $\bar{w}^b$, which occurs when the prior probability of the good-type agent is sufficiently high. Panel (b) of Figure 3 depicts this scenario: the principal offers the bad-type agent an incentive term of $\bar{w}^b$, together with an information rent, to prevent the bad type from mimicking the contract with $\check{w}^g > \bar{w}^b$ intended for the good-type agent.

In what follows, I focus on the more interesting case where $\hat{w}_0^g > \bar{w}^b$, for which separating contracts emerge in equilibrium. The case where $\hat{w}_0^g \leq \bar{w}^b$ is discussed in Appendix A.

Let us now consider the case where disclosure is voluntary and $f > 0$. Proposition 2 shows that if the principal could observe the agent's type, then once the disclosure cost becomes binding, she could offer a contract with an incentive term $\hat{w}^i(f) > \hat{w}_0^i$ to induce disclosure, provided that doing so yields non-negative profit. This generates a threshold $\bar{f}^i$ for the disclosure cost, above which contracting with type i is no longer profitable. Since the good-type agent yields higher gain from any prosocial contract, $V^g(w) > V^b(w)$, the participation constraint of the bad-type agent binds at lower disclosure costs than that of the good type. This means that contracting with the bad-type agent becomes unprofitable at a lower disclosure cost compared to the good type, i.e; $\bar{f}^g > \bar{f}^b$.

These observations suggest that, with private agent types, three types of equilibria may arise: *full disclosure*, where both types accept a contract and disclose the prosocial outcome; *partial disclosure*, where only the good type does so; and *non disclosure*, where neither type engages in prosocial disclosure. Proposition 2 characterizes these cases.

Proposition 2. *Assume that* $\hat{w}_0^g > \bar{w}^b$.

1. *There exist a threshold function* $\hat{f}(\lambda)$ *such that, for any* λ,

   a) *if* $f \in [0, \hat{f}(\lambda)]$, *both agent types accept a contract and commit to disclosure.*

   b) *if* $f \in (\hat{f}(\lambda), \bar{f}^g]$, *only the good-type agent accepts a contracts and commits to disclosure.*

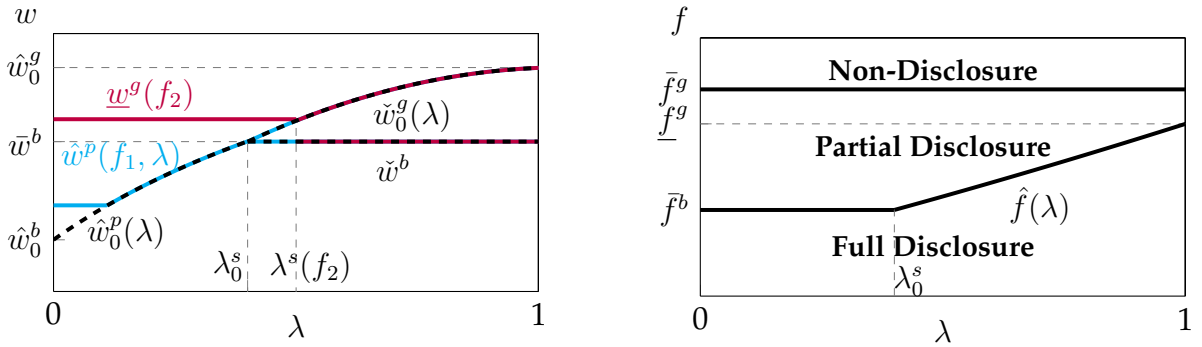
   c) *if* $f > \bar{f}^g$, *neither agent type accepts a contracts and commits to disclosure.*

2. *The threshold* $\hat{f}(\lambda)$ *is weakly increasing in* λ.

As shown in Figure 4(a), when disclosure is costless, the principal offers a pooling contract $w_0^p(\lambda)$ for $\lambda < \lambda_0^s$ and separating contracts $\check{\chi}_0^i(\lambda)$ for $\lambda > \lambda_0^s$. For sufficiently low disclosure costs, $f < \underline{f}^b = V^b(\hat{w}_0^b)$, these contracts induce both types to participate and disclose. However, when f rises to the interval $[\underline{f}^b, \bar{f}^b]$, the contract $w_0^p(\lambda)$ might not induce the participation of the bad-type agent, particularly if λ is low. In this case, the principal raises the incentive term in the pooling contracts to secure the bad type's disclosure. These contracts are depicted by $w^p(f_1, \lambda)$ in Figure 4(a).

Now suppose that $\bar{f}^b < f < V^b(\hat{w}_0^g)$. In this case, prosocial contracting with the bad type is no longer profitable, but the bad type would still accept the principal's optimal contract for the good type, $\hat{w}_0^g$. To exclude the bad-type agent, the principal can offer the incentive term $\underline{w}^g(f)$ such that $V^b(\underline{w}^g(f)) = f$, making the bad type indifferent between participation and rejection. This excluding contract is depicted by $\underline{w}^g(f_2)$ in Figure 4(a).

Note that $\underline{w}^g(f)$ rises with f, suggesting that when the disclosure cost exceeds $\bar{f}^b$, the principal can offer higher incentive terms to the good-type agent while excluding the bad-type. In this setting, the cost of disclosure tightens the incentive compatibility constraint. It allows the principal to offer a higher incentive term compared to the case when disclosure cost is zero, without an information rent. For sufficiently high λ, however, the principal may find the separating contracts $\check{\chi}_0^i(\lambda)$ more profitable than this excluding contract. Hence, there exist a threshold $\lambda^s(f)$ above which the principal offers separating contracts with an incentive term above $\underline{w}^g(f)$ for the good-type agent and an information rent for the bad-type. Since these separating contracts require a larger λ to become



a) Equilibrium Contracts

b) Disclosure Profile

Figure 4: Contracts and Disclosure under Adverse Selection Type I and $\hat{w}_0^g > \bar{w}^b$

optimal as $\underline{w}^g(f)$ increases with f, $\lambda^s(f)$ also increases with f.

If the disclosure cost falls in the interval $[V^b(\hat{w}_0^g), \underline{f}^g]$, the principal can offer the optimal contingent transfer for the good type, $\hat{w}_0^g$, without attracting the bad type. Since this contract does not attract the bad type, it can be profitably offered for any prior distribution of types. In this case, the cost of disclosure fully eliminates the adverse selection problem, as contracting with the bad-type is no longer profitable for the principal and her optimal contract for the good-type agent does not induce bad-type's disclosure.

Lastly, when $f > \underline{f}^g$, the principal must raise the incentive term to $\hat{w}^g(f)$, defined by $V^g(\hat{w}^g(f)) = f$, to ensure good-type participation. Once $f > \bar{f}^g$, contracting becomes unprofitable with either type, and no contract is offered. Figure 4(b) depicts the resulting disclosure profiles: $\hat{f}(\lambda)$ marks the threshold below which full-disclosure equilibria arise, while $\bar{f}^g$ is the cut-off above which a non-disclosure equilibrium emerges.

Remark 1: A key feature of equilibria under this form of adverse selection—which arises from strong complementarity between the two tasks for either the principal or the agent—is the use of disclosure as a screening device that enables the principal to offer stronger incentives to the intrinsically motivated agent. In the partial-disclosure equilibria described above and illustrated in Figure 4(a), the principal provides higher incentive terms to the good-type agent compared to the benchmark case with costless disclosure. Section 4 examines the welfare implications of a mandatory disclosure regulation that removes this screening role.

3.2.2 Adverse Selection Type II; When contracting with the bad-type is more profitable

This section analyzes equilibrium contracts when the intrinsic utility of the good-type agent over the outcomes makes prosocial contracting with the good-type agent less profiting than with the good-type for the principal. As shown in Lemma 1, a strong free-riding effect or negative cross task effect can reduce the profit of principal from increasing good-type agent's effort on the task with verifiable outcome. In this section, I consider cases in which the principal prefers a higher incentive term for the bad-type agent; $\hat{w}^b > \hat{w}^g$. In such an environment, the principal is willing to offer a higher incentive term to the bad-type agent to induce his participation, i.e. $\bar{w}^b > \bar{w}^g$.

When agent types are not observable, these optimal contracts for the principal are not incentive compatible; the good-type agent has an incentive to mimic the bad type in order to benefit from the higher incentive terms designed for him.

Lemma 3 shows that for any incentive term w, the good-type agent exerts higher effort on the task with verifiable outcome, $e_v^g(w) > e_v^b(w)$ and derives a higher gain from contracting $V^g(w) > V^b(w)$. As discussed in the previous section, this implies that when the principal prefers a higher incentive term for the bad type, separating the two types through incentive-compatible contracts is infeasible. In this case, the principal is constrained to offer a single contract to maximize her expected profit. The principal's problem in this setting writes:

$$\max_{w} \sum_{i} m^i \pi(e^i(w))$$

$$\text{s.t.} \qquad e^i = \text{argmax}_e \left\{ ew + U^i(e) \right\} \quad \text{(Incentive Constraint)},$$

$$V^i(w) - f \geq 0 \quad \text{(Participation Constraint)}$$

First consider the case where disclosure is mandatory or $f = 0$, so both types disclosure is ensured. The following lemma characterizes the equilibrium contracts in this case.

Lemma 6. *Assume that $\hat{w}_0^b > \hat{w}_0^g$, and $f = 0$. In equilibrium, the principal offers a pooling contract $w_0^p(\lambda) \in [\hat{w}_0^g, \hat{w}_0^b]$, strictly decreasing in λ.*

In this form of adverse selection, the principal's profit from the good type declines once the incentive term exceeds $\hat{w}_0^g$, whereas her profit from the bad type increases for $w < \hat{w}_0^b$. Accordingly, depending on the prior probability of the bad type, $m^b = 1 - \lambda$, the principal offers a pooling contract $w_0^p(\lambda) \in [\hat{w}_0^g, \hat{w}_0^b]$. Since contracting with the bad type yields higher profit for the principal, $w_0^p(\lambda)$ rises with m^b (falls with λ). These pooling contracts are illustrated in Figure 5 (a).

Now consider the case where disclosure is costly, i.e., $f > 0$. Since the good-type agent derives a higher gain from contracting, any contract that induces disclosure by the bad type automatically satisfies the participation constraint of the good type. Hence, under private agent types and costly disclosure, the equilibrium can take one of three forms: full disclosure, in which both types accept the same contract and commit to disclosure; partial disclosure, in which only the good type accepts a contract and discloses; and non-disclosure, in which no contract is accepted. The following proposition characterizes the equilibrium contracts under adverse selection type II and costly disclosure.

Proposition 3. *Assume that* $\hat{w}_0^b > \hat{w}_0^g$.

1. *There exist a threshold function* $\hat{f}(\lambda)$ *such that, for any* λ,

   a) *if* $f \in [0, \hat{f}(\lambda)]$, *both agent types accept the same contract and commit to disclosure.*

   b) *if* $f \in (\hat{f}(\lambda), \bar{f}^g]$, *only the good-type agent accepts a contract and commits to disclosure.*

   c) *if* $f > \max\{\hat{f}(\lambda), \bar{f}^g\}$, *neither agent type accepts a contracts and commits to disclosure.*

2. *The threshold* $\hat{f}(\lambda)$ *is strictly decreasing in* λ *(increasing in* m^b*).*

The optimal pooling contract under no disclosure cost $w_0^p(\lambda)$ can construct an equilibrium if it satisfies the participation of the bad-type agent. Since $w_0^p(\lambda)\ in[\hat{w}_0^g, \hat{w}_0^b]$, if the disclosure cost is low enough such that $f < V^b(\hat{w}_0^g)$, all pooling contracts induce a full disclosure equilibrium.

Suppose that the pooling contract $w_0^p(\lambda)$ fails to satisfy the participation constraint of the bad-type agent but continues to induce participation by the good type. Since the

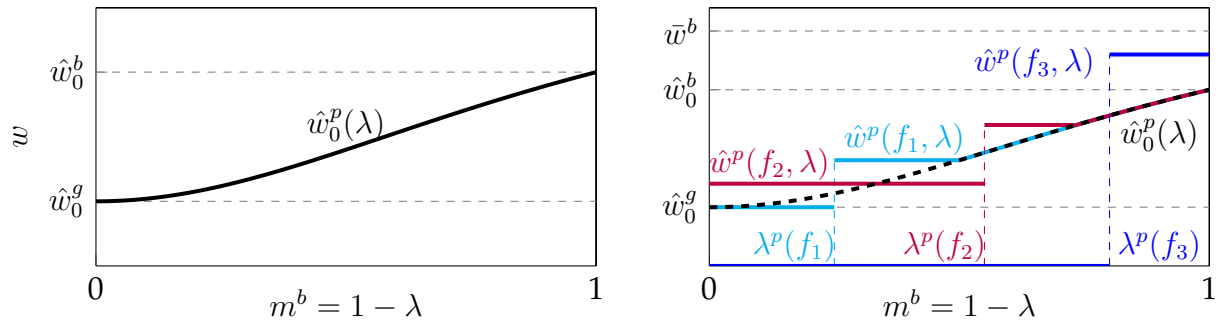


Figure 5: Equilibrium Contract Under Adverse Selection Type II

participation of the good type can always be secured at a lower incentive term than that required for the bad type, i.e. $\hat{w}^g(f) < \hat{w}^b(f)$, the principal faces a trade-off: she may either raise the incentive term to $\hat{w}^b(f)$ to induce disclosure from both types, or offer the lower incentive term $\hat{w}^g(f)$ that attracts only the good type. Intuitively, the prior probability of the bad type, m^b, governs this choice. A full-disclosure equilibrium is optimal only when m^b is sufficiently large, otherwise the principal prefers partial disclosure.

For instance, consider $f_1 \in [V^b(\hat{w}_0^g), V^g(\hat{w}_0^g)]$. At this cost of disclosure, while the pooling contract $\hat{w}^p(\lambda)$ may fail to satisfy the participation constraint of the bad type—particularly when m^b is low—it nevertheless ensures the participation of the good type. As illustrated by $\hat{w}^p(f_1, \lambda)$ in Figure 5 (b), the principal offers the contract $\hat{w}^b(f)$, which induces disclosure from both types, only when the share of bad-type agents in the prior is sufficiently large. Conversely, when the prior probability of the bad type lies below the threshold $\lambda^p(f_1)$, the principal prefers to offer $\hat{w}_0^g$, thereby excluding the bad type.

As long as the principal is willing to engage in contracting with both agent types, i.e. $f < \bar{f}^b$ and $f < \bar{f}^g$, she faces the same trade-off: offering $\hat{w}^b(f)$ to induce participation by both types, or offering $\hat{w}^g(f)$ only to the good type. For instance, Figure 5(b) illustrates this through the contracts at $\lambda^p(f_2)$: the principal must offer $\hat{w}^g(f_2) > \hat{w}_0^g$ to ensure participation by the good type, while a still higher incentive term $\hat{w}^b(f_2)$ is required to induce disclosure by both types. A full-disclosure equilibrium arises when the share of bad-type agents exceeds the threshold $\lambda^p(f_2)$. Importantly, this threshold increases with the disclosure cost, since the principal requires a larger probability of the more profitable type (the bad type, in this case) to justify inducing a full-disclosure equilibrium.

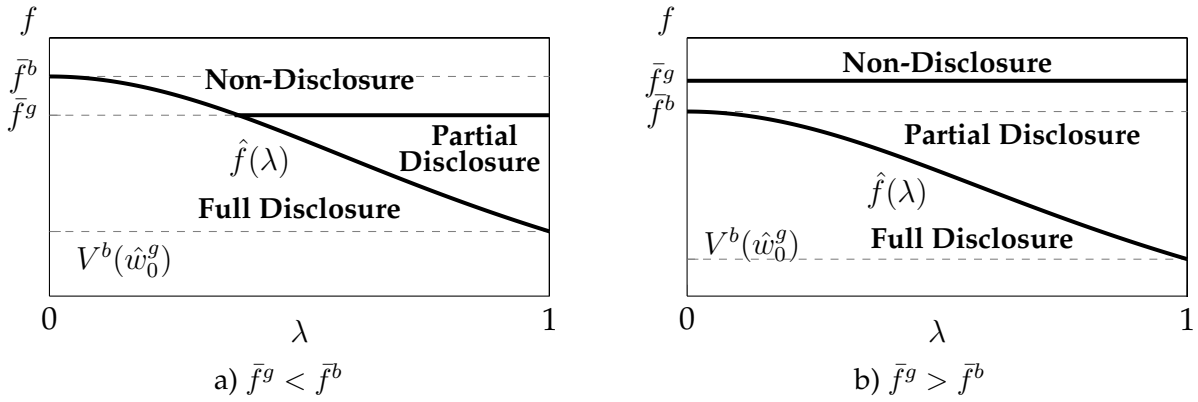


Figure 6: Disclosure Profile Under Adverse Selection Type II

While $\bar{w}^g > \bar{w}bg$ in this setting, since the good-type agent has a higher gain from any incentive term, i.e. $V^g(w) > V^b(w)$, we can have cases in which $\bar{f}^g < \bar{f}^b$ or $\bar{f}^g > \bar{f}^b$. For instance, contracts $\lambda^p(f_3)$ in Figure 5(b) depicts equilibrium contracts for a disclosure cost f_3 such that $f_3 \in [\bar{f}^g, \bar{f}^b]$. In this case, the principal can not profit from any contract accepted by the good-type agent, but can profit from contracting with the bad-type agent by offering a contract $\hat{w}^b(f) > \bar{w}^g$. Since such a contract is costly to the principal if accepted by a good-type agent, the principal is willing to offer it only if the share of bad-type agent in the prior is sufficiently high, i.e. $\lambda < \lambda^p(f_3)$.

Lastly, in cases where $\bar{f}^g > \bar{f}^b$, the disclosure cost may be sufficiently high such that contracting with the bad type is no longer profitable, while a profitable contract $\hat{w}^g(f)$ that induces disclosure by the good type still exists. Since such a contract does not attract the bad type, the principal can offer it irrespective of the prior distribution. Put differently, if $\bar{f}^g > \bar{f}^b$, then for any $f \in [\bar{f}^b, \bar{f}^g]$, a partial-disclosure equilibrium arises independently of the prior over types. Figure 6 illustrates the two disclosure profiles corresponding to the cases $\bar{f}^g < \bar{f}^b$ and $\bar{f}^g > \bar{f}^b$.

Remark 2: The screening role of voluntary disclosure under this type of adverse selection differs fundamentally from that in adverse selection type I. In this environment, the principal may exclude the bad-type agent to limit the incentives offered to the good-type agent, either because of negative cross-task effects or strong free-riding considerations. The next section examines the impact of mandatory disclosure regulation, which removes

this screening role.

4 Welfare Analysis

This section examines the welfare implications of prosocial disclosure and contracting. The main objective is to study the welfare implications of regulations that mandate the disclosure of the verifiable outcome $\mathcal{V}$[23].

In this model, a mandatory disclosure regulation affects welfare through two channels: a direct mechanical effect stemming from changes in total expenditure on disclosure, and an indirect effect through their influence on the equilibrium contracts. To distinguish between these, I define two welfare measures. The first is *Contracting efficiency* or *Gross Welfare*, which captures the welfare derived from the principal's expected benefit and the agent's utility in the equilibrium, while excluding disclosure costs. The second is *Net Welfare*, which additionally incorporates the cost of disclosure[24].

Formally, I define the gross welfare or efficiency of a contract X between the principal and agent i as:

$$GW^i(X) = \pi^i(X) + V^i(X),$$

and the net welfare of such a contract as:

$$NW^i(X) = GW^i(X) - f.$$

The following lemma formalizes the gross welfare or efficiency of a contract with incentive term X between the principal and an i-type agent.

Lemma 7.

[23] Prominant examples include Corporate Sustainability Reporting Directive (CSRD) and the Sustainable Finance Disclosure Regulation (SFDR) in the European Union

[24] One rationale for distinguishing between the two welfare measures lies in the potential externalities of the prosocial outcome, which are explicitly excluded from the model. A social planner may mandate the disclosure of a verifiable prosocial outcome to incentivize greater effort toward such outcomes and to amplify their positive externalities. In that case, a policy that raises prosocial effort can improve overall social welfare, even if the imposed disclosure expenses outweigh the total surplus accrued by the principal and the agent through contracting.

Another reason to distinguish between the two welfare metrics concerns the nature of the disclosure cost. For instance, if the disclosure cost represents a payment to a monopolistic third party that verifies the disclosed information, its negative effect in the planner's objective function would be less than the cost for the agent.

1. *Efficiency of a contract with the bad-type agent reaches its maximum with the incentive term* $w^b_{FB} = \bar{w}^b$*, where* $\bar{w}^b > \hat{w}^b_0 > 0$*, if* $Y_E > 0$*.*

2. *Efficiency of a contract with the good-type agent reaches its maximum at the incentive term* w^g_{FB}*, where* $w^g_{FB} > \bar{w}^g > 0$*, if* $Y_1(\dot{a}^g) - D_{12}Y_2(\dot{e}^g) > 0$*.*

In the moral hazard setting, the principal can influence the agent's effort only by sharing part of the project's value with him. Thus, contracting generates surplus only if providing the agent with a positive incentive increases the project's value to the principal above the autarky level. As discussed in Section 3.1, a sufficiently strong negative cross-task effect can eliminate the surplus from any such contract. Consequently, as Lemma 7 shows, prosocial contracting improves welfare only when the verifiable outcome is valuable to the principal and the negative cross-task effect remains limited.

However, the agent's limited liability introduces a source of inefficiency in this contracting environment. Since incentivizing agent's effort is possible only through profit sharing, the principal's objective differs from the total welfare; she is interested in motivating the agent toward higher efforts as long as it increases her net profit. Specifically, while the principal's profit—her net share of the project value—declines once the incentive term exceeds her optimum $\hat{w}^i$, the total surplus of the contract continues to increase over the interval $[\hat{w}^i, \bar{w}^i]$. In fact, for the bad-type agent, the maximum surplus is attained at $\bar{w}^b = Y_E$, a point at which the principal's profit falls to zero and the agent earns the total surplus.

As discussed in Section 3.1, the effort exerted by the good-type agent in autarky creates a free-riding effect: the principal has no incentive to share the portion of the project value that she would obtain without contracting. This free riding generates extra inefficiency in contracts with the good-type agent. Lemma 7 shows that contract efficiency with the good-type agent is maximized at an incentive term $w^g_{FB} > \bar{w}^b$, where the principal's profit falls strictly below her autarky payoff. In other words, gross surplus peaks when the principal not only shares the incremental value created by contracting but also transfers the autarky value of the project to the agent.

4.1 Mandatory Disclosure Regulation

This section analyzes the effect of a regulatory mandate requiring agents to verifiably disclose the prosocial outcome $\mathcal{V}$. Under such regulation, disclosure is compulsory, and its cost is borne by the agent regardless of any contractual arrangement with the principal. A disclosure mandate relaxes the agent's participation constraint. Therefore, the equilibrium contracts under mandatory disclosure are equivalent to those under a voluntary disclosure regime with zero disclosure cost.

The effect of this mandate on prosocial contracting depends critically on the contracting environment and the equilibrium that would arise under a voluntary regime. I consider three cases in turn: observable types, adverse selection type I, and type II.

Observable types: Consider first the case where agent types are observable. As shown in Section 3.1, when agent's participation constraint is not binding, the principal offers a contract with incentive term $\hat{w}_0^i$. For $f \in (\hat{f}^i, \bar{f}^i)$, voluntary disclosure requires a higher incentive term $\hat{w}^i(f) > \hat{w}_0^i$, while for $f > \bar{f}^i$ it ceases altogether. In such cases, when voluntary disclosure fails to induce prosocial contracting, a mandatory regime still sustains the contract $\hat{w}_0^i$, though at the cost f borne by the agent. The next lemma formalizes the impact of a mandatory disclosure regulation on prosocial contracting and welfare when agent types are observable.

Proposition 4. *Assume that agent types are observed by the principal.*

1. *When voluntary regime results in non-disclosure, mandating disclosure can improve gross welfare while reducing net welfare.*

2. *When voluntary regime results in disclosure, mandating disclosure might reduce both gross and net welfare.*

With observable types, when voluntary disclosure fails to induce contracting, mandating disclosure can sustain a contract with incentive term $\hat{w}_0^i$, which raises effort on the verifiable task and improves both the agent's and the principal's payoff. Such a mandate increases gross welfare, but the gain is insufficient to offset the disclosure cost—otherwise

the contract would emerge voluntarily. Thus, mandatory disclosure enhances prosocial effort and expected outcomes, albeit at the expense of imposing the disclosure cost on the agent.

In cases where prosocial disclosure and contracting arise under a voluntary regime, mandatory disclosure can reduce both gross and net welfare. When disclosure is voluntary and the cost is binding, the principal increases the agent's share by offering a higher incentive term to secure participation. Specifically, for $f \in (\hat{f}^i, \bar{f}^i)$, voluntary disclosure leads to a contract with $\hat{w}^i(f) > \hat{w}_0^i$, whereas mandatory disclosure fixes the incentive at $\hat{w}_0^i$. As established in Lemma 7, contract efficiency increases over $[\hat{w}_0^i, \bar{w}^i]$, implying that the contract under voluntary disclosure yields higher gross and net welfare than its counterpart under mandatory disclosure.

The intuition comes from the inefficiency created by limited liability. Because the principal's objective is to maximize her own profit rather than total surplus, she sets the incentive term below the level that would maximize efficiency. However, when disclosure is voluntary and costly, the principal must raise the agent's share of the project value to induce participation. This shift aligns the contract more closely with the efficient level, thereby reducing the inefficiency that stems from the divergence between the principal's objective and overall welfare.

This constitutes an important channel through which mandatory disclosure can harm efficiency and welfare. The principal might offer higher incentives under a voluntary regime to induce agent's disclosure. Therefore, mandatory disclosure can reduce the incentives by relaxing the participation constraint of the agent.

Adverse Selection Type I: Consider the environment where prosocial contracting with the good-type agent is more profitable for the principal, particularly due to a strong complementarity of the tasks for the principal and the good-type agent. As discussed in Section 3.2.1, voluntary disclosure may lead to full, partial, or non-disclosure equilibria. The following proposition examines the effect of a disclosure mandate in each case.

Proposition 5. *Assume that agent types are private and* $\hat{w}_0^g > \hat{w}_0^b$

1. *When the voluntary regime results in non-disclosure, mandatory disclosure increases gross welfare but reduces net welfare.*

2. *When the voluntary regime results in partial disclosure, mandatory disclosure may increase or decrease gross welfare but reduces net welfare.*

3. *When the voluntary regime results in full disclosure, mandatory disclosure weakly decreases gross and net welfare.*

As shown in Figure 5 (b), under this form of adverse selection, a non-disclosure equilibrium arises when the cost of disclosure is sufficiently high to render prosocial contracting with either type unprofitable for the principal, i.e. when $f > \bar{f}^g > \bar{f}^b$. In this case, mandating disclosure can generate contracts, either separating or pooling, that rely on information which would not be provided under a voluntary regime. These contracts increase effort on the task with verifiable outcome and raise both the agent's and the principal's payoff, thereby enhancing gross welfare. However, this improvement is insufficient to offset the disclosure cost, implying that mandatory disclosure ultimately reduces net welfare.

Let us consider the effect of disclosure mandate in cases where voluntary regime results in partial disclosure equilibria; as discussed in section 3.2.1, when prosocial contracting is more profitable with the good-type, there exist partial disclosure equilibria in which only the good-type engages in prosocial disclosure and contracting. In such cases, mandating disclosure has two effects on the equilibrium contracts; it enables contracting with the bad-type, and it alters the contract offered to the good-type. As illustrated in Figure 4 (a), the latter effect is negative, whether $\hat{w}^g > \bar{w}^b$ or $\hat{w}^g < \bar{w}^b$[25]. In this environment, the good-type agent secures a more favorable contract when the bad-type is excluded from prosocial contracting under the voluntary regime.

This means that, when voluntary regime results in partial disclosure, mandatory disclosure reduces the contracting efficiency with the good-type agent, while generating welfare from contracting with the bad-type agent. As noted in Section 3.2.1, under this type of adverse selection, specifically when $\hat{w}^g > \bar{w}^b$, voluntary disclosure serves as a

[25]See Appendix A.

screening device for the principal, allowing her to offer contracts to the intrinsically motivated agent without attracting the less profitable type. Mandatory disclosure eliminates this screening mechanism and diminishes the screening role of disclosure. Consequently, while mandating disclosure boosts the bad-type agent's effort on the task with verifiable outcome, it reduces the good-type agent's efforts by exacerbating the adverse selection problem.

The overall effect of mandatory disclosure in this case on gross welfare can be either positive or negative, depending on the relative magnitude of the welfare generated by contracting with the bad-type agent relative to the loss from reduced efficiency of the contract with the good-type agent. In particular, when task complementarity for the good-type agent is strong, so that prosocial contracting significantly increases effort on both tasks, the negative impact of mandatory disclosure on the good-type agent may outweigh the welfare gains from contracting with the bad-type agent. In such cases, a mandatory disclosure regulation reduces gross welfare. Conversely, if the complementarity effect is not strong enough and the bad-type agent's effort is sufficiently valuable, the welfare created by contracts enabled with the bad-type agent can dominate the loss from reduced efficiency with the good-type agent.

Moreover, when a partial disclosure equilibrium arises under a voluntary regime, mandating disclosure reduces net welfare. Even if the welfare created from enabled contracts with the bad-type outweighs the loss from weaker contracting with the good-type, so that gross welfare rises, the gain is insufficient to offset the additional disclosure costs. In fact, the cost of disclosure in this case exceeds the welfare generated by the newly enabled bad-type contract. Since mandatory disclosure also worsens the contract offered to the good-type, it cannot generate a net welfare gain in this environment.

Lastly, consider the case where the voluntary regime results in full disclosure. If mandating disclosure affects the equilibrium contracts, the effect is negative. As in the observable-types setting, under a voluntary regime the principal may offer a higher incentive than under mandatory disclosure to motivate the agent to disclose. By removing the need to provide such motivation, mandatory disclosure reduces the incentive term and thereby lowers the efficiency of equilibrium contracts.

Adverse Selection Type II: Under this type of adverse selection, the principal prefers to offer a lower incentive term to the good-type than the bad-type agent, driven by either a strong free-riding effect or a negative cross-task effect. Note that, although both forces can make contracting with the good-type unprofitable for the principal, their implications for welfare differ. A negative cross-task effect lowers the welfare frontier of contracting, whereas the free-riding effect arises solely from the divergence between the principal's objective and social welfare. Consequently, even though contracts with the bad-type may be more profitable for the principal, contracting with the good-type can yield higher social value when the free-riding effect is sufficiently strong.

As discussed in Section 3.2.2, this environment admits full, partial, and non disclosure equilibria under a voluntary regime. The following proposition characterizes the welfare implications of a disclosure mandate in this setting.

Proposition 6. *Assume that agent types are private and* $\hat{w}_0^b > \hat{w}_0^g$

1. *When the voluntary regime results in non-disclosure, mandatory disclosure results increases gross welfare might increase or decrease net welfare.*

2. *When the voluntary regime results in partial disclosure, mandatory disclosure may increase or decrease gross and net welfare.*

3. *When the voluntary regime results in full disclosure, mandatory disclosure weakly decreases gross and net welfare.*

An important implication of this proposition is that, unlike the cases of observable types or adverse selection type I, under this type of adverse selection, a disclosure mandate can sustain contracts that yield a higher net welfare than those arising in the voluntary regime. In other words, mandatory disclosure can provide contractible information that generates welfare gains exceeding the disclosure costs it imposes.

Let us consider a full non-disclosure equilibrium in this setting. As illustrated in Figure 6, such an equilibrium can arise when contracting with the good-type agent is not profitable for the principal, i.e. $f > \bar{f}^g$. In cases where $f < \bar{f}^b$ (as depicted in panel (a) of Figure 6), contracting with the bad-type agent remains profitable. However, the principal

may still refrain from inducing disclosure, since doing so would expose her to losses from the good-type agent accepting the same contract. In this case, the contract sustained under mandatory disclosure can result not only in a higher gross welfare, but also a higher net welfare.

In this setting, a mandatory disclosure contract raises the prosocial effort of the bad-type agent, thereby increasing gross welfare. When $f < \bar{f}^b$, the net welfare generated by this contract can be positive, even though the bad-type agent's individual payoff is negative. At the same time, the good-type agent also exerts higher effort under the same contract, which can further enhance both gross and net welfare, particularly when the free-riding effect is strong. Overall, the efficiency gains from mandatory disclosure may more than offset the disclosure cost, yielding a higher net welfare relative to the voluntary regime.

Let us now consider a partial disclosure equilibrium in this setting. Proposition 6 establishes that, in such a case, mandatory disclosure may increase or decrease both gross and net welfare. The mandate generates two distinct effects: first, it provides information that enables contracting with the bad-type agent, which unambiguously raises gross welfare; second, it alters the contract offered to the good-type agent. The latter effect can either enhance or diminish efficiency, depending on whether the pooling contract under mandatory disclosure entails a higher or lower incentive term relative to the excluding contract under the voluntary regime.

For instance, consider the contracts $\hat{w}^p(f_1, \lambda)$ depicted in Figure 5(b). When the share of bad-type agents in the prior is low, the principal offers the contract $\hat{w}_0^g$ to the good-type agent, excluding the bad-type. Under mandatory disclosure, the contract becomes $\hat{w}_0^p(\lambda) > \hat{w}_0^g$, implying higher incentive terms for both agent types and, consequently, higher gross welfare. Importantly, Proposition 6 shows that this increase in gross welfare can be large enough to offset the additional disclosure cost, such that net welfare may also rise under mandatory disclosure. In particular, when a strong free-riding effect exists in contracting with the good-type agent, the welfare gains from the higher incentive offered under mandatory disclosure can fully compensate for the increased disclosure expenses.

Conversely, there are cases in which mandatory disclosure leads to a contract with

a lower incentive term for the good-type agent compared to the voluntary regime. For instance, consider the contracts $\hat{w}^p(f_1, \lambda)$ depicted in Figure 5(b); when the share of bad-type agents in the prior is low, mandatory disclosure can reduce the incentive term offered to the good-type agent. In such cases, mandatory disclosure may increase or decrease gross welfare relative to the voluntary regime, as it lowers the efficiency of the contract with the good-type agent while generating welfare by enabling contracting with the bad-type agent. The overall effect on gross and net welfare depends on the relative magnitude of these two opposing effects.

Finally, when voluntary disclosure results in a full disclosure equilibrium, mandatory disclosure can sustain an equilibrium with a lower incentive term, as illustrated in Figure 5(b). Under a mandatory regime, the principal no longer needs to raise the incentive term to induce the agent's disclosure. This reduction in the incentive term lowers both gross and net welfare. Thus, as in the case of observable types or adverse selection type I, when disclosure and contracting with both agent types occur voluntarily and the disclosure cost is binding, mandating disclosure can reduce the efficiency and welfare.

5 Conclusion

This paper develops a theoretical framework to analyze prosocial disclosure and contracting in a principal–agent model with multiple hidden actions and privately known agent types. In this setting, the disclosure of verifiable information about a prosocial outcome enables the principal to incentivize prosocial effort by linking rewards to measurable performance. The design and efficiency of such contracts depend critically on the agent's intrinsic utility over both verifiable and unverifiable outcomes, as well as on the degree of complementarity between these outcomes for the principal.

The main contribution of this paper is to analyze how mandatory disclosure regulation influences prosocial contracting across environments shaped by agents' private types. I show that its impact is generally ambiguous and highly sensitive to heterogeneity in agents' intrinsic values. Notably, I identify cases where mandatory disclosure can either enhance or undermine both prosocial effort and overall welfare.

From a social planner's perspective, the impact of such regulation can be anticipated from the prosocial contracting that emerges under a voluntary regime. When voluntary disclosure results in full disclosure and the cost of disclosure is binding, a mandatory disclosure regulation unambiguously reduces contracting efficiency and net welfare. When voluntary disclosure results in non-disclosure, a disclosure mandate can induce prosocial contracts that increase effort and, in certain cases, enhance net welfare.

The most nuanced case arises when the voluntary regime leads to partial disclosure. In such cases, the effect of a disclosure mandate on contracting efficiency and welfare depends sensitively on the agent's private types. In particular, the effect is more likely to be negative when prosocial and financial objectives are strong complements for the principal or the intrinsically motivated agents. Conversely, mandatory disclosure is more likely to enhance welfare in the presence of a strong free-riding effect in contracting with the intrinsically motivated agents.

This analysis underscores the need for sophisticated, and potentially sector-specific, disclosure mechanisms for prosocial performance that take into account the interdependence of firms' multiple hidden actions and types. Future research could extend this framework by investigating optimal prosocial performance metrics, analyzing dynamic settings with richer contract structures, or empirically testing the model's predictions.

References

Aghamolla, Cyrus, and Byeong-Je An, "Mandatory vs. voluntary ESG disclosure, efficiency, and real effects," *Nanyang Business School Research Paper*, (2023).

Akerlof, George A., and Rachel E. Kranton, "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, 19 (2005), 9–32, doi:10.1257/0895330053147930, available at: https://www.aeaweb.org/articles?id=10.1257/0895330053147930.

Albuquerque, Rui, Yrjö Koskinen, and Chendi Zhang, "Corporate social responsibility and firm risk: Theory and empirical evidence," *Management science*, 65 (2019), 4451–4469.

Auzepy, Alix, Christina E Bannier, and Fabio Martin, "Are sustainability-linked loans designed to effectively incentivize corporate sustainability? A framework for review," *Financial Management*, 52 (2023), 643–675.

Barbalau, Adelina, and Federica Zeni, "The optimal design of green securities," *Available at SSRN 4155721*, (2022).

Basu, Sudipta, Justin Vitanza, Wei Wang, and Xiaoyu Ross Zhu, "Walking the walk? Bank ESG disclosures and home mortgage lending," *Review of Accounting Studies*, 27 (2022), 779–821.

Bénabou, Roland, and Jean Tirole, "Individual and corporate social responsibility," *Economica*, 77 (2010), 1–19.

Besley, Timothy, and Maitreesh Ghatak, "Profit with Purpose? A Theory of Social Enterprise," *American Economic Journal: Economic Policy*, 9 (2017), 19–58, doi: 10.1257/pol.20150495, available at: https://www.aeaweb.org/articles?id=10.1257/pol.20150495.

Bond, Philip, and Yao Zeng, "Silence is safest: Information disclosure when the audience's preferences are uncertain," *Journal of Financial Economics*, 145 (2022), 178–193, doi:https://doi.org/10.1016/j.jfineco.2021.08.012, available at: https://www.sciencedirect.com/science/article/pii/S0304405X2100369X.

Bonham, Jonathan, and Amoray Riggs-Cragun, "Motivating green innovation through ESG performance shares and ESG-contingent income tax rates," *Chicago Booth Research Paper*, (2025).

Bénabou, Roland, and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, 96 (2006), 1652–1678, doi:10.1257/aer.96.5.1652, available at: https://www.aeaweb.org/articles?id=10.1257/aer.96.5.1652.

Cassar, Lea, "Job mission as a substitute for monetary incentives: Benefits and limits," *Management Science*, 65 (2019), 896–912.

Cohen, Shira, Igor Kadach, Gaizka Ormazabal, and Stefan Reichelstein, "Executive compensation tied to ESG performance: International evidence," *Journal of Accounting Research*, 61 (2023), 805–853.

Fehrler, Sebastian, and Michael Kosfeld, "Pro-social missions and worker motivation: An experimental study," *Journal of Economic Behavior and Organization*, 100 (2014), 99–110, doi:https://doi.org/10.1016/j.jebo.2014.01.010, available at: https://www.sciencedirect.com/science/article/pii/S016726811400016X.

Gladilina, Irina, Hafis Hajiyev, Irina Vaslavskaya, Elena Kirillova, Olga Dymchenko, Emil Hajiyev, Olga Averina, and Rustem Shichiyakh, "Adapting ESG Principles to Contracting Practices: Towards Sustainable Business Agreements." *International Journal of Sustainable Development & Planning*, 19 (2024).

Goldstein, Itay, Alexandr Kopytov, Lin Shen, and Haotian Xiang, "On ESG investing: Heterogeneous preferences, information, and asset prices," (2022).

Heinkel, Robert, Alan Kraus, and Josef Zechner, "The effect of green investment on corporate behavior," *Journal of financial and quantitative analysis*, 36 (2001), 431–449.

Holmstrom, Bengt, and Paul Milgrom, "Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design," *The Journal of Law, Economics, and Organization*, 7 (1991), 24–52.

Kim, Sehoon, Nitish Kumar, Jongsub Lee, and Junho Oh, "ESG lending," *Journal of Financial Economics*, (2021).

Loumioti, Maria, and George Serafeim, "The issuance and design of sustainability-linked loans," *Available at SSRN*, (2022).

Moharram, AH, Hafiza Aishah Hashim, WALEED M Alahdal, and SHAYUTI BINTI MOHAMED Adnan, "Should ESG disclosure be mandatory? An overview," *Journal of Sustainability Science and Management*, 19 (2024), 221–236.

Oehmke, Martin, and Marcus M Opp, "A Theory of Socially Responsible Investment," *The Review of Economic Studies*, 92 (2024), 1193–1225, doi:10.1093/restud/rdae048, available at: https://doi.org/10.1093/restud/rdae048.

Pedersen, Lasse Heje, Shaun Fitzgibbons, and Lukasz Pomorski, "Responsible investing: The ESG-efficient frontier," *Journal of financial economics*, 142 (2021), 572–597.

Weksler, Ran, and Boaz Zik, "Disclosure in Markets for Ratings," *American Economic Journal: Microeconomics*, 15 (2023), 501–526.

Yang, Zhi, Nguyen Thi Thu Huong, Nguyen Hoang Nam, Nguyen Thi Thuy Nga, and Cao Thi Thanh, "Greenwashing behaviours: Causes, taxonomy and consequences based on a systematic literature review," *Journal of Business Economics and Management (JBEM)*, 21 (2020), 1486–1507.

Zhang, Dongyang, "Green financial system regulation shock and greenwashing behaviors: Evidence from Chinese firms," *Energy Economics*, 111 (2022), 106064, doi:https://doi.org/10.1016/j.eneco.2022.106064, available at: https://www.sciencedirect.com/science/article/pii/S0140988322002304.

Appendix

A Omitted Cases

Adverse Selection Type I ($\hat{w}_0^g \leq \bar{w}^b$). This section examines the effect of disclosure costs on equilibrium contracts in environments characterized by adverse selection type I, where separating contracts are everywhere less profitable than pooling ones (i.e., when $\hat{w}_0^g < \bar{w}^b$). As illustrated in Figure 3a, when the disclosure cost is zero, the principal offers a pooling contract $w^p(\lambda) \in [\hat{w}_0^b, \hat{w}_0^g]$. Such a pooling contract can constitute an equilibrium if the disclosure cost is sufficiently low so that the bad-type agent is willing to accept it. The following proposition characterizes the equilibrium contracts as the disclosure cost varies.

Proposition 7. *Assume that $\hat{w}_0^g \leq \bar{w}^b$.*

1. *There exist a threshold function $\hat{f}(\lambda)$ such that, for any λ,*
   a) *if $f \in [0, \hat{f}(\lambda)]$, both agent types accept a contract and commit to disclosure.*
   b) *if $f \in (\hat{f}(\lambda), \bar{f}^g]$, only the good-type agent accepts a contracts and commits to disclosure.*
   c) *if $f > \bar{f}^g$, neither agent type accepts a contracts and commits to disclosure.*
2. *The threshold $\hat{f}(\lambda)$ is strictly decreasing in λ.*

Assume the cost of disclosure exceeds $\underline{f}^b$, so that the principal has to ensure the participation of the bad-type agent by an incentive term $\hat{w}^b(f) > \hat{w}_0^b$. As shown in Figure 7(a), as long as $\hat{w}^b(f)$ remains under $\hat{w}_0^g$, the principal provides enough incentive for the bad-type to engage in prosocial disclosure and contracting, regardless of the prior. In this range of disclosure cost, i.e. $f \in [\underline{f}^b, V^b(\hat{w}_0^g)]$, raising the incentive term of the pooling contract increases principal's profit from contracting with the good-type agent while ensuring the participation of the bad-type agent. These equilibria are depicted as $\hat{w}^p(f_1, \lambda)$ in Figure 7.

Now assume the disclosure cost is above $V^b(\hat{w}_0^g)$, such that the incentive term that ensures bad-type's disclosure is higher than the optimal contract with the good-type agent,

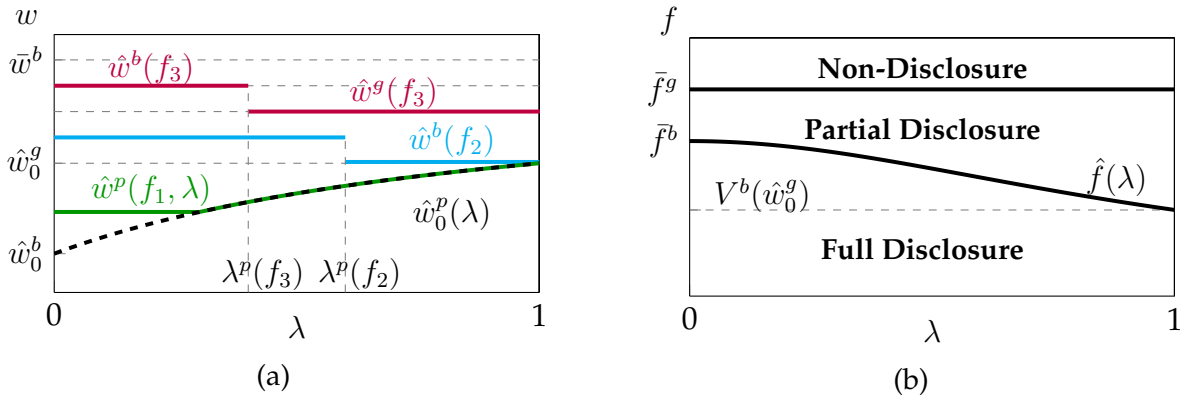


Figure 7: Equilibrium Contract Under Adverse Selection Type I and $\hat{w}_0^g \leq \bar{w}^b$

i.e $\hat{w}^b(f) > \hat{w}_0^g$. In this case, if the principal offers $\hat{w}^b(f)$ to ensure the participation of the bad-type agent, the good-type agent prefers to accept this contract rather than the optimal contract $\hat{w}_0^g$ intended for him. This suggests that, for $f > V^b(\hat{w}_0^g)$, the direction of the adverse selection becomes reversed; the good-type agent prefers to mimic the bad-type agent.

Suppose that $f \in [V^b(\hat{w}_0^g), \bar{f}^b]$. In this interval, inducing bad-type's disclosure by offering $\hat{w}^b(f) > \hat{w}_0^g$ is profitable, but incurs a cost to the principal by attracting the good-type agent. The principal then has two options: either satisfying the bad-type agent's participation by offering $\hat{w}^b(f)$, or offering the optimal contract $\hat{w}_0^g$ to the good-type agent, excluding the bad-type agent. As shown in Figure 7(a), the principal prefers to induce bad-type agent's disclosure if the share of them in the prior is sufficiently high (λ is sufficiently low). This is depicted as $\hat{w}^b(f_2)$ in Figure 7(a), where $\lambda^p(f_2)$ marks the threshold in λ below which full the principal induces bad-type's disclosure.

Note that, $\hat{w}^b(f)$ increases with f in the interval $[V^b(\hat{w}_0^g), \bar{f}^b]$, and a larger share of bad-type agent (lower λ) is required to make it profitable for the principal to induce their disclosure. This suggests that the threshold in the disclosure cost that separates full-disclosure and partial disclosure equilibria, $\hat{f}(\lambda)$, is decreasing in the share of the good-type agent in the prior, λ. This threshold is depicted in Figure 7(b)[26].

When the disclosure cost exceeds $\bar{f}^b$, contracting with the bad-type agent becomes

[26]Equivalently, the threshold $\lambda^p(f)$ falls as f increases in this interval. This can be seen by comparing $\lambda^p(f_2)$ and $\lambda^p(f_3)$ in Figure 7(a).

unprofitable. The principal prefers a contract $\hat{w}^g(f)$ that will be accepted only by the good-type agent, as long as $f < \bar{f}^g$. As disclosure cost rises above $\bar{f}^g$, contracting with the good-type agent also becomes unprofitable, resulting in a non-disclosure equilibrium, as depicted in Figure 7(b).

B Omitted Proofs

Proof. of Lemma 1:

Let $\Delta Y \equiv Y_{12}(e) = Y_{vn} - Y_v - Y_n$,

and $\Delta B \equiv B_{12}(e) = B_{vn} - B_v - B_n$.

1 , 2) First consider the bad-type agent. For any incentive term w, since $e_n^b(w) = 0$, the solution to (2) for the bad type is $e_v^b(w) = w$. The principal's profit from offering w to the bad-type agent writes:

$$\pi^b(w) = w(Y_n - w),$$

which is maximized at $\hat{w}_0^b = \frac{Y_n}{2}$

Now consider the good-type agent. Assuming that the solution to (2) is interior and denoting it as $e^i = e^i(w) = (e_v^i(w), e_n^i(w))$, the first order condition of (2) writes:

$U_1^i(a) + w = 0$, and

$U_2^i(a) = 0$,

Let us denote $d \equiv -U_{12}^g = s - \Delta B$. Then the solution to the good-type agent problem writes:

$$e_v^g(w) = \frac{B_v + w - B_n d}{1 - d^2},$$
$$e_n^g(w) = \frac{B_n - (B_v + w)d}{1 - d^2},$$

The first order derivative of the principal objective function in (1) writes:

$$\frac{\partial \pi^i(w)}{\partial w} = Y_1(e^i)\frac{\partial e_v^i}{\partial w} + Y_2(e^i)\frac{\partial e_n^i}{\partial w} - w\frac{\partial e_v^i}{\partial w} - e_v^i(w)$$

Replacing for $\frac{\partial e_v^i}{\partial w}$ and $\frac{\partial e_n^i}{\partial w}$ yields:

$$\frac{\partial \pi^i(w)}{\partial w} = \frac{Y_1(e^i) - w - Y_2(e^i)d}{1 - d^2} - e_v^i(w)$$

The first order derivative of the principal's profit, $\frac{\partial \pi^g(w)}{\partial w}$, is positive at $w = 0$ iff

$$(Y_1(\tilde{e}^g) - \tilde{e}_v^g) - d(Y_2(\tilde{e}^g)) - \tilde{e}_v^g d) > 0.$$

Note that, $Y_1(e^g) = Y_v + \Delta Y e_n^g$ and $Y_2(e^g) = Y_n + \Delta Y e_v^g$. The solution to the principal's problem for the good-type agent's contract then writes:

$$\hat{w}_0^g = \frac{(Y_v - B_v) - d(Y_n - B_n) + \Delta Y \frac{(1+d^2)B_n - 2dB_v}{1-d^2}}{2(1 + \Delta Y \frac{d}{1-d^2})}$$

3) First assume that $\Delta B = \Delta Y = 0$. In that case $B_v > 0$ (by assumption 1) and $s > 0$ imply that $\hat{w}_0^g < \hat{w}_0^b$.

Second, assume that $\Delta Y = 0$, and ΔB is sufficiently large such that $d < 0$. In that case we can have $\hat{w}_0^g > \hat{w}_0^b$ if d is sufficiently large in absolute value, and $Y_n > B_n$.

Next, assume that $\Delta Y > 0$, and $\Delta B = 0$. In that case, we will have $\hat{w}_0^g < \hat{w}_0^b$ if s is large enough such that $e_n^g(\hat{w^b}) = 0$. If $e_n^g(\hat{w^b}) > 0$, then the principal can prefer a higher incentive term for the good-type agent $\hat{w}_0^g > \hat{w}_0^b$, for sufficiently large ΔY.

Thus, we can have $\hat{w}^g > \hat{w}^b$ only if ΔB or ΔY are sufficiently large.

□

Proof. of Proposition 1:

Consider the principal's optimal contract for the type i agent with $f = 0$, denoted as $\hat{w}_0^i$. Assume $\hat{w}_0^i > 0$. For $f \leq \underline{f}^i \equiv V^i(\hat{w}_0^i)$, this contract induces the disclosure of the agent and hence constitute an equilibrium.

Now assume that $f \geq \underline{f}^i$. If $\hat{w}_0^i > 0$, the principal can increase the share of the agent from the total surplus to induce his participation. Consider an incentive term x^i, such that $V^i(x^i) = f$. The principal can earn positive from offering this contract if $\pi^i(x) > 0$.

Let $\bar{w}^i$ denote the solution to $\pi^i(\bar{w}^i) = 0$. The principal can offer a contract $w^i = V^{i-1}(f)$ and gain positive profit if and only if $x^i < \bar{w}^i$. This implies that there exist a threshold $\bar{f}^i$ which is the solution to

$$\pi^i(V^{i-1}(f)) = 0,$$

above which contracting is not profitable for the principal.

It remains to show that this contract is strictly more profitable for the principal than satisfying agent's participation constraint with a lump-sum transfer t. The most profitable such contract must have the optimal incentive term $\hat{w}_0^i$, which yields a higher profit for the principal than any other incentive term. If a contract with an unconditional transfer t^i satisfies the agent's participation constraint with equality, then $V^i(\hat{w}_0^i) + t^i = f$. Principal's

profit form such contract then writes:

$$\pi^i(\hat{w}_0^i) - t^i = \pi^i(\hat{w}_0^i) - (f - V^i(\hat{w}_0^i)).$$

Principal's profit from a contract $x^i = V^{i^{-1}}(f)$ is higher than the contract with a lump-sum transfer if and only if

$$\pi^i(V^{i^{-1}}(f)) > \pi^i(\hat{w}_0^i) - (f - V^i(\hat{w}_0^i)).$$

Replacing f with $V^i(x^i)$ and rearranging, this condition can be written as

$$\pi^i(x^i) + V^i(x^i) > \pi^i(\hat{w}_0^i) + V^i(\hat{w}_0^i).$$

This means that, if the contract $x^i = V^{i^{-1}}(f)$ yields a higher total surplus (gross welfare) than $\hat{w}_0^i$, offering a lump-sum transfer to satisfy agent's participation constraint yields a lower profit than x^i. I establish this result in the proof of lemma 7, which states that the total surplus of contracting is increasing in the range $[\hat{w}^i, \bar{w}^i]$. Moreover, in the proof of lemma 2, I show that $V^i(w)$ is strictly increasing in w, which implies that $\hat{w}^i(f) = V^{i^{-1}}(f)$ is strictly increasing in f.

□

Proof. of Lemma 2:

For any incentive term w, the agent chooses his effort levels to maximize his value function: $V^i(w) = \text{Max}_e\{we_v + U(e)\}$. Derivation with respect to w and employing the envelope theorem yields: $\frac{dV^i(w)}{dw} = e_v^i(w)$.

From the proof of lemma 1, $\frac{\partial e_v^i(w)}{\partial w} > 0$. Hence $V^i(w)$ is strictly increasing and convex in w.

□

Proof. of Lemma 3:

From the proof of lemma 1,

$$\frac{\partial e_v^g(w)}{\partial w} = \frac{1}{1 - d^2},$$
$$\frac{\partial e_n^b(w)}{\partial w} = 1.$$

Hence, any incentive term w induces higher marginal effort from the good-type agent. Given that $\tilde{e}_v^g > \tilde{e}_v^b$, at any incentive term w, the good-type agent exerts higher effort e_v. Since, $\frac{dV^i(w)}{dw} = e_v^i(w)$, the good-type agent has a higher gain from any incentive term.

□

Proof. of Lemma 4:

a) The incentive compatibility constraints can hold if and only if there exist X^i such that:

$$V^b(\check{w}^g) - V^b(\check{w}^b) \leq \check{t}^b - \check{t}^g \leq V^g(\check{w}^g) - V^g(\check{w}^b).$$

Lemma 3 ensures that this condition is satisfied for any $\check{w}^g > \check{w}^b$.

b) Since the incentive compatibility constraint of the good-type agent can be satisfied without an unconditional transfer, $\check{t}^g = 0$, the transfer $\check{t}^b$ required to satisfy the incentive compatibility of the bad-type agent must satisfy

$$\check{t}^b \geq V^b(\check{w}^g) - V^b(\check{w}^b).$$

Consider contracts $(0, \check{w}^g)$ and $(\check{t}^b, \check{w}^b)$. Offering these contracts yields a higher profit for the principal compared to offering a pooling contract $w^p = \check{w}^g$ to both agent types if:

$$\pi^b(\check{w}^b) - \check{t}^b \geq \pi^b(\check{w}^g).$$

Replacing for $\check{t}^b$ and rearranging, yields:

$$\pi^b(\check{w}^b) + V^b(\check{w}^b) \geq \pi^b(\check{w}^g) + V^b(\check{w}^g)$$

lemma 7 shows that the total surplus of contracting, $\pi^b(w) + V^b(w)$, is increasing in the range $[\hat{w}^b, \bar{w}^b]$, which implies that the separating contracts yield a higher profit than pooling contracts if and only if $\check{w}^g > \bar{w}^b$. □

Proof. of Lemma 5:

1) By lemma 4, if $\hat{w}^g \leq \bar{w}^b$, pooling contracts are more profitable than a menu of separating contracts. In this case, the first order condition of principal's profit writes:

$$\frac{\partial[\sum_i m^i \pi(a^i(w))]}{\partial w} = \lambda \frac{\partial \pi^g(w)}{\partial w} + (1-\lambda)\frac{\partial \pi^b(w)}{\partial w}.$$

Note that $\frac{\partial \pi^g(w)}{\partial w}$ is positive and increasing in $(\hat{w}_0^b, \hat{w}_0^g)$ while $\frac{\partial \pi^g(w)}{\partial w}$ is negative and decreasing in this interval. Hence, the solution to the principal's problem, $\hat{w}^p$, goes from $\hat{w}_0^b$ at $\lambda = 0$ to $\hat{w}_0^g$ at $\lambda = 1$.

2) If $\hat{w}_0^g > \bar{w}^b$, separating contracts become more profitable than pooling if the principal is willing to offer $\check{w}^g > \bar{w}^b$ to the good-type agent. Consider the pooling contract $\hat{w}^p = \text{argmax}_w \{\sum_i m^i \pi^i(w)\} \in [\hat{w}_0^b, \hat{w}_0^b]$. Since this contract is increasing in λ, there exist a threshold λ_0^s for which $\hat{w}^p = \bar{w}^b$. For $\lambda > \lambda_0^s$, the principal maximizes her profit by choosing separating contracts $(0, w^g)$ and (t^b, w^b), where $t^b = V^b(w^g) - V^b(w^b)$, as characterized in lemma 4. The principal profit then writes:

$$\sum_i m^i \pi(e^i(X^i)) = \sum_i m^i \pi(e^i(w^i, 0)) - m^b t^b.$$

The first order condition with respect to w^b writes:

$$\frac{\partial \sum_i m^i \pi^b(w^b)}{\partial w^b} = (1-\lambda)(Y_1(e^b(w^b)) - w^b) = (1-\lambda)(Y_v - w^b) = 0.$$

Hence $\check{w}^b = Y_v = \bar{w}^b$.

The first order condition with respect to w^g writes:

$$\frac{\partial \sum_i m^i \pi^b(w^b)}{\partial w^g} = \lambda \frac{\pi^g(w^g)}{\partial w^g} - (1-\lambda)(\frac{\partial t^b}{\partial w^g}) = \lambda \frac{\pi^g(w^g)}{\partial w^g} - (1-\lambda)(e^b(w^g)) = 0.$$

As λ approaches to 1, the solution to this first order condition goes to $\hat{w}_0^g$. Therefore, $\check{w}^g(\lambda)$ increases from $\bar{w}^b$ at λ_0^s to $\hat{w}_0^g$ at $\lambda = 1$.

□

Proof. of Proposition 2 and 7:

a) Consider the case where $\hat{w}_0^g > \bar{w}^b$.

First assume $f < \bar{f}^b$. For $\lambda < \lambda_0^s$, if the pooling contract $\hat{w}_0^p$ induces the disclosure of the bad-type agent, the good-type agent also accepts the contract, since $V^g(w) > V^b(w)$ for any incentive term w. Hence $\hat{w}_0^p$ constitute an equilibrium if $f < V^b(\hat{w}_0^p)$.

For $f < \bar{f}^b$ contracting with the bad-type is profitable, but the principal must offer a contract with an incentive term equal to or above $\hat{w}^b(f)$ to induce the participation of the

bad-type. For $\lambda < \lambda_0^s$, the principal solves

$$\max_{\{w \in [\hat{w}^b(f), \bar{w}^b]\}} \sum_i m^i \pi^i(w)$$

Since $\frac{\partial \pi^b(w)}{\partial w}$ is negative at $\hat{w}^b(f)$, the probability of the good-type in the prior, λ, must be sufficiently high for the principal to offer a contract with a higher term. Hence, there exist a threshold $\lambda < \lambda_0^s$ below which the principal offers $\hat{w}^b(f)$ and above which she offers a pooling contract $\hat{w}_0^p$ which approaches $\bar{w}^b$ as λ goes to λ_0^s.

For $\lambda > \lambda_0^s$, as long as $f < \bar{f}^b$, the separating contracts induce the disclosure of both types because the bad-type's gain from the contract he receives equals $V^b(\check{w}^g)$ and $\check{w}^g > \bar{w}^b$. Hence for $f < \bar{f}^b$, a full disclosure equilibrium emerges for any λ.

Next assume $f \in [\bar{f}^b, \bar{f}^g]$. In that case contracting with the bad-type is no longer profitable for the principal. Define $\underline{w}^g(f)$ such that $V^b(\underline{w}^g(f)) = f$, which is the highest incentive term the principal can offer the good-type without attracting the bad-type agent. The principal can also offer the separating contracts $\check{X}^i(\lambda)$ such that $\check{w}^g(\lambda) > \underline{w}^g(f)$, if offering these contracts is more profitable than $\underline{w}^g(f)$:

$$\lambda \pi^g(\check{w}^g(\lambda)) + (1-\lambda)(\pi^b(\check{w}^b) - \check{t}^b(\lambda)) > \lambda \pi^g(\underline{w}^g(f)).$$

Note that since $\pi^b(\check{w}^b) - \check{t}^b(\lambda) < 0$, λ must be sufficiently high for the separating contract to be more profitable than the excluding contract $\underline{w}^g(f)$. Hence, there exist a threshold $\lambda^s(f)$ such that $\check{w}^g(\lambda^s(f)) = \underline{w}^g(f)$. For λ below this threshold, the principal offers $\underline{w}^g(f)$ inducing only the participation of the good-type, and above that she offers separating contracts $\check{X}^i(\lambda)$ such that $\check{w}^g(\lambda) > \underline{w}^g(f)$, resulting in full disclosure. Note that since $V^b(\check{w}^b) + \check{t}^b(w) = V^b(\check{w}^g)$, and $\check{w}^g(\lambda) > \underline{w}^g(f)$, the contract $\check{X}^b(\lambda)$ satisfies bad-type's participation constraint.

Define $\hat{f}(\lambda)$ as the inverse function of $\lambda^s(f)$. Since $\underline{w}^g(f)$ increases with f, $\lambda^s(f)$ also increases with f. This implies that the threshold in the disclosure cost $\hat{f}(\lambda) \in [\bar{f}^b, \bar{f}^g]$, above which a partial disclosure equilibria emerges, is increasing in λ.

Note that for $f \in [V^b(\hat{w}_0^g), \bar{f}^g]$, we have $\underline{w}^g(f) \geq \hat{w}_0^g$. This means that the principal can offer $\hat{w}_0^g$ to the good-type without attracting the bad type. Lastly, for $f > \bar{f}^g$, contracting with both types become unprofitable for the principal and hence a non-disclosure equi-

libria emerges.

b) Consider the case where $\hat{w}_0^g < \bar{w}^b$.

The contract $\hat{w}^p \in [\hat{w}_0^b, \hat{w}_0^g]$ characterized in lemma 5 induces the disclosure of both agent types if $f < V^b(\hat{w}_0^b)$.

First assume that $f \in [V^b(\hat{w}_0^b), \bar{f}^b]$. In this interval, the principal can gain positive profit from contracting with the bad-type by an incentive term $\hat{w}^b(f) \in (\hat{w}_0^b, \bar{w}^b]$ as characterized in proposition 1.

For $f < V^b(\hat{w}_0^g)$, we have $\hat{w}^b(f) < \hat{w}_0^g$. Hence, the principal solves:

$$\max_{\{w \in [\hat{w}^b(f), \hat{w}_0^g\}} \sum_i m^i \pi^i(w)$$

Since $\frac{\partial \pi^b(w)}{\partial w}$ is negative at $\hat{w}^b(f)$, the probability of the good-type in the prior, λ, must be sufficiently high for the principal to offer a contract with an incentive term above $\hat{w}^b(f)$. Hence, there exist a threshold in λ below which the principal offers $\hat{w}^b(f)$ and above which she offers the pooling contract $\hat{w}_0^p$ which approaches $\bar{w}^b$ as λ goes to λ_0^s. Hence, a full-disclosure equilibrium emerges in this interval.

Now assume $f = V^b(\hat{w}_0^g)$. At this disclosure cost the principal's optimal contract under observable types is identical for both types. Hence the principal offers $\hat{w}_0^g$ inducing both types participation.

For $f \in [V^b(\hat{w}_0^g), \bar{f}^b]$, the principal can earn positive profit from contracting with the bad-type by offering $\hat{w}^b(f) > \hat{w}_0^g$. In this case the principal is willing to offer a higher incentive term to the bad-type, because $V^g(w) > V^b(w)$ implies that the good-type's participation can be satisfied at a lower incentive term $\hat{w}^g(f) < \hat{w}^b(f)$. Hence, the principal must choose between offering $\hat{w}^b(f)$ to both types, or only attracting the good-type borrower with $\hat{w}^g(f)$. The principal induces both types' disclosure if

$$\lambda \pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f)) > \lambda \pi^g(\hat{w}^g(f)).$$

Since $\pi^g(\hat{w}^b(f)) < \pi^g(\hat{w}^g(f))$, principal prefers full-disclosure if λ is sufficiently low. Hence, there exist a threshold $\hat{\lambda}^p(f)$ below which the principal offers $\hat{w}^g(f)$ resulting in partial disclosure, and above which she offers $\hat{w}^b(f)$ inducing both types' disclosure.

Define $\hat{f}(\lambda)$ as the inverse function of $\hat{\lambda}^p(f)$. Since $\hat{w}^b(f)$ increases with f, $\pi^g(\hat{w}^b(f))$ and consequently $\lambda^s(f)$ decreases with f. This implies that the threshold in the disclosure cost $\hat{f}(\lambda) \in [V^b(\hat{w}_0^g), \bar{f}^b]$, above which a partial disclosure equilibria emerges, is decreasing in λ.

For $f \in [\bar{f}^b, \bar{f}^g]$, contracting with the bad-type becomes unprofitable. Since $\hat{w}^g(f) < \hat{w}^b(f)$, the principal can offer $\hat{w}^g(f)$ to the good-type without attracting the bad-type, resulting in partial disclosure.

Lastly, for $f > \bar{f}^g$, contracting with both agent types becomes unprofitable and a non-disclosure equilibrium emerges.

□

Proof. of Lemma 6:

By lemma 4, only pooling equilibria are feasible in this environment. The first order condition of principal's profit writes:

$$\frac{\partial[\sum_i m^i \pi^i(w)]}{\partial w} = \lambda \frac{\partial \pi^g(w)}{\partial w} + (1-\lambda)\frac{\partial \pi^b(w)}{\partial w}.$$

Note that $\frac{\partial \pi^b(w)}{\partial w}$ is positive and increasing in $(\hat{w}_0^g, \hat{w}_0^b)$ while $\frac{\partial \pi^g(w)}{\partial w}$ is negative and decreasing in this interval. Hence, the solution to the principal's problem, $\hat{w}^p$, goes from $\hat{w}_0^g$ at $\lambda = 1$ to $\hat{w}_0^g$ at $\lambda = 0$.

□

Proof. of Proposition 3

a) First assume that $\bar{f}^g < \bar{f}^b$.

Consider the pooling contract $\hat{w}_0^p \in [\hat{w}_0^g, \hat{w}_0^b]$ that is offered in equilibrium with $f = 0$. For $f < V^b(\hat{w}_0^g)$, this contract induces the participation of both types and hence constitutes an equilibrium.

For $f > V^b(\hat{w}_0^g)$, the contract $\hat{w}_0^p$ fails to satisfy the participation constraint of the bad-type agent if $V^b(\hat{w}_0^p) < f$. First assume that $f \in [V^b(\hat{w}_0^g), \bar{f}^g]$, so that the principal can earn positive profit from contracting with the good-type agent by an incentive term $\hat{w}^g(f) \geq \hat{w}_0^g$. In this case, if λ is such that $V^b(\hat{w}_0^p) < f$, the principal has two choices; either offering $\hat{w}^g(f)$ to the good-type agent excluding the bad-type agent, or offering $\hat{w}^b(f) > \hat{w}^g(f)$ and

induce both types' disclosure. The latter is more profitable for the principal if:

$$\lambda\pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f)) > \lambda\pi^g(\hat{w}^g(f)).$$

Since $\pi^g(\hat{w}^b(f)) < \pi^g(\hat{w}^g(f))$, we need to have λ sufficiently low for the principal to find $\hat{w}^b(f)$ more profitable. Hence, there exist a threshold $\lambda_1^p(f)$ below which the principal offers $\hat{w}^b(f)$ resulting in full disclosure, and above which she offers $\hat{w}^g(f)$ resulting in a partial disclosure equilibrium.

Now assume that $f \in [\bar{f}^g, \bar{f}^b]$, such that contracting with the good-type is no longer profitable, but the principal can earn positive profit by offering $\hat{w}^b(f)$ to the bad-type. Since the good-type agent accepts the contract $\hat{w}^b(f)$, the principal finds it profitable to offer this contract if:

$$\lambda\pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f)) > 0.$$

Note that, since $f > \bar{f}^g$, we have $\hat{w}^b(f) > \hat{w}^g(f)\bar{w}^g$, which implies that $\pi^g(\hat{w}^b(f)) < 0$. Hence, λ must be low enough for the principal to find offering $\hat{w}^b(f)$ profitable. Therefore there exist a threshold $\lambda_2^p(f)$ below which the principal offers $\hat{w}^b(f)$ resulting in full disclosure and above which she offers no contract resulting in non-disclosure.

Note that for f such that $\hat{w}^b(f) = \bar{w}^g$ which implies that $\pi^g(\hat{w}^b(f)) = 0$, the thresholds $\lambda_1^p(f)$ and $\lambda_2^p(f)$ coincide. Hence, we have a continuous threshold function $\lambda^p(f)$ such that for $f < \bar{f}^g$, we have $\lambda^p(f) = \lambda_1^p(f)$, and for $f > \bar{f}^g$, we have $\lambda^p(f) = \lambda_2^p(f)$.

Since $\hat{w}^b(f)$ increases with f, the threshold $\lambda^p(f)$ above which offering this contract is optimal is decreasing in λ. Define $\hat{f}(\lambda)$ as the inverse function of $\lambda^p(f)$. Since $\lambda^p(f)$ is decreasing in f, then $\hat{f}(\lambda)$ is decreasing in λ.

Lastly, for $f > bar f^b$, contracting with both agent types becomes unprofitable and the principal offers no contract, resulting in non-disclosure.

b) Assume that $\bar{f}^g > \bar{f}^b$.

The pooling contract $\hat{w}_0^p \in [\hat{w}_0^g, \hat{w}_0^b]$ that is offered in equilibrium with $f = 0$ induces the participation of both types and hence constitutes an equilibrium, if $f < V^b(\hat{w}_0^g)$.

For $f > V^b(\hat{w}_0^g)$, the contract $\hat{w}_0^p$ fails to satisfy the participation constraint of the bad-type agent if $V^b(\hat{w}_0^p) < f$. First assume that $f \in [V^b(\hat{w}_0^g), \bar{f}^b]$, so that the principal can earn positive profit from contracting with the bad-type agent by an incentive term $\hat{w}^b(f)$.

In this case, if λ is such that $V^b(\hat{w}_0^p) < f$, the principal has two choices; either offering $\hat{w}^g(f) \geq \hat{w}_0^g$ to the good-type agent excluding the bad-type agent, or offering $\hat{w}^b(f) > \hat{w}^g(f)$ and induce both types' disclosure. The latter is more profitable for the principal if:

$$\lambda\pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f)) > \lambda\pi^g(\hat{w}^g(f)).$$

Since $\pi^g(\hat{w}^b(f)) < \pi^g(\hat{w}^g(f))$, we need to have λ sufficiently low for the principal to find $\hat{w}^b(f)$ more profitable. Hence, there exist a threshold $\lambda^p(f)$ below which the principal offers $\hat{w}^b(f)$ resulting in full disclosure, and above which she offers $\hat{w}^g(f)$ resulting in a partial disclosure equilibrium.

Since $\hat{w}^b(f)$ increases with f, the threshold $\lambda^p(f)$ above which offering this contract is optimal is decreasing in λ. Define $\hat{f}(\lambda)$ as the inverse function of $\lambda^p(f)$. Since $\lambda^p(f)$ is decreasing in f, then $\hat{f}(\lambda)$ is decreasing in λ.

Next assume that $f \in [\bar{f}^b, \bar{f}^g]$. In this interval, contracting is no longer profitable with the bad-type, but the principal can earn positive profit profit by offering $\hat{w}^g(f)$ to the good-type. Since this contract does not attract the bad-type, the principal can offer it for any λ, resulting in partial disclosure.

Lastly, for $f > \bar{f}^g$, contracting with both agent types becomes unprofitable and the principal offers no contract, resulting in non-disclosure.

□

Proof. of Lemma 7:

Let w_{FB}^i denote the first incentive term that maximizes the contract efficiency with the type i agent:

$$w_{FB}^i = \text{argmax}_w\{Y(e^i(w)) + U^i(e^i(w))\},$$

where $e^i(w)$ is as given in the proof of lemma 1:

$$e^i(w) = \text{argmax}_w\{we_v^i(w) + U^i(e^i(w))\}.$$

Consider the incentive term w such that $w = \frac{Y(e^i(w))}{e_v^i(w)}$. Under this incentive term, the agent's objective function becomes identical to gross welfare. Hence, w_{FB}^i is the solution to:

$$w = \frac{Y(e^i(w))}{e_v^i(w)} \quad (\star)$$

1) For the bad-type agent, since $e_n^b(w) = 0$ for any w, we have:

$$w_{FB}^b = \frac{Y(e^i(w_{FB}^b))}{e_v^i(w_{FB}^b)} = Y_v.$$

Note that, at $w = Y_v$, the profit of the principal falls to zero, since he transfers all of the surplus to the agent. Therefore, for the bad-type agent, $w_{FB}^b = \bar{w}^b = Y_v > \hat{w}_0^b = \frac{Y_v}{2}$. This implies that the efficiency of the contract with the bad-type agent is increasing in the interval $[\hat{w}_0^b, \bar{w}^b]$.

2) For the good-type agent, the solution to $w = \frac{Y(e^g(w))}{e_v^g(w)}$ writes:

$$\frac{Y_v - dY_n + \Delta Y \frac{(1+d^2)B_n - 2dB_v}{1-d^2}}{1 + 2\Delta Y \frac{d}{1-d^2}},$$

which is strictly lower than $\hat{w}_0^g$ given in the proof of lemma 1.

Now consider the incentive term $\bar{w}^g$ at which $\pi^g(\bar{w}^g) = 0$. $\bar{w}^g$ is the solution to:

$$w = \frac{Y(e^g(w)) - Y(\tilde{e}^g)}{e_v^g(w)}.$$

Comparing to w_{FB}^g given by $(\star)$, it is straightforward that $\bar{w}^g < w_{FB}^g$ because $Y(\tilde{e}^g) > 0$, which is ensured by assumption 1. Only in the case where $Y_1(\tilde{e}^g) + dY_2(\tilde{e}^g) < 0$, we have $\hat{w}_0^g = \bar{w}^g = w_{FB}^g = 0$.

Hence, the efficiency of the contract with the good-type agent is increasing in the interval $[\hat{w}_0^g, \bar{w}^g]$, if $\hat{w}_0^g > 0$.

□

Proof. of Proposition 4:

1) As shown by proposition 1, under observable types, non-disclosure results when $f > \bar{f}^i = V^i(\bar{w}^i)$. Note that, since $\pi^i(\bar{w}^i) = 0$, in this cases the disclosure cost is higher than the maximum welfare achievable by a contract that yields non-negative profit for the principal; $f > GW^i(\bar{w}^i)$.

Mandatory disclosure results in the contract $\hat{w}_0^i$. Since, by lemma 7,

$$0 \leq GW^i(\hat{w}_0^i) < GW^i(\bar{w}^i) < f,$$

if $\hat{w}_0^i > 0$, mandatory disclosure creates gross welfare but results in negative net welfare.

2) By proposition 1, under voluntary disclosure, if $f \in [\hat{f}, \bar{f}]$, the principal offers an incentive term $\hat{w}^i(f) \in [\hat{w}_0^i, \bar{w}^i]$. In those cases, mandatory disclosure results in a contract with incentive term $\hat{w}_0^i < \hat{w}^i(f)$ which yields a lower gross and net welfare, since contract efficiency is increasing in $[\hat{w}_0^i, \bar{w}^i]$.

□

Proof. of Proposition 5:

1) As shown in proposition 2, under adverse selection type II, a non-disclosure equilibrium emerges when $f > \bar{f}^g > \bar{f}^b$. Under mandatory disclosure, the principal offers either a pooling contract or a menu of separating contracts that generate gross welfare. However, by proposition 4, since $f > GW^i(\bar{w}^i)$, the surplus generated by the contract(s) under mandatory disclosure is lower than the disclosure cost and hence result in negative net welfare.

2) To prove the result regarding gross welfare, I specify two cases for which mandatory disclosure results in higher or lower gross welfare.

Consider the case where $\hat{w}_0^g > \bar{w}^b$ and assume $f \in [\bar{f}^b, \bar{f}^g]$. As shown by proposition 2, for $\lambda < \lambda^s(f)$, the principal offers $\underline{w}^g(f)$ to the good-type, excluding the bad-type.

First suppose that $\lambda_0^s < \lambda < \lambda^s(f)$, so that mandatory disclosure results in separating contract $\check{w}_0^i(\lambda)$. The contract under mandatory disclosure results in higher gross welfare if:

$$\lambda GW^g(\check{w}_0^g(\lambda)) + (1-\lambda)GW^b(\bar{w}^b) > \lambda GW^g(\underline{w}^g(f)).$$

Assume that $Y_v \to 0$ so that $GW^b(\bar{w}^b) \to 0$. Note that we can have $\hat{w}_0^g > \hat{w}_0^b \to 0$, if ΔY is sufficiently high. In that case, since $\check{w}_0^g(\lambda) < \underline{w}^g(f)$, the contracts under mandatory disclosure results in lower gross welfare than the excluding contract under voluntary disclosure.

Next suppose that $\lambda < \lambda_0^s < \lambda^s(f)$, such that mandatory disclosure results in a pooling contract $\hat{w}_0^p$. The contract under mandatory disclosure results in higher gross welfare if:

$$\lambda GW^g(\hat{w}_0^p) + (1-\lambda)GW^b(\hat{w}_0^p) > \lambda GW^g(\underline{w}^g(f)).$$

Assume that $\lambda \to 0$. In that case, the right hand side approaches zero while the left hand

side is strictly positive for $Y_v > 0$. Hence, in this case, mandatory disclosure results in higher gross welfare.

Regarding net welfare, note that when $\hat{w}_0^g > \bar{w}^b$, partial disclosure results when $f > \bar{f}^b$, which implies that the contract with the bad-type results in negative net welfare. Since mandatory disclosure decreases the incentive term offered to the good-type and hence reduces net welfare of contracting with good type, the overall effect of mandating disclosure on net welfare is negative in these cases.

It remains to show that the effect of mandatory disclosure on net welfare is negative when $\hat{w}_0^g < \bar{w}^b$. By Proposition 7, in this case, a partial disclosure equilibrium results when:

$$\lambda \pi^g(\hat{w}^g(f)) > \lambda \pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f)),$$

and the principal offers $\hat{w}^g(f) \geq \hat{w}_0^g$. Under voluntary disclosure, the contract $w_0^p(\lambda)$ as characterized in Lemma 5 is offered in equilibrium where $\hat{w}_0^p(\lambda) < \hat{w}^g(f) < \hat{w}^b(f)$.

Note that since $\hat{w}^g(f) > \hat{w}_0^p(\lambda)$, we have $V^g(\hat{w}^g(f)) > V^g(\hat{w}_0^p(\lambda))$. So we can write:

$$\lambda[\pi^g(\hat{w}^g(f)) + V^g(\hat{w}^g(f)) - f] > \lambda[\pi^g(\hat{w}^b(f)) + V^g(\hat{w}_0^p(\lambda)) - f] + (1-\lambda)\pi^b\hat{w}^b(f)).$$

Moreover, note that since $V^b(\hat{w}^b(f)) = 0$, we have $GW^b(\hat{w}^b(f)) = \pi^b(\hat{w}^b(f))$. Since $\hat{w}_0^p(\lambda) < \hat{w}^b(f)$, we have $GW^b(\hat{w}_0^p(\lambda)) < GW^b(\hat{w}^b(f))$. Hence we can write:

$$\lambda[\pi^g(\hat{w}^g(f))+V^g(\hat{w}^g(f))-f] > \lambda[\pi^g(\hat{w}^b(f))+V^g(\hat{w}_0^p(\lambda))-f]+(1-\lambda)[\pi^b(\hat{w}^b(f))+V^b(\hat{w}^b(f))-f],$$

$$\lambda NW^g(\hat{w}^g(f)) > \lambda NW^g(\hat{w}_0^p(\lambda)) + (1-\lambda)NW^b(\hat{w}_0^p(\lambda)).$$

Therefore, in this case, net welfare is higher with the excluding contract $\hat{w}^g(f))$ under voluntary disclosure compared to the pooling contract $\hat{w}_0^p(\lambda)$ under mandatory disclosure.

3) If voluntary disclosure results in a full-disclosure equilibrium, mandating disclosure either does not change the equilibrium contract or decrease the incentive term for both agent types. By lemma 7, a lower incentive term in the interval $[\hat{w}^i, \bar{w}^i]$ generates lower gross welfare and consequently lower net welfare.

□

Proof. of Proposition 6:

1) First consider non-disclosure equilibria that emerge when $f > \max\{\bar{f}^b, \bar{f}^g\}$. In those

cases, mandatory disclosure results in a contract $\hat{w}_0^p$ which generates positive gross welfare. However, both contracts generate negative net welfare since $GW^i(\hat{w}_0^p) < GW^i(\bar{w}^i) > f$. Therefore, in this cases mandatory disclosure increases gross welfare but reduces net welfare.

Next, I specify cases where voluntary regime results in non-disclosure, but mandatory disclosure results in a contract with positive net welfare.

Consider the case $\bar{f}^b > \bar{f}^g$, and assume $\bar{f}^g < \hat{f}(\lambda) < f < \bar{f}^b$. By proposition 3, a non-disclosure equilibrium emerges in this case, if $\pi^g(\hat{w}^g(f)) < 0$, and:

$$\lambda\pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f)) < 0.$$

Mandatory disclosure results in a contract $\hat{w}_0^p \in [\hat{w}_0^g, \hat{w}_0^b]$. I specify conditions under which this contract creates positive net welfare.

First, note that since $f < \bar{f}^b$, net welfare of the contract $\hat{w}_0^p$ with the bad-type can be positive. While we have $V^b(\hat{w}_0^p) < f$, so that this contract does not induce the participation of the bad-type, we have $\pi^b(\hat{w}_0^p) > 0$. Hence, net welfare of this contract with the bad-type can be positive if $\pi^b(\hat{w}_0^p)$ is sufficiently large.

Second, note that since $w_{FB}^g > \bar{w}^g$, we can have a contract $w \in [\bar{w}^g, w_{FB}^g]$ such that $\pi^g(w) < 0$, but $GW^g(w) > 0$. If for such a contract, we have $V^g(w) > f$, it is sufficient to have $NW^g(w) > 0$. Hence, if the contract emerging under mandatory disclosure $\hat{w}_0^p$ lies in the interval $[\bar{w}^g, w_{FB}^g]$, it can result in positive net welfare when accepted by the good-type agent.

Therefore, if the following sufficient conditions hold together, voluntary disclosure results in non-disclosure but the contract under mandatory disclosure results in positive net welfare:

$$(i)\ \ \hat{w}_0^p \in [\bar{w}^g, w_{FB}^g],\ \ and\ \ V^g(\hat{w}_0^p) > f.$$

$$(ii)\ \ NW^b(\hat{w}_0^p)) > 0,$$

$$(iii)\ \ V^b(\hat{w}_0^p) < 0\ \ and\ \ \lambda\pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f)) < 0.$$

The following numerical example shows that these conditions can hold at the same time. Consider the following value of parameters:

$$Y_v = 0.5, Y_n = 0.2, Y_{vn} = 0.2 \quad B_v = 0.4, B_n = B_{vn} = 0, \quad \lambda = \frac{1}{2}$$

$$C(e,h) = \frac{1}{2}e_v^2 + \frac{1}{2}e_v e_n + \frac{1}{2}e_n^2$$

In autarky, we have $\tilde{e}_v^g = 0.4, \tilde{e}_n^g = \tilde{e}_v^b = \tilde{e}_n^b = 0$.

For any incentive term w, we have $e_v^g(w) = 0.4 + w, e_v^b(w) = w, e_n^g(w) = e_n^b(w) = 0$.

Under observable types, we have $\hat{w}_0^g = 0.05, \bar{w}^g = 0.1, w_{FB}^g = 0.5, \hat{w}_0^b = 0.25, \bar{w}^b = w_{FB}^b = 0.5$.

Under mandatory disclosure, the principal offers:

$$\hat{w}_0^p = \text{argmax}\{\frac{1}{2}(w+0.4)(0.5-w) + \frac{1}{2}w(0.5-w)\} = 0.15,$$

which results in $V^g(\hat{w}_0^p) \approx 0.7125, \pi^g(\hat{w}_0^p) \approx -0.0075, GW^g(\hat{w}_0^p) \approx 0.06375$ and $V^b(\hat{w}_0^p) \approx 0.01125, \pi^b(\hat{w}_0^p) \approx 0.0525, GW^b(\hat{w}_0^p) \approx 0.06375$. Also note we have $V^g(\bar{w}^g) = 0.005$.

Assume $f = 0.05$, such that $f > V^g(\bar{w}^g)$ implying that at this disclosure cost contracting with the good-type agent is no longer profitable for the principal, but $f < GW^g(\hat{w}_0^p) < V^g(\hat{w}_0^p)$ such that if $\hat{w}_0^p$ is offered, the good-type agent accepts it and it results in positive net welfare $NW^g(\hat{w}_0^p) \approx 0.0137$. Moreover, we have $NW^b(\hat{w}_0^p) \approx 0.0137$, so this contract results in positive net welfare if accepted by the bad-type agent. Hence, mandatory disclosure results in positive net welfare.

However, at $f = 0.05$ we have $V^b(\hat{w}_0^p) < f$, implying that this contract does not induce the disclosure of the bad-type under voluntary disclosure. To satisfy the participation constraint of the bad-type agent, the principal has to offer a contract at least as good as $\hat{w}^b(f) \approx 0.316$ for which $V^b(0.316) = 0.05$. If the principal offers $\hat{w}^b(f)$, her profit will be $\frac{1}{2}\pi^g(\hat{w}^b(f)) + \frac{1}{2}\pi^b(\hat{w}^b(f)) \approx -0.005 < 0$. Hence, in this numerical example, a non-disclosure equilibrium emerges under voluntary regime, while mandatory disclosure can result in positive gross and net welfare.

2) I specify cases where mandatory disclosure result in higher or lower gross and net welfare, when the voluntary regime results in partial disclosure.

Consider the contract under mandatory disclosure $\hat{w}_0^p \in [\hat{w}_0^g, \hat{w}_0^b]$. Assume that $\lambda \to 1$ so that $\hat{w}_0^p \to \hat{w}_0^g$. If $f \in [\hat{f}^g, \bar{f}^g]$, this contract does not induce the disclosure of the good-type agent, but the principal can earn positive profit by offering $\hat{w}^g(f) < \bar{w}_0^g$ to the good type agent. In this case, as shown in proposition 3, under voluntary disclosure for $\lambda < \lambda^p(f)$, the principal offers $\hat{w}^g(f)$, resulting in partial disclosure. Under mandatory

disclosure, both agents receive the contract $\hat{w}_0^p \to \hat{w}_0^g$. Since in this case $\hat{w}_0^p < \hat{w}^g(f)$, gross and net welfare of the contract with the good-type is lower under mandatory disclosure. Moreover, since $\lambda \to 1$, the welfare created by the contract with the bad-type agent is infinitesimally small. So in such a case, mandatory disclosure results in lower gross and net welfare.

Next, I specify cases where mandatory disclosure enhances gross and net welfare. For instance, consider the case where $V^b(\hat{w}_0^p) < f < V^g(\hat{w}_0^g < \bar{f}^b]$, so the principal must offer $\hat{w}^b(f) < \bar{w}^b$ to induce the the participation of the bad-type, while the good-type agent's participation constraint is satisfied by incentive terms equal or above $\hat{w}_0^g$. As characterized by proposition 3, there exist a threshold $\lambda^p(f)$ below which the principal offers $\hat{w}_0^g$, excluding the bad-type agent. In this case, the contract under mandatory disclosure has a higher incentive term than the one under voluntary disclosure; $\hat{w}_0^p) > \hat{w}_0^g$. If $\hat{w}_0^p < w_{FB}^g$ (which holds true for $\lambda \to 1$), the increase in the incentive term by mandating disclosure enhances gross welfare of the contract with the good-type agent, while creating welfare from contracting with the bad-type. Hence, in this case, mandatory disclosure improves gross welfare compared to voluntary disclosure.

In general, in cases where voluntary disclosure results in partial disclosure with a contract $\hat{w}^g(f)$ offered to the good type, if $\hat{w}_0^p) > \hat{w}^g(f)$, gross welfare increases by a disclosure mandate. In such cases, although $V^b(\hat{w}_0^p) < f$, we can have $\pi^b(\hat{w}_0^p)$ sufficiently large such that $NW^b(\hat{w}_0^p)) > 0$. Hence, the sufficient condition for voluntary regime to result in partial disclosure and mandatory disclosure to enhance net welfare is:

$$(i) \quad V^b(\hat{w}_0^p) < f, \quad and, \quad \lambda\pi^g(\hat{w}^g(f)) > \lambda\pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f))$$

$$(ii) \quad \hat{w}_0^p \in [\hat{w}^g(f), w_{FB}^g]$$

$$(iii) \quad NW^b(\hat{w}_0^p)) > 0$$

The following numerical example shows that these conditions can hold true simultaneously. Consider the following parameters:

$$Y_v = 0.5, Y_n = 0.2, Y_{vn} = 0.2 \quad B_v = 0.3, B_n = B_{vn} = 0, \quad \lambda = \frac{1}{2}$$

$$C(e,h) = \frac{1}{2}e_v^2 + \frac{1}{2}e_v e_n + \frac{1}{2}e_n^2$$

In autarky, we have $\tilde{e}_v^g = 0.3, \tilde{e}_n^g = \tilde{e}_n^b = \tilde{e}_n^b = 0$.

For any incentive term w, we have $e_v^g(w) = 0.3 + w, e_v^b(w) = w, e_n^g(w) = e_n^b(w) = 0$.

Under observable types, we have $\hat{w}_0^g = 0.1, \bar{w}^g = 0.2, w_{FB}^g = 0.5, \hat{w}_0^b = 0.25, \bar{w}^b = w_{FB}^b = 0.5$.

Under mandatory disclosure, the principal offers:

$$\hat{w}_0^p = \text{argmax}\{\frac{1}{2}(w+0.3)(0.5-w) + \frac{1}{2}w(0.5-w)\} = 0.175,$$

which results in $V^g(\hat{w}_0^p) \approx 0.0678, \pi^g(\hat{w}_0^p) \approx 0.00437, GW^g(\hat{w}_0^p) \approx 0.07217$ and $V^b(\hat{w}_0^p) \approx 0.01531, \pi^b(\hat{w}_0^p) = 0.05687, GW^b(\hat{w}_0^p) \approx 0.07217$.

Assume that $f = 0.06$, such that while $NW^i(\hat{w}_0^p) > 0, \forall i \in \{b, g\}$, we have $V^b(\hat{w}_0^p) < V^g(\hat{w}_0^p) < 0$, meaning that this contract does not induce neither types' disclosure at this cost. The principal can earn positive profit from good-type by offering $\hat{w}^g(f) = 0.1582 < \bar{w}^g$ such that $V^g(\hat{w}^g(f)) = f$. She can also earn positive profit by offering $\hat{w}^b(f) = 0.3464 < \bar{w}^b$ such that $V^b(\hat{w}^b(f)) = f$.

If the principal offers $\hat{w}^g(f)$, her profit will be $\lambda\pi^g(\hat{w}^g(f)) \approx 0.0033$. If she offers $\hat{w}^b(f)$, her profit will be $\lambda\pi^g(\hat{w}^b(f)) + (1-\lambda)\pi^b(\hat{w}^b(f)) \approx 0.00125$. Therefore, under voluntary disclosure, the principal prefers to offer $\hat{w}^g(f)$, resulting in partial disclosure.

Note that, since $\hat{w}_0^p > \hat{w}^g(f)$, and $\hat{w}_0^p < w_{FB}^g$, gross welfare is higher under mandatory disclosure. Also, $NW^b(\hat{w}_0^p)) = GW^b(\hat{w}_0^p)) - f \approx 0.07217 - 0.06 > 0$, net welfare is also higher under mandatory disclosure. Since $V^g(\hat{w}^g(f)) = f$, net welfare under voluntary disclosure is $\lambda NW^g(\hat{w}^g(f)) = \lambda\pi^g(\hat{w}^g(f)) \approx 0.0033$. Under mandatory disclosure, net welfare increases to $\lambda GW^g(\hat{w}_0^p) + (1-\lambda)GW^b(\hat{w}_0^p) - f \approx 0.01217$. Thus, in this numerical example, where a partial disclosure equilibrium emerges under voluntary regime, mandatory disclosure can result in higher gross and net welfare.

3) If voluntary disclosure results in a full-disclosure equilibrium, mandating disclosure either does not change the equilibrium contract or decrease the incentive term for both agent types. By lemma 7, a lower incentive term in the interval $[\hat{w}^i, \bar{w}^i]$ generates lower gross welfare and consequently lower net welfare.

□