AI and the Changing Nature of Journalism

Abhishek Ray[*] Cesar Martinelli[†]

March 4, 2024

Abstract

The growing use of generative AI in journalism has highlighted major quality assurance concerns such as editorial oversight and AI use transparency. Yet, how oversight and transparency will incentivize limited or assistive AI use is unclear. We address this concern by proposing a novel principal-agent framework incorporating oversight and transparency to incentivize assistive AI usage. Unlike popular discourse, our analysis shows that AI usage screening is less effective than manual oversight of AI-produced content. We also demonstrate that transparency with readers about AI use is ineffective in encouraging consumption from news outlets using AI assistively rather than generatively. We further analyze the implications of costly transaction costs, realizable property rights, occasional shirking, and responsible AI use. Our insights on encouraging assistive AI use apply to all these situations, but we show how each aspect relates to policy issues. Finally, we demonstrate the trust vs. accountability dilemma and contribute to the current regulation debate on using AI in journalism.

Keywords: journalism, generative ai, agency theory, game theory

[*]Costello College of Business, George Mason University
[†]Department of Economics, George Mason University

1 Introduction

As artificial intelligence (AI) usage grows in the institutional press (Tobitt, 2023; Guaglione, 2023), journalism is increasingly under scrutiny. Increasing generative AI use has raised content originality issues, with readers questioning news reports' authorship and whether reporters appropriately gathered and reported facts (Economist, 2024). However, limited (assistive) AI use for editing and fine-tuning that allows journalists to investigate, write, and report has garnered less attention (Broussard et al., 2019). Such limited use might benefit the profession and democracy, which relies on informed voters. Yet, while stakeholders voice accountability (e.g., oversight, transparency) demands (Guess, 2024), it is unclear how enforceable these demands will be.

The major problem underlying enforceable AI usage norms is the widespread accessibility and affordability of generative AI tools (Túñez-López et al., 2021). Commercial news organizations, which have been losing income, find this ubiquity appealing. AI is now used to write news stories, saving firms money on journalist expenses (Nel, 2023). For example, News Corp executive chair Michael Miller disclosed that a tiny team used AI to produce 3,000 weekly articles in June. The four-person team checks for "hallucinations" or false information an LLM model develops when it fails to foresee an acceptable input response (Hurst, 2023). Such tools may improve productivity, but the same principles can be used for misleading or false journalism. For example, NewsGuard has uncovered hundreds of AI-generated fake news sites in multiple languages since April, generating millions in ad revenues (Yaseen, 2023).

The reaction to such increasing AI usage has been swift but varied (ICIJ, 2023). According to ethics researchers, one extreme has been to advocate a total ban on generative AI usage. More mainstream media organizations led by experienced journalists have called for accountability measures promoting limited (or assistive) use – e.g., human oversight of AI-produced content and training journalists on AI usage (Cools and Diakopoulos, 2023). Some organizations (e.g., Newsquest Media Group) have even started hiring 'AI-assisted reporters' adept at "utilizing AI technology to create national, local, and hyper-local content for our news brands, while also applying their traditional journalism skills" (Hurst, 2023). Overall, the public wants AI usage transparency by labeling human- and AI-produced articles.

Given this background, the fundamental question is: *Can limited or assistive AI use be fostered in journalism where generative AI can save costs, thus improving profits?* The answer is unclear since information asymmetries exist in a journalist's relationship to the news organization (e.g., editor, manager) and the news organization's relationship to the reader (Napoli, 1997; Fengler and Ruß-Mohl, 2008). News organizations (or principals) typically supervise journalists (agents) to ensure their content production supports the ownership's ideological or economic goals. Yet, journalists may be granted some leeway because the principal lacks direct supervision resources, and monitoring is not always costless. Besides, misaligned journalist-editor goals in boosting content attention may also hinder content production.

Conversely, consumers (principals) rely on news organizations (agents) for relevant and reliable information. However, news organizations face a trade-off between maximizing revenue vs. investing in information acquisition. Pandering to the consumer's perceptions (e.g., using framing) may attract more attention, thereby increasing readership and revenues, but at the cost of accuracy in reporting Fengler and Russ-Mohl (2008). Inaccurate or biased reporting may impact the reader's awareness, translating into sub-optimal decisions at the voting booth. Given recent evidence – e.g., 49% of news organizations said they use generative AI tools for their journalists – coupled with the dwindling trust of consumers in the news, the trust gap between consumers and news organizations has only widened in recent years (Shapiro, 2023).

We answer this fundamental question using agency theory. Specifically, using a novel principal-agent class of models incorporating trust and accountability measures (i.e., oversight and transparency), we examine two categories of models distinguished by their principal-agent dynamics: (1) *Model 1*: Principal (Manager/Editor), Agent (Journalist); (2) *Model 2*: Principal (Public), Agent (Journalist). In these models, agents possess private types contingent on their way of technology utilization (i.e., generative or assistive AI usage) for journalistic content generation. Given hidden types and actions, the analysis focuses on cases where the principal may incentivize the agent to a specific technology usage type instead of another.

Accountability through oversight (Model 1) and transparency (Model 2) as contractual mechanisms distinguish our approach from the literature. The equilibrium analysis reveals several important insights. Model 1 shows that encouraging assistive AI usage requires setting incen-

tives on effort and AI use type. Moreover, regardless of AI use screening, AI-related content should always have editorial oversight. Model 2 shows that regardless of transparency, a news organization using assistive rather than generative AI will gain greater public attention if it spends a comparable amount of time on non-AI regular journalistic tasks as a generative AI-using news organization. We then discuss several essential model extensions and a trust vs. accountability dilemma. We demonstrate that journalists' use of assistive AI may not always increase public attention. Our insights contribute to the growing discussion on AI journalism regulations.

We structure the paper as follows. Section 2 lays out the contribution of our work to extant literature. Section 3 explains our overall framework, and Subsections 3.1 and 3.2 explain how each principal-agent relationship is structured. Section 4 establishes the insights and provides intuition using the model analysis and evidence from the field. We then relax assumptions in our base model in Section 5 to explore further enriching insights in line with Model 1 and 2 analysis. Then, we draw on our overall research to provide policy recommendations in Section 6 to demonstrate the fundamental trust vs. accountability dilemma in AI-flavored journalism. Finally, we conclude with how to extend our work and test our insights in future studies.

2 Related Literature

Our research is related to two main areas of study: the increasing use of AI and machine learning (ML) tools in journalism and the analysis of news organizations from an agency-theoretic perspective. The following subsections will provide an overview of how our research contributes to these fields.

2.1 AI Usage in Journalism

Using AI, news organizations are automating specific news reports (Reuters, 2023; Shapiro, 2023). AI generally organizes and delivers news in human-readable ways (Yaseen, 2023). The algorithm employed usually processes enormous amounts of data, chooses pre-programmed article structures, positions crucial points, and inserts names, statistics, and figures (Guess, 2024). AI speeds up and scales news output, supports journalists, and enables optimization and personalization. Yet, it cannot develop news sources and regional contacts to verify the claim, as

journalists do (Diakopoulos, 2020; ICIJ, 2023). Thus, scholarship and practice should study human-centered AI in journalism by understanding journalists' roles and functions in hybrid workflows, how journalists and AI can interact and collaborate, and how these new technologies may affect human autonomy, agency, and labor economics (CNTI, 2023).

Our model fills the gap in understanding how news organizations should contract journalists in a hybrid AI work environment. Through Models 1 and 2 (Section 3), we address the contractual nature of how journalists and news organizations can be incentivized to use AI in an assistive rather than a generative manner. Further, through our unifying extensions in Section 6, we show that when we consider relaxing assumptions related to how journalists use AI, our insights hold and, in some cases, provide an improved understanding of the hybrid journalist-AI work environment.

2.2 Principal Agent in Journalism

Information asymmetry is critical in principal-agent relationships (Laffont, 2003). Both parties to a market transaction want a deal. They usually need more information regarding the other's preferences. The agent—usually the subordinate or seller—knows more about the product or service than the principal, usually the boss or client. This imbalance may cause mistrust and, in some cases, no contracts (Laffont and Maskin, 1981; Martimort and Laffont, 2009). Many media transactions have similar information asymmetries, even when journalists and sources have no contract (Napoli, 1997; Fengler and Ruß-Mohl, 2008). For instance, using public relations professionals or other source material as agents, journalists might launch an investigation as principals. In turn, the news organizations as principals may engage journalists as agents to investigate an issue. Finally, the public as principals rely on news organizations as agents for important news that determines their public choices (Fengler and Russ-Mohl, 2008).

We use this fundamental framework to build our principal-agent analysis. Specifically, we consider the journalists–news organizations–the public chain of relationships and consider two principal-agent models. One model considers the news organization as the principal and the journalist as the agent. In contrast, the other considers the public the principal and the news organization the agent. In doing so, we address the primary problem with burgeoning AI use in journalism: how increasing AI use changes how news organizations use journalists to improve

efficiency while maintaining trustworthy relationships with the public. Our model insights add to the ongoing debate on regulating AI use in journalism (ICIJ, 2023; CNTI, 2023).

3 Model Preliminaries

We begin by specifying features of our modeling framework (refer Figure 1) that are common across the two models. Then, in Subsections 3.1 and 3.2, we delve into the context-specific features of each model. Table 1 lists all the variables used across both models.

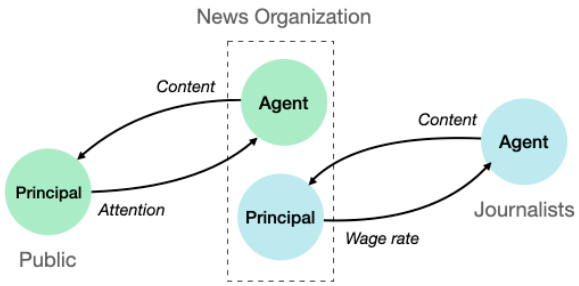


Figure 1: Principal Agent Framework combining the public, news organizations, and journalists.

Each agent (news organization or journalist) in the framework is characterized using a type $i \in \{S, NS\}$, where $S =$ shirker while $NS =$ non-shirker regarding AI usage. Training in AI models for journalistic work is becoming more common, sometimes even required (Oh, 2018; Jamil, 2021). Previous research suggests that skill upgrades may lead to employee shirking and varying skill levels (Carlin, 1989; Carmichael, 1989; Eguchi, 2004). Therefore, we consider two agent types, characterized using the proportion of effort $\theta_i \in (0, 1), i \in \{S, NS\}$ dedicated to journalistic duties and responsibilities that do not involve using AI.

Agents (θ_S or shirkers) who spend less time on journalistic work and more on AI are considered to use it for generative purposes and have low marginal production costs. For instance, earlier generative AI use by journalists in news organizations (e.g., Gannett, AP) resulted in more articles produced but with more mistakes and copyright issues (Guaglione, 2023). In contrast, agents who spend more time on journalistic work and less time using AI (θ_{NS} or non-shirkers) are considered to use assistively and have higher marginal production costs. For instance, some news organizations (e.g., The Atlantic, National Public Radio) have limited the AI use of journalists to assistive purposes using guidelines (Nelson, 2023). The principal knows the distribution ($\mu_i : \sum_{i \in \{S,NS\}} \mu_i = 1$) of agent types, optimizes the expected payoff

under individual rationality and incentive compatibility constraints. The principal further sets accountability measures, given whether the agent will exert the desired effort–limited AI use.

Table 1: Summary of model variables used across Models 1 and 2.

Model Preliminaries	
Parameter	**Description**
$i \in \{S, NS\}$	Index for agent type. $S =$ Shirker, $NS =$Non-Shirker.
$\theta_i \in (0, 1)$	Proportional allocation to non-AI work.
$\mu_i \in (0, 1)$	Principal's beliefs regarding agent type.
Model 1	
$\tilde{p}_i \in \mathbb{R}_{\geq 0}$	Full information first-best level wage rates.
$p_i^\circ \in \mathbb{R}_{\geq 0}$	Pooled wage rates for all journalist types.
$p_i \in \mathbb{R}_{\geq 0}$	Contractual per-unit effort wage rates under adverse selection issues.
$q_i > 0$	Quality of content produced by journalist.
$R(q_i)$	Revenue from content of quality q_i.
$e_i > 0$	Agent effort differentiated by type.
$c_i \in \mathbb{R}_{\geq 0}$	Agent marginal cost of production differentiated by type.
Model 2	
$v_a \in \mathbb{R}_{\geq 0}$	Public's value from information that aligns with their views.
$v_f \in \mathbb{R}_{\geq 0}$	Public's value from information that conflicts with their views.
$\tilde{x}_i > 0$	Full information attention paid to the agent, differentiated by type.
$x_i > 0$	Attention paid to the agent, differentiated by type, under asymmetric information.
γ_i	Probability of producing content that aligns with the public's view.
$a_H, b_H \in (0, 1)$	Public's valuation of content produced without transparency that aligns with their views.
$a_F, b_F \in (0, 1)$	Public's valuation of content produced without transparency that conflicts with their views.
$g_S, g_{NS} \in (0, 1)$	Public's valuation of content produced with transparency that aligns with their views.
$r_i \in \mathbb{R}_{\geq 0}$	Marginal revenue from attention for agent.
$k_i \in \mathbb{R}_{\geq 0}$	Marginal cost of producing content for agent.

Trust and accountability are essential to offering agents contracts. Mills and Reiss (2017) state that formal effort or output control is enforceable based on a contract or institutional rules. In contrast, trust is based on generally accepted social norms and motivates norm-conforming behavior through observation and informal acknowledgment. Lewicki and Brinsfield (2015) and Ostrom (2010) further show that trust allows people to make intuitive judgments and evaluations based on few basic norms or rules, reducing the need for accountability through effortful monitoring and frequent re-analysis of a situation or relationship. When trust and goal commitment are high, external monitoring is unnecessary and costly.

Accordingly, we interpret accountability and trust in the framework's context as follows (similar to e.g., Vanacker and Belmas, 2009). First, in both models, we represent trust when

the principal uses no accountability measure. Second, in Model 1, we consider the principal's accountability measure using (*editorial*) oversight. The principal ensures that if the agent uses generative AI and is suspected to be of type θ_S, the product is edited and corrected to remove all inaccurate information. Finally, in Model 2, we consider the principal's accountability measure using transparency. That is, the principal ensures that the agent discloses AI usage. Such disclosures impact consumption payoffs, so the decision to be transparent is non-trivial. The following subsections provide a contextual description of each feature and context-specific variables.

3.1 Model 1: News Organization–Journalist

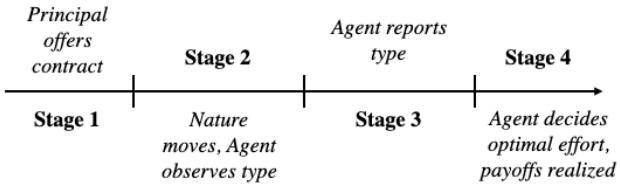


Figure 2: The Basic Principal-Agent Model for News Organizations (the Principal) and Journalists (the Agent)

In Stage 1, the principal announces a contract – with or without AI usage oversight – consisting of a per-unit effort wage rate $p_i, i \in \{S, NS\}$ for each agent type. The principal aims to maximize their return $R(q_i)$, linear in quality, by screening for agent types given that the agent's reported types distribution is known. However, the returns depend on the quality of the content produced by the agent (journalist). Specifically, an agent type S produces content of quality $q_S = \theta_S e_S - (1 - \theta_S)e_S$ while an agent type NS produces content of quality $q_{NS} \equiv e_{NS} = \theta_{NS}e_{NS} + (1 - \theta_{NS})e_{NS}$. These characterizations presume that type S agent's AI effort decreases the quality of non-AI tasks. Type NS's AI efforts complement non-AI efforts in producing quality content q_{NS}.

Each agent type has a marginal cost c_i of efforts spent doing non-AI work. Going by industry evidence (Pavlik, 2023; Tobitt, 2023), we assume the marginal cost of production for agent S is lesser than the agent type NS, $c_S < c_{NS}$. In setting wage rates for the two types, the principal considers incentivizing truthful reporting by setting incentive compatibility (IC) constraints. Consider the agent type S paid p_S, typified by effort allocation θ_S to non-AI work and the marginal cost of effort c_S. The IC constraint should be such that S should have no incentive

to deviate and change their allocation to θ_{NS} and get paid p_{NS} at the same marginal cost of c_S. Similarly, consider an agent type NS, as typified by the allocation θ_{NS} to non-AI work and the marginal cost of effort c_{NS}. The IC constraint should ensure they have no incentive to deviate to $i = S$, change their allocation to θ_S, and get paid p_S at the same marginal cost of c_{NS}.

The overall Model 1 problem can be represented using the following maximization problem.

$$\begin{aligned} \underset{p_S, p_{NS}}{\text{maximize}} \ \pi_P &= (\mu_S)\left(R(q_S) - p_S e_S\right) + (\mu_{NS})\left(R(q_{NS}) - p_{NS} e_{NS}\right) \\ \text{s.t. } \pi_i(p_i, \theta_i) &= p_i e_i - c_i(\theta_i e_i)^2 \geq 0, i \in \{S, NS\} \quad (\textbf{IR}), \\ \pi_i(p_i, \theta_i) &\geq \pi_i(p_{-i}, \theta_{-i}), i, -i \in \{S, NS\} \quad (\textbf{IC}), \\ e_i &\in \mathbb{R}_{\geq 0} \in \text{argmax}(\pi_i(p_i, \theta_i)). \end{aligned} \tag{1}$$

The revenue $R(q_S)$ changes given oversight implementation. With oversight, the principal improves content quality by editing out poor-quality information, implying $R(q_S) = q_S + (1 - \theta_S)e_S$. Without oversight, $R(q_S) = q_S$; but, $R(q_{NS}) = q_{NS}$ either way. The equilibrium analyzes contracts with or without screening or oversight, resulting in four combinations. While it can be argued that screening is also a way of ensuring accountability with no trust, we consider trust limited to cases without oversight measures, as per journalistic editorial practices (Cools and Diakopoulos, 2023).

3.2 Model 2: Public–News Organization

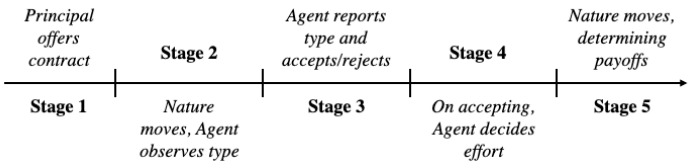


Figure 3: The Basic Principal-Agent Model for News Organizations (the Agent) and Public (the Principal)

This model highlights how the public consumes news content based on expected utility, considering conflicting or aligning reportage. Despite news organizations using AI, the public is often unaware, leading to decreased content quality on disclosing AI use (Pavlik, 2023). Thus, enforcing transparency of AI use in content creation may lower content consumption value (Longoni et al., 2022; Scire, 2023). This is the basis for the following structure that com-

pares public consumption to news organization profit maximization. Overall, the public derives positive value from the news because it values the information that aligns with their views v_a and information that conflicts with their views v_f, such that $v = v_a - v_f \geq 0$. Without transparency, when the agent allocates $\theta_S e_S$ (or, $\theta_{NS} e_{NS}$) proportion of effort to non-AI journalistic work, then with probability γ_S (or, γ_{NS}), produced content aligns with the principal's views and provides a utility of a_H (or, b_H such that $a_H > b_H$), thus increasing the principal's utility to $v_a + a_H - v_f$ (or, $v_a + b_H - v_f$). Conversely, with probability $1 - \gamma_S$ (or, $1 - \gamma_{NS}$), produced content conflicts with the principal's views and provides disutility of a_F (or, b_F such that $a_F > b_F$), thus decreasing the principal's utility as $v_a - v_f - a_F$ (or, $v_a - v_f - b_F$).

With transparency, the principal is more inclined to reject exaggerated claims by the agent (Longoni et al., 2022). Indeed, news consumers have upped their demands for transparency (Scire, 2023), resulting in a few news organizations disclosing generative AI usage in some articles. We theorize that with transparency, the principal has an attenuated impact of news that aligns with their views. Specifically, when the agent exerts effort θ_S (or, θ_{NS}), then with probability γ_S (or, γ_{NS}), the produced content aligns with the principal's views and provides a utility of $g_S < a_H$ (or, $g_{NS} < b_H$), thus increasing the principal's utility as $v_a + g_S - v_f$ (or, $v_a + g_{NS} - v_f$). Conversely, with probability $1 - \gamma_S$ (or $1 - \gamma_{NS}$), the produced content conflicts with the principal's views. Still, the principal derives no disutility from this since the generative AI transparency gives the principal a reason to reject any claims that do not align with their views. Thus, the principal's utility is $v_a - v_f$.

The principal decides the attention allocation, based on agent types likelihood μ_i, and their effort e_i to provide content that may or may not coincide with the principal's views. Attention is considered a scarce, costly, transferable commodity and the currency in the news consumption context (Myllylahti, 2020). This incentive to the agent is used to produce content at a marginal cost that differs given the agent type. Without loss of generality, we assume marginal costs k_i of exerting non-AI effort, such that $k_S < k_{NS}$. Further, the agents can monetize the attention received, depending on their type. That is, the marginal revenue from attention is r_i such that $r_S < r_{NS}$. Evidence suggests generative AI content creators engage consumers differently (Shapiro, 2023). Generative AI content creators have low referral traffic from associated

sources, resulting in low ad earnings (Brewster et al., 2023). Indeed, AdExchanger and NewsGuard have warned content providers to stop using generative AI-generated content to improve revenues (Vargas, 2023).

The overall Model 2 problem can be represented using the following maximization problem.

$$
\begin{aligned}
&\underset{p_S, p_{NS}}{\text{maximize}}\ \hat{\pi}_P = (\mu_S)\left(\mathbb{E}_{\gamma_S} U(v_a, v_f, a_H, a_F, g_S, e_S)\right) + (\mu_{NS})\left(\mathbb{E}_{\gamma_{NS}} U(v_a, v_f, b_H, b_F, g_{NS}, e_{NS})\right) \\
&\text{s.t. } \pi_{\theta_S}(a_H, a_F, g_S, r_S) = \begin{cases} (\gamma_S a_H + (1-\gamma_S)a_F)x_S r_S - k_S \theta_S e_S \geq 0, & \textbf{IR} \text{ no transparency} \\ (\gamma_S g_S)x_S r_S - k_S \theta_S e_S \geq 0, & \textbf{IR} \text{ with transparency} \end{cases} \\
&\qquad \pi_{\theta_{NS}}(b_H, b_F, g_{NS}, r_{NS}) = \begin{cases} (\gamma_{NS} b_H + (1-\gamma_{NS})b_F)x_{NS} r_{NS} - k_{NS} \theta_{NS} e_{NS} \geq 0, & \textbf{IR} \text{ no transparency} \\ (\gamma_{NS} g_{NS})x_{NS} r_{NS} - k_{NS} \theta_{NS} e_{NS} \geq 0, & \textbf{IR} \text{ with transparency} \end{cases} \\
&\qquad \pi_{\theta_S}(a_H, a_F, r_S) \geq \pi_{\theta_{NS}}(a_H, a_F, r_{NS}), (\textbf{IC} \text{ no transparency}) \\
&\qquad \pi_{\theta_S}(g_S, r_S) \geq \pi_{\theta_{NS}}(g_S, r_{NS}), (\textbf{IC} \text{ with transparency}) \\
&\qquad \pi_{\theta_{NS}}(b_H, b_F, r_{NS}) \geq \pi_{\theta_S}(b_H, b_F, r_S), (\textbf{IC} \text{ no transparency}), \\
&\qquad \pi_{\theta_{NS}}(g_{NS}, r_{NS}) \geq \pi_{\theta_S}(g_{NS}, r_S), (\textbf{IC} \text{ with transparency}), \\
&\qquad e_i \in \mathbb{R}_{\geq 0} \in \text{argmax}(\pi_i(p_i, \theta_i)), i \in \{S, NS\}.
\end{aligned}
\tag{2}
$$

where the utilities for the principal change, given transparency. Without transparency, the following are the utilities when the principal pays attention to different agent types:

$$
U(v_a, v_f, a_H, a_F, b_H, b_F) = \begin{cases} (\gamma_S(v_a + a_H - v_f) + (1-\gamma_S)(v_a - v_f - a_F))e_S - x_S^2, & \text{Type} : \theta_S \\ (\gamma_{NS}(v_a + b_H - v_f) + (1-\gamma_{NS})(v_a - v_f - b_F))e_{NS} - x_{NS}^2, & \text{Type} : \theta_{NS} \end{cases} \tag{3}
$$

With transparency, the following are the utilities:

$$
U(v_a, v_f, g_S, g_{NS}) = \begin{cases} (\gamma_S(v_a + g_S - v_f) + (1-\gamma_S)(v_a - v_f))e_S - x_S^2, & \text{Type} : \theta_S \\ (\gamma_{NS}(v_a + g_{NS} - v_f) + (1-\gamma_{NS})(v_a - v_f))e_{NS} - x_{NS}^2, & \text{Type} : \theta_{NS} \end{cases} \tag{4}
$$

The equilibrium analyzes contracts with or without transparency. Marginal revenue (i.e., r_i) and marginal costs (i.e., k_i) play critical roles in determining how attention benefits the agent. Consequently, this model analyzes how to ensure the agent contracts with limited rather than generative AI usage, given that payoffs may differ.

The following equilibrium analysis begins by setting the complete information benchmark for each model to contrast how imperfect information affects contractual decisions. Subsequently, the analysis reveals insights regarding contracts with and without accountability measures and how accountability incorporation impacts the principal's payoff.

4 Equilibrium Analysis

Standard principal-agent problem analysis starts with the full information benchmark (first-best contract). First-best contracts are efficient and allow the principal to contract with the agent without uncertainty. Adverse selection and moral hazard make principal-agent contracts unpredictable and ineffective. The equilibrium analysis of each model will be explained in subsections 4.1 and 4.2 using this background and relevant values.

4.1 Model 1: News Organization–Journalist

The following lemma establishes the full-information benchmark (first-best contract) for Model 1. We denote the first-best wage rates with screening as $\tilde{p}_i, i \in \{S, NS\}$.

Lemma 1. *First-best contractual wage rate for an agent of type $i \in \{S, NS\}$ is as follows. For agent type NS, $\tilde{p}_{NS} = \frac{1}{2}$ with or without oversight. For agent type S, with oversight $\tilde{p}_S = \frac{\theta_S}{2}$ when $\theta_S \in (0, \theta_{NS})$. However, without oversight, contracting with type S is possible only when $\theta_S \in (\frac{1}{2}, \theta_{NS})$ such that $\tilde{p}_S = \theta_S - \frac{1}{2}$.*

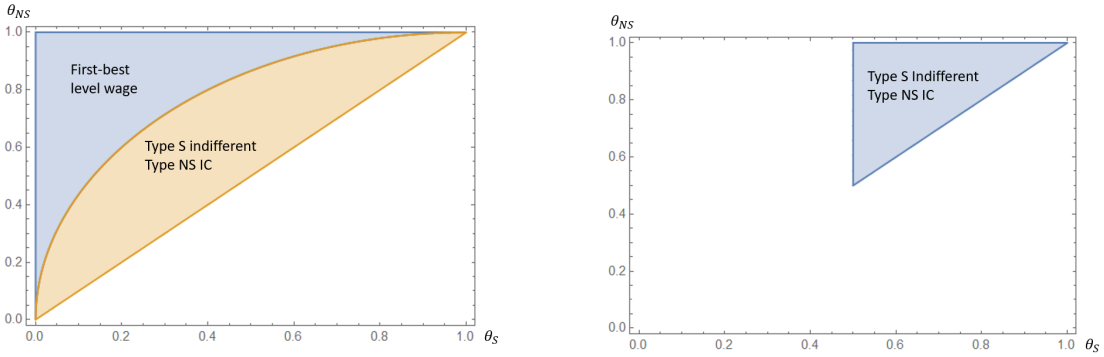
The proof is in Appendix B. First-best contracts allow the principal to set wage rates for each agent type, maximizing return based on the agent's optimum payoff. Under oversight measures, the principal keeps the type S agent's wages proportional to their revealed type (viz., non-AI effort). Without oversight, the principal contracts only when both agents exert non-AI efforts higher than $\frac{1}{2}$ such that $\theta_{NS} > \theta_S > \frac{1}{2}$. Essentially, the principal maximizes its payoff by considering oversight an essential recourse to prevent revenue loss and contracts with agent type NS in all circumstances, paying a higher wage than type S. However, without oversight, the principal contracts with agent type S only when the latter guarantees some minimum non-AI effort.

However, optimal contracts in the full information scenario are seldom feasible. The ambiguity about the agent types results in the principal aiming to maximize an expected return rather than a certain one. In addition, maximizing returns is subject to the agent's individual rational-

ity and incentive compatibility constraint. Thus, the principal sets 'second-best' contracts that satisfy both conditions. The following characterizes the optimal wage rates under the adverse selection problem of agent types $i \in \{S, NS\}$.

Proposition 1. *Contracts that ensure incentive compatibility and first-best level wages are possible only with oversight and when* $\theta_{NS} \in (\sqrt{(2-\theta_S)\theta_S}, 1)$. *Otherwise, type* S *agents are only weakly incentivized while type* NS *agents find it incentive-compatible to reveal their true type in contracts that are either: (1) with oversight and when* $\theta_{NS} \in (\theta_S, \sqrt{(2-\theta_S)\theta_S})$, *or (2) without oversight and when* $\theta_S \in \left(\frac{1}{2}, \theta_{NS}\right)$.

The proof, along with complete expressions for $p_i^*, i \in \{S, NS\}$ for each case, are in Appendix C. Proposition 1 has two key takeaways: first, even under adverse selection issues, under certain conditions, the principal can offer contracts with first-best level wages with oversight to agents of either type. Typically, with adverse selection, the principle cannot condition compensation on effort without accounting for agent type. Hence, the first best is usually not possible. Second, without oversight, the principal has no choice but to offer contracts that require type S agents to exert a minimum non-AI effort at the cost of making them weakly incentivized to reveal their true type. In contrast, type NS agents are offered contracts with higher wage rates than type S. Given the adverse selection, these insights show that incentivizing assistive AI usage indirectly depends on managing content production using accountability measures.



(a) Contract with Oversight

(b) Contract without Oversight

Figure 4: Optimal contracts with screening for Model 1. Shaded regions show the values of $\theta_i, i \in \{S, NS\}$ for which contracts are possible with adverse selection issues.

Illustrative principal-agent problems in contingent delegation (Gan et al., 2023) or contingent contracts (Deb and Mishra, 2014) suggest dominant-strategy contracts when agent payoff

depends on both reported types and principal payoff. Principals favor coordination, and agents prioritize adaptation in such coordinated decision-making scenarios. In the context of Model 1, the principal-agent relationship is similar in light of AI usage for content production. Non-AI work proportions provide supplementary signals regarding the agent's output quality. Thus, in line with the informativeness principle (Holmström, 2017), Proposition 1 insights make a case for discriminatory pay based on AI usage, given that the principal is not guaranteed strong incentive compatibility. Yet, there are two essential caveats: (1) the level of pay is contingent on how much the agent's efforts may harm the principal's returns; (2) the principal has a guaranteed amount of effort when there is no recourse to insuring against agent's sub-par output.

These insights answer two crucial issues concerning AI usage and journalism. First, Serdouk and Bessam (2023) shows that the economic aspect of AI adoption suggests boosting profitability and effectiveness while reducing costs (including staff wages). Our insights suggest a more nuanced approach. Screening for AI usage and editorial oversight is critical to ensuring journalists are incentivized to use AI in an assistive rather than a generative manner. Contrary to recent calls for automated systems for editing and content supervision (Chin-Rothmann, 2023), quality oversight is crucial to incentivize the effective use of AI. Second, Moran and Shaikh (2022) and Broussard et al. (2019) suggest that intelligent bots and human journalists can be integrative rather than competitive. Newsrooms are often good places to learn tech skills before moving to jobs that pay better. For instance, the American Society of News Editors (ASNE) says that 43% of full-time editors left U.S. newsrooms between 2008 and 2017 (Grieco, 2020). Yet, preserving content distinctiveness by rewarding more journalistic labor and less AI-related work may incentivize talent retention in news organizations.

Nonetheless, screening for AI usage involves background checks, resume reviews, and re-evaluation post-training in AI use. These mechanisms may need time and, in some cases, may prove costly (Bartling et al., 2012). Moreover, news organizations may have pay scales that pool journalists according to their responsibilities or job descriptions (Fu, 2023). In such cases, oversight is the organization's only recourse against generative AI usage that may deteriorate content quality. The following proposition lays out pooled wage rates $p_i^\circ, i \in \{S, NS\}$ for journalists of either type without screening.

Proposition 2. *Regardless of oversight, agent-type NS wages are always lower than first-best levels when the principal does not screen. However, for agent type S, with oversight, $p_S^\circ > \tilde{p}_S,\ \forall\ \theta_S \in (0, \theta_{NS})$. Similarly, without oversight, $p_S^\circ > \tilde{p}_S$ only when $\mu \in \left(0, \frac{c_S}{c_{NS}+c_S}\right)$ and $\theta_S \in \left(\frac{\theta_{NS} c_{NS} \mu}{2\theta_{NS} c_{NS} \mu + c_S(1-\mu)}, \theta_{NS}\right)$.*

The proof along with complete expressions for $p_i^\circ, i \in \{S, NS\}$ are in Appendix D. Combining the wages of all journalists, regardless of their usage of AI, organizations may incentivize generative AI usage by paying above first-best levels. Specifically, even with oversight, the principal can over-compensate agent type S for generative AI usage. Without oversight, the principle compensates agent type S at levels greater than the first-best level only if the principal believes the agent uses AI less for generative and more for assistive purposes, even if the agent's AI usage is sure to exceed some threshold. Thus, screening and oversight result in appropriate compensation for journalistic efforts, while oversight may not be helpful without screening.

The standard principal-agent theory advocates limiting agent discretion as much as possible (Bartling et al., 2012). Otherwise, the agent may exploit its freedom to shirk or need expensive monetary incentives to operate in the principal's interest. Indeed, Huang and Cappelli (2010) study screening vs. monitoring in jobs where meticulousness and work ethic screening is costly. They show that screening can attract more dedicated workers who need less supervision and performance incentives, which may conflict with employee monitoring. Proposition 2 reflects some of this intuition in the context of journalism, which involves a certain work ethic that may conflict with generative AI usage (Pavlik, 2023). Under certain conditions, combining wages without screening may fail to incentivize the required level of diligence, particularly in assistive over generative AI use.

These insights address the following key issues concerning pay equity and the advent of AI in journalism. Labor unions in journalism and other industries are mobilizing to safeguard their jobs from AI (Fu, 2023). They are negotiating contracts with explicit language about generative AI use. However, given how rapidly AI tools and the nature of reporting evolve, incentivizing desired AI usage are critical to any such contracts. Christophe Deloire, secretary general of Reporters Without Borders, has noted that AI poses considerable challenges to journalism and media, yet incentives to utilize these tools ethically and prudently are lacking (ICIJ, 2023).

We add to this discussion through Propositions 1 and 2. Specifically, Proposition 1 shows that screening without oversight for AI usage patterns may not be enough to incentivize assistive over generative AI usage. Conversely, Proposition 2 shows that oversight may be risky without screening since generative AI usage may be incentivized through over-compensation of type S agents (journalists).

Nevertheless, the principal's decision to use oversight can depend on whether screening can be considered a substitute. For instance, if the principal screens for AI usage and improves hiring of more NS type agents, oversight may not be required. Indeed, editorial oversight is standard in most news organizations, but AI-usage oversight would add quality assurance checks (Cools and Diakopoulos, 2023). Thus, knowing when to utilize oversight is critical for leading news organizations trying to cut costs. The following proposition addresses this issue.

Proposition 3. *Regardless of screening, the principal should always use oversight.*

The proof is provided in Appendix E. Proposition 3 essentially states that oversight is the *dominant* strategy of the principal, regardless of screening. Evidence from news organizations bears out this insight. Recently established guidelines in several news organizations emphasize substantial human involvement and supervision in AI use, including further editing and fact-checking of outputs before publication. News organizations reject replacing journalists with machines and emphasize human decision-making while employing generative AI techniques. For instance, Swedish and Norwegian news outlets Aftonbladet and VG, respectively, mandate that "a human has reviewed all material published" (Schori, 2023), and "all use of generative AI must be manually approved before publication," (VG, 2023).

More prominent organizations, such as Reuters, view oversight as seeking "meaningful human involvement, and to develop and deploy AI products and use data in a manner that treats people fairly" (Reuters, 2023). The Guardian's standards require generative AI to be linked to a "specific benefit and the explicit permission of a senior editor" (Viner and Bateson, 2023). Overall, the trend towards inserting more human intervention in monitoring AI-incorporated content production is becoming standard practice in news organizations worldwide, further establishing the unique nature of journalism as a profession and AI's impact on it. Our analysis and the Proposition 3 result further reinforces this guideline.

However, a key takeaway from Proposition 3, combined with earlier insights, is that oversight incorporation results in different implications for wages received by both agent types. As a result, when the principal screens but forgoes oversight, the type NS agent may accept lower wages than first-best levels while the type S agent is overcompensated. Conversely, when the principal uses screening and oversight, the type NS agent receives higher wages than type S, and the latter almost always receives lower wages than before. Thus, compensating agents can be challenging. A well-compensated type NS agent can boost productivity, while an overcompensated type S agent can impact content quality. Therefore, making an informed decision based on skills, experience, and performance is crucial.

Given that AI usage can affect news organizations' relationships with journalists, several news organizations' AI principles also require transparency with readers. We establish insights along this aspect in the following subsection 4.2 using Model 2.

4.2 Model 2: Public–News Organization

The following lemma establishes the full-information benchmark (first-best) for Model 2. Note that here, *contract* implies the amount of attention the principal (public) pays to the agent (news organization), which in turn incentivizes the agent to exert desired *type* of effort.

Lemma 2. *The first-best attention level for an agent of type $i \in \{S, NS\}$ is such that the agent is indifferent to participation. For agent type i, following characterizes first-best attention-levels without transparency:*

$$\tilde{x}_i = \begin{cases} \frac{r_S}{2\theta_S k_S}\left(v(a_H\gamma_S + a_F(1-\gamma_S)) + (a_H\gamma_S)^2 - (a_F(1-\gamma_S))^2\right), & i = S, \\ \frac{r_{NS}}{2\theta_{NS} k_{NS}}\left(v(b_H\gamma_{NS} + b_F(1-\gamma_{NS})) + (b_H\gamma_{NS})^2 - (b_F(1-\gamma_{NS}))^2\right), & i = NS \end{cases}$$

while the first-best attention-levels with transparency is $\tilde{x}_i = \frac{r_i}{2\theta_i k_i}\left(g_i\gamma_i(g_i\gamma_i + v)\right)$.

The proof is in Appendix F. The key observation from Lemma 2 is that attention to an agent type is impacted by the proportion of non-AI effort by the agent (θ_i) and a convex combination of utility and disutility due to the content. The utility is influenced by the likelihood of it aligning with the principal's views. Variations in this likelihood alter the level of attention given to different types of agents, *ceteris paribus*. Note that without transparency, there is a greater

range of utility vs. disutility ($\{a_H, a_F\}$ or $\{b_H, b_F\}$) that the public enjoys from the content compared to with transparency ($\{g_S, g_{NS}\}$). Hence, without transparency, the principal might pay greater attention to the source, producing content more aligned with the principal's views. Therefore, as explained in the following proposition, this essential trade-off between likelihood of view alignment and utility vs. disutility ranges impacts attention to the different agent types.

Proposition 4. *Under asymmetric information regarding agent types, first-best attention levels are impossible. Further, regardless of transparency, even when $r_{NS} > r_S$, agent type NS is guaranteed greater attention than S only when $\theta_S \in \left(\theta_{NS}\left(\frac{r_S}{r_{NS}}\right), \theta_{NS}\right)$. Otherwise, agent type S always receives greater attention than NS.*

The proofs with complete expressions of all attention levels are in Appendix G. Proposition 4 has several essential insights. First, recall that attention is a costly and scarce commodity *paying* attention to one agent over another is non-trivial. In Model 2, paying attention also considers the agent type – the proportion of efforts to non-AI work. Since the principal does not have perfect information on this aspect, the utility enjoyed when allocating attention to one type can change when the agent employs more effort toward AI vs. non-AI work. Second, observe that Proposition 4 shows that when agent types S has non-AI proportion of efforts closer to type NS, essentially *leveling* the field for the principal to choose, the principal always goes with NS. Since agent NS gives more balanced information than S, this choice of NS dominates.

More generally, Proposition 4 suggests that when a risk-averse principal faces asymmetric information and risk-neutral agents, equilibrium contracts involve the *risky* agent type being indifferent, weakly incentive-compatible to participation and the other being incentivized to reveal their type truthfully. These contracts reduce risk from allocating attention to agents with more varying outputs that may be detrimental to the principal. With costly attention to allocate and asymmetric information, the principal hopes to give as little attention as possible and gain the maximum expected benefit from the agent's output. Yet, the output from type S agent has the maximum variability and expected utility from content. Thus, the only parameters that could change such allocations would be the type θ_i and the alignment likelihoods γ_i.

Finally, note that regardless of transparency, only the proportion θ_i of effort to non-AI work impacts attention levels to NS. This finding reflects survey data (Hofeditz and Mirbabaie, 2021)

and adds to the more extensive debate around AI transparency (Nel, 2023; Subramaniam, 2023). The conflict between labeling risk and reader preconceptions about AI muddles the transparency decision. Publishers can be more reluctant to label AI-generated news text, given the likelihood of a complete boycott or backlash by readers (Fu, 2023). Conversely, readers might need to be more uniformly knowledgeable about AI's capabilities. Therefore, the fundamental question for news readers and organizations is the discernible difference between human and generative AI writing. Public AI literacy must drive transparency beyond simple labels (Diakopoulos, 2020), but achieving such literacy is complex with scarce, costly attention.

Given the established relationships between news organizations, journalists, and the public, we now delve into some additional cases that the base models do not consider. In the following Section 5, we explore some specific features which, when added to the models, reveal further insights that tie into the overall policy implications discussed in Section 6.

5 Extensions

This section considers relaxing relevant assumptions and adding structure to Models 1 and 2. The overall theme of these modifications is two-fold: testing whether our insights hold up under such relaxations and whether typical institutional factors modify the equilibrium contracts across both models.

5.1 Costly Transaction Costs

In Model 1, we consider the principal capable of using oversight without any costs. But, from an institutional perspective, oversight is similar to an accountability mechanism with certain costs associated with its enforcement (North, 1986). We classify these costs as the transaction costs of the principal in Model 1. Specifically, suppose that when the principal uses oversight to improve the quality of output by $(1-\theta_S)e_S$, the costs incurred are $k((1-\theta_S)e_S)^2$, where $k > 0$ is the marginal cost of oversight quality assurance. The following lemma characterizes the contractual wages.

Lemma 3. *Contracts are possible only when* $\theta_{NS} \in (\sqrt{1-(1-\theta_S)^2}, 1)$. *Incentive compatibility and first-best wage rates are possible* $k \in (0, \frac{\theta_S c_S \mu\left(\theta_{NS}^2-(2-\theta_S)\theta_S\right)}{(1-\theta_S)^2})$. *Otherwise, type* S *is only weakly incentivized, while type* NS *finds it incentive-compatible to participate.*

The proof with the expressions for wage rates is provided in Appendix H.1. Transaction costs limit the space of contracts offered to both agent types. In particular, because such expenses exist only when the principal uses oversight, the proportion $\theta_i, i \in \{S, NS\}$ of non-AI work that the principal contracts with is now more restricted than before (Lemma 1). One interpretation of this result is that the principal 'hedges' losses due to costly oversight by ensuring some minimum amount of non-AI work the type NS agent can do. The principal needs a guarantee of at least $\sqrt{1-(1-\theta_S)^2}$ level of non-AI work by type NS and offers first-best level wages only when the marginal cost of oversight is lower than some finite threshold and can be as low as cost-less. Nonetheless, if this cost is guaranteed to be at least some finite value, the principal does not offer first-best level wage but instead provides a contract that is enough to incentivize the type NS to reveal their type truthfully. The type S is indifferent to truthful revelation but gets enough incentive to participate.

Still, how do such transaction costs impact the decision to use oversight? The following proposition considers this question.

Proposition 5. *When transaction costs are low, the principal should use oversight. However, when transaction costs are high, oversight can be used only when the type S agent has low marginal production costs, $c_S \in \left(0, \frac{2\theta_S}{\theta_{NS}^2+\theta_S^2}\right)$.*

The proof is in Appendix H.2. Proposition 5 provides a contrast to the earlier Proposition 3. Essentially, without transaction costs, the principal was always better off using oversight, regardless of screening. But, with transaction costs, oversight usage is possible only when such costs are low or when the agent type S has low marginal costs. The latter indicates the principal safeguards against spurious generative AI content that an agent type S may generate quickly and burden quality assurance tasks. Conversely, when the type S agent has higher marginal production costs, the principal is better off not employing oversight. In such situations, since first-best incentive-compatible contracts are not possible, the principal settles for offering contracts that make the type S agent indifferent. In contrast, type NS finds it incentive-compatible to participate (Lemma 3).

The ramifications of Proposition 5 highlight the conflict between human oversight and AI safeguards in news organizations. Many news organizations utilize AI tools to report and eval-

uate news quickly and compete in the information marketplace (Yaseen, 2023). Yet, due to the rapid evolution of generative AI tools and the spread of misinformation, news organizations must be able to adapt to these developments for effective manual oversight (Nahmias and Perel, 2021). For instance, malicious actors can hack fact-checking tools to validate fake news, creating a risk from news consumption (Opdahl et al., 2023). Thus, tool-based oversight on such aspects may strain fact-checkers and journalists, resulting in insufficient time for manual assessment of claims and sources. Besides, tools-based evaluation of user-generated content on social media platforms is challenging because images are often renamed, re-compressed, resized, and even cropped, reducing accuracy and blurring the fine-grained distinctions that verification tools rely on.

Still, it can be argued that once the cost of using AI gets factored into the agent's efforts, the contracts can become more equitable. That is, if the agent executes journalistic duties and uses AI *responsibly*, the principal may have lower risks in contracting with either agent type. The following extension tackles this issue.

5.2 Responsible AI Usage

Model 1 focuses solely on the costs incurred by the agent for activities that do not include AI use. Nevertheless, responsible AI use involves ensuring AI system outputs are sufficiently deliberated on before use in any reporting (Buhmann and Fieseler, 2021). In light of such practices, it is unclear if the principal contracts with(out) oversight or screening measures. The following proposition sheds light on this issue.

Proposition 6. *Contracts are impossible when the principal screens agents who responsibly use AI. The principal can instead contract without screening and always use oversight.*

The proof with complete derivations of all expressions is in Appendix H.3. Proposition 6 offers the crucial insight that screening does not result in a contract when principals are faced with contracting with journalists with known responsible AI use habits. The agent accepts no wage rates based on its non-AI working proportion of efforts and considers remuneration based on the entire effort. Thus, contracts that *pool* agents together for the same wage rates result in successful contracts. Yet, our analysis indicates that oversight remains an important contractual

feature.

Proposition 6 also connects with the more recent guidelines set by journalist groups and select news organizations on using generative AI responsibly (Yaseen, 2023). Specifically, our analysis correctly points out that even when journalists might use AI responsibly and conduct the necessary evaluations, oversight remains an essential additional step before publishing the article online or otherwise. Indeed, supervising generative AI content to ensure it fits the news organization's standards and goals is imperative since an editorial process is only sometimes enough, according to The Center for News, Technology and Innovation report (CNTI, 2023).

However, one aspect of the journalist's efforts towards creating content is the propensity to shirk vs. use AI responsibly. Until now, Model 1 assumed that once an agent is a shirker (θ_S), that always determines the output. Moreover, Model 1 assumes that shirking is probabilistic, with no agent-specific intentions to shirk captured in their decision-making. The following extension handles relaxing both these assumptions.

5.3 Occasional Shirking

In the occasional shirking scenario, we analyze a case where an agent, with probability λ shirks devotes θ_S proportion of effort towards non-AI work. For occasional shirking, Model 1 changes in two ways. First, agent types are no longer relevant since the agent is of one type (NS) but has a non-zero likelihood of devoting $\theta_S < \theta_{NS}$ proportion of non-AI work, thereby devoting different overall efforts. Essentially, the agent's payoff is now of the following type.

$$\pi_{NS}(p, \theta_i) = \begin{cases} p_S e_S - c_{NS}((\lambda\theta_S + (1-\lambda)\theta_{NS})e_S)^2, & \text{when occasionally shirking,} \\ p_{NS}e_{NS} - c_{NS}(\theta_{NS}e_{NS})^2, & \text{otherwise.} \end{cases}$$

Second, the principal maximizes payoff subject to incentivizing the desired effort e_{NS} without any incentive compatibility. This is because shirking, assumed to be a stochastic rather than deterministic occurrence, cannot be reported by the agent. The principal can only incentivize the efforts that discourage shirking while encouraging non-shirking efforts (similar to moral hazard scenarios). The following proposition describes equilibrium contracts for this case.

Proposition 7. *When the agent types S occasionally shirks with probability $\lambda \in (0, 1)$, the*

principal can offer first-best contracts with screening that pay type S lower than type NS, regardless of oversight.

The proof is in Appendix H.4. Proposition 7 reflects the critical insight from Proposition 1 – AI usage has the potential of introducing pay disparities, given the potential impact of shirking on revenues. Notably, echoing Oh (2018), Proposition 7 shows that while pay disparity existed earlier, technical acumen related to AI skills can further exacerbate such disparities. Despite cost-cutting efforts, a newsroom cannot live without its writer-reporters. Yet data reporters, designers, and developers are often neglected. With the burgeoning AI usage, journalists are expected to undertake various technical training programs[1] that equip them with the basics of AI use. Yet, most journalists are skeptical of the potential of AI making their jobs easier vs. usurping their positions in the organization. Thus, news organizations can incentivize journalists to undertake such training programs with higher pay, even when shirking may be possible. Understanding how such technology works will help journalists recognize fakes and fight disinformation from spreading, as a survey of 52 news organizations shows (Merrefield, 2023).

We now turn to Model 2, and within that model, we consider making transparency disclosure part of the subscription fee. Only when the principal pays does the agent disclose AI use. The following subsection illustrates how the equilibrium contract changes given this modification.

5.4 Realizable Property Rights

We explore the issue of journalists selling their content on independent publishing platforms. Journalists are increasingly using their reporting resources and capabilities and, instead of using them for news organizations, are establishing their property rights to access. Substack is one example of such a movement. Property rights are typically set using some access fee or price. While Model 2 does not have any such base fee, we consider a specific question in this context, motivated by data transparency issues in online content consumption: *does pricing the disclosure of AI use improve attention allocation for a risk-averse principal with costly attention*? To answer this, we modify Model 2 and introduce a fee w_i for the agent $i \in \{S, NS\}$ when AI use is disclosed, $\pi_{\theta_i} = (\gamma_i g_i) x_i r_i + w_i - k_i \theta_i e_i$.

Introducing a flat fee for transparency concerns introduces an additional source of revenue

[1] See, e.g., `https://www.lse.ac.uk/media-and-communications/polis/JournalismAI/Training`

for the agent, given the loss in attention they might expect from such disclosures. The agent still sets effort to be indifferent to participation, but the attention levels impact whether the agent contracts at first or second-best levels. The following proposition establishes the equilibrium contract for this case.

Proposition 8. *When transparency disclosure comes with a price, first- or second-best levels of attention to both agent types are possible only when* $\theta_S \in \left(\theta_{NS}\left(\frac{r_S}{r_{NS}}\right), \theta_{NS}\right)$. *Further, if* $k_S \in \left(\hat{k}, k_{NS}\right)$, $v \in (1, \bar{v})$ *and* $w_{NS} \in (0, \hat{w})$ *then first-best levels of attention are possible.*

The proof with the complete expressions for $\hat{k}, \hat{w}, \bar{v}$ is in Appendix H.5. Observe that in contrast to Proposition 4, in this case, transparency price makes it possible to have the first-best attention levels to both agent types. However, such levels are only possible when agents S and NS have a similar proportion of non-AI work, along with high marginal production costs of agent type S and a price for transparency for NS with an upper limit. Suppose the agent type S incurs higher costs in content production and type NS cannot overcharge for transparency disclosure. In that case, the principal can allocate the first-best attention levels to both agent types. Essentially, the price for transparency becomes the way to recover lost revenue for agent types. At the same time, limits on such fees give the principal means to allocate attention that maximizes its payoff.

Nevertheless, with a price for transparency, both agent types are guaranteed some non-negative attention levels when $\theta_S \in \left(\theta_{NS}\left(\frac{r_S}{r_{NS}}\right), \theta_{NS}\right)$. One of them is the first-best, but the other is the second-best level when $\theta_S \in \left(\theta_{NS}\left(\frac{r_S}{r_{NS}}\right), \theta_{NS}\right)$ but $k_S < \hat{k}$ or when $w_{NS} > \hat{w}$, such that exactly one of the agent types is weakly incentivized to reveal their true type. Transparency disclosure at a price may motivate the principal to contract with both agent types when their production costs or transparency price differs significantly but not their payoffs, even under equal non-AI work proportions. The converse scenario, where non-AI work proportions are significantly different ($\theta_S \in \left(0, \theta_{NS}\left(\frac{r_S}{r_{NS}}\right)\right)$), thus differentiating their overall payoffs and content, highlights that the fees along with differences in content quality on transparency allows the principal to refuse contracting with either agent.

The larger insight from this analysis points to the recent debate on consumer transparency regarding how much their data is worth. AI-generated content is raising the value of human-

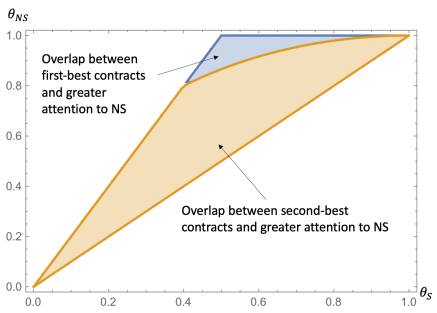
generated content (Economist, 2024). Readers have to decide between machine-produced content that may look human-produced. Given the transparency of AI use, putting a price on the attention may ease that decision. Specifically, paying to disclose AI use substitutes away from attention allocated in Model 2. In essence, the principal puts a price on part of their attention and pays the agent. While the debate has centered around full disclosure on how user data is used, our analysis indicates that the price for transparency presents the possibility that the principal can be much better off without such a price. Thus, our research suggests supporting evidence for the Data Value and Transparency Act 2019, proposed by Senators Mark Warner and Josh Hawley (Warzel, 2019).

6 Policy Implications

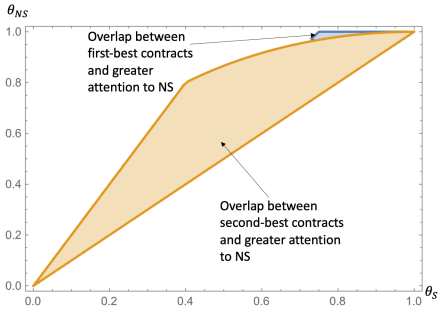
Our policy implications consider two major dimensions: trust vs. accountability. Keeping with Section 3, we consider accountability as the absence of trust. Further, we base our insights on Model 1 and 2 analysis considered in Sections 4.1 and 4.2. Our overall analysis strategy is to connect the two models and assume that if one agent type dominates in hiring in Model 1, then in Model 2, that agent is the dominant type.

Suppose the news organization screens and hires more journalists who typically do a significant proportion of non-AI work (refer to Figures 5a and 5b). Then, Model 1 analysis shows that when these journalists differ significantly from shirkers, first-best contracts with accountability (i.e., oversight), where the principal and agent maximize their payoffs as best as possible, are possible. Still, note that the news organization essentially contracts with these journalists *without trust*. Through Model 2, this news organization is now a non-shirker for AI use but may not receive more public attention. Public attention is only guaranteed when its non-AI work proportions are comparable, regardless of responsibility (i.e., transparency), to news companies that shirk when using AI. Generally, suppose a news organization has a majority of non-shirker journalists with first-best contracts and accountability. In that case, the public may pay more attention to a shirker news organization where transparency does not matter.

Conversely, suppose the news organization hires more journalists with a lesser proportion of non-AI work (refer Figures 6a and 6b). Then, Model 1 analysis shows that regardless of accountability, second-best contracts are possible when the hired journalists are similar in non-



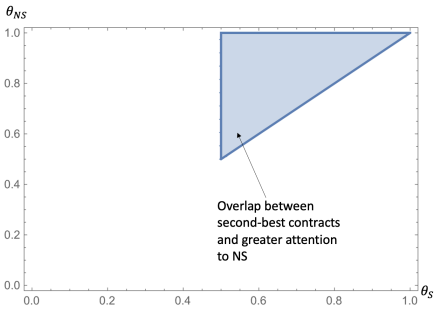
(a) Contract with oversight overlap with greater attention to NS contracts, $\frac{r_S}{r_{NS}} = 0.5$



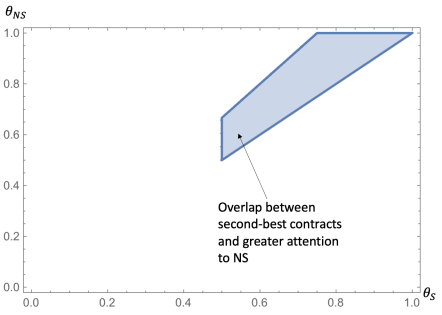
(b) Contract with oversight overlap with greater attention to NS contracts, $\frac{r_S}{r_{NS}} = 0.75$

Figure 5: Overlap between $\theta_i, i \in \{S, NS\}$ ranges comparing across Models 1 and 2.

AI work proportions to non-shirkers. Note that in this case, *trust* does not matter – the news organization relies on incentives to nudge AI use. Consequently, across Models 1 and 2, second-best contracts might be more prevalent than first-best, and Model 2 suggests the principal might pay more attention to the non-shirker news organization. Yet, there are scenarios when, even in this case, if the shirker news organization's non-AI work proportions are meager compared to the non-shirker's, Model 2 suggests the shirker organization receives more attention than the other. Thus, our end-to-end analysis of Models 1 and 2 indicates that a dilemma exists at the heart of journalism's production and consumption in the burgeoning AI age.



(a) Contract without oversight overlap with greater attention to NS, $\frac{r_S}{r_{NS}} = 0.5$



(b) Contract without oversight overlap with greater attention to NS, $\frac{r_S}{r_{NS}} = 0.75$

Figure 6: Overlap between $\theta_i, i \in \{S, NS\}$ ranges comparing across Models 1 and 2.

Trust vs. Accountability Dilemma: We hypothesize that the growing use of AI in journalism may lead readers to consume increasingly AI-produced news without accountability, even when the news organization does not trust its journalistic processes and invests in accountability mechanisms and hiring AI-skilled journalists. Further, we propose that when news organizations compete with others in the market by incentivizing more AI content production (i.e., lower non-

AI work proportions), then they may trust journalists with producing more content by slacking on accountability, not offer them contracts that are optimal given their effort costs and might further risk driving readers to consume content from competing organizations that increasingly make low-quality AI-generated content without accountability.

Considering this dilemma, we offer the following potential ways in which journalistic guidelines of AI use can evolve. We draw on insights from Section 5 and the main analysis in Section 4. First, we propose that oversight costs be controlled such that work on editorial oversight can be distributed over several rather than a few personnel. Second, we suggest that experienced journalists in news organizations be trained for responsible AI use while younger or newer journalists are trained and screened for AI use whenever possible. Finally, we propose that news organizations wishing to be transparent about AI use should monetize the reader's consumption to ensure that AI use disclosure retains rather than repels readers.

7 Conclusion & Future Directions

This paper proposes an agency-theoretic framework to incentivize desired AI use in news organizations and journalists. Our analysis was divided into two models – Model 1 and 2 (Subsections 3.1 and 3.2) – where we considered news organizations and journalists (Model 1) and the public and news organizations (Model 2). Our analysis considers designing an incentive system that, in each case, the principal can use to incentivize the agent to exert the desired type of effort – assistive AI use. Our fundamental analysis establishes crucial results that indicate the trade-off between trust vs. accountability measures when incorporating AI use in journalism. We extend the base model in several ways, relaxing some assumptions regarding oversight costs and occasional shirking. Still, we show that the fundamental insights from our base model hold.

An essential next step in our analysis is adding another actor to the framework – news sources (agent) relationship with journalists (principal). Napoli (1997) shows that this relationship is essential to trustworthy journalism. In contrast, Fengler and Russ-Mohl (2008) indicates how such connections are usually relational *contracts* that might not have financial incentives but depend on the principal's goodwill with the agent. With burgeoning AI use, this relationship allows analyzing how information acquisition costs may still be crucial, relevant, and, therefore, determinant of hybridizing AI in journalistic work. Essentially, till AI can form human

connections independent of supervision and ensure trustworthy news sourcing, the journalist remains an essential and irreplaceable cog in the news organization's wheel.

Another step forward would be combining Models 1 and 2 and analyzing the overall equilibrium region of non-AI work for possible contracts, given agent types. The current analysis examines the non-AI work proportions separately and combines it in Section 6 with qualitative implications. However, there is an opportunity to develop a model with a *dual-nature* principal and agent relationship. Miller (2005) shows the relevance of such relationships to the political sphere (i.e., bureaucrats with politicians and the public). In contrast, Lee and Taylor (2014) shows the relevance of such relationships to the general corporate setting. Future work could address the significance of this in journalism.

Finally, we can test our insights in an experimental setting. Using lab experiments that consider oversight or transparency of writing generated using AI assistively or generatively, we can tease out how incentives impact desired AI use. Conlon and Parks (1990) studied how monitoring for behavior in traditional principal-agent settings can impact contingent-performance pay. Subsequently, studies such as Jacquemet et al. (2008) studied the effects of optimism biases when monitoring illegal activities. Still, journalism's uniquely human nature and its relevance to the health of democracy show that experimental testing of derived principal-agent insights is imperative in light of burgeoning AI use.

References

Bartling, B., Fehr, E., and Schmidt, K. M. 2012. "Screening, competition, and job design: Economic origins of good jobs," *American Economic Review* (102:2), pp. 834–864.

Brewster, J., Fishman, Z., and Xu, E. 2023. "Funding the Next Generation of Content Farms - Misinformation Monitor: June 2023 - NewsGuard," .
URL `https://www.newsguardtech.com/misinformation-monitor/june-2023/`

Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., and Chuan, C.-H. 2019. "Artificial Intelligence and Journalism," *Journalism & Mass Communication Quarterly* (96:3), pp. 673–695.
URL `https://doi.org/10.1177/1077699019859901`

Buhmann, A., and Fieseler, C. 2021. "Towards a deliberative framework for responsible innovation in artificial intelligence," *Technology in Society* (64), p. 101,475.
URL `https://www.sciencedirect.com/science/article/pii/S0160791X20312781`

Carlin, P. S. 1989. "Why the incidence of shirking varies across employers," *Journal of Behavioral Economics* (18:2), pp. 61–73.
URL `https://www.sciencedirect.com/science/article/pii/0090572089900016`

Carmichael, H. L. 1989. "Self-enforcing contracts, shirking, and life cycle incentives," *Journal of Economic Perspectives* (3:4), pp. 65–83.

Chin-Rothmann, C. 2023. "Navigating the risks of artificial intelligence on the digital news landscape," Tech. rep., Center for Strategic & International Studies.
URL https://www.csis.org/analysis/navigating-risks-artificial-intelligence-digital-news-landscape

CNTI 2023. "Artificial Intelligence in Journalism - Center for News, Technology and Innovation," .
URL https://innovating.news/article/ai-in-journalism/

Conlon, E. J., and Parks, J. M. 1990. "Effects of monitoring and tradition on compensation arrangements: An experiment with principal-agent dyads," *Academy of Management Journal* (33:3), pp. 603–622.

Cools, H., and Diakopoulos, N. 2023. "Writing guidelines for the role of AI in your newsroom? Here are some, er, guidelines for that," .
URL https://www.niemanlab.org/2023/07/writing-guidelines-for-the-role-of-ai-in-your-newsroom-here-are-some-er-guidelines-for-that/

Deb, R., and Mishra, D. 2014. "Implementation with contingent contracts," *Econometrica* (82:6), pp. 2371–2393.

Diakopoulos, N. 2020. "Accountability, Transparency, and Algorithms," *The Oxford handbook of ethics of AI* (17:4), p. 197.

Economist, T. 2024. "AI-generated content is raising the value of trust," .
URL https://www.economist.com/leaders/2024/01/18/ai-generated-content-is-raising-the-value-of-trust

Eguchi, K. 2004. "Trainers' dilemma of choosing between training and promotion," *Labour Economics* (11:6), pp. 765–783. The Organization of Labour within Firms.
URL https://www.sciencedirect.com/science/article/pii/S0927537104000260

Fengler, S., and Russ-Mohl, S. 2008. "The crumbling hidden wall: Towards an economic theory of journalism," *Kyklos* (61:4), pp. 520–542.

Fengler, S., and Ruß-Mohl, S. 2008. "Journalists and the information-attention markets: Towards an economic theory of journalism," *Journalism* (9:6), pp. 667–690.

Fu, A. 2023. "As AI enters newsrooms, unions push for worker protections," .
URL https://www.poynter.org/business-work/2023/artificial-intelligence-writers-guild-unions-journalism-jobs/

Gan, T., Hu, J., and Weng, X. 2023. "Optimal contingent delegation," *Journal of Economic Theory* (208), p. 105,597.

Grieco, E. 2020. "US newspapers have shed half of their newsroom employees since 2008," *Pew Research Center Briefs* .

Guaglione, S. 2023. "'Nothing Is Ready for Prime Time': Journalists Push Back Against Publications' Race to Have Newsrooms Use Generative AI Tools," *DigiDay* .
URL https://digiday.com/media/nothing-is-ready-for-primetime-journalists-push-back-against-publications-race-to-have-newsrooms-use-generative-ai-tools/

Guess, R. 2024. "AI generates debate over newsroom ethics," Tech. rep., Voice of America.
URL https://www.voanews.com/a/ai-generates-debate-over-newsroom-ethics-/7450743.html

Hofeditz, L., and Mirbabaie, M. 2021. "Do You Trust an AI-journalist? A Credibility Analysis of News Content with AI-Authorship." in *Proceedings of the Twenty-Ninth European Conference on Information Systems (ECIS 2021)*, .

Holmström, B. 2017. "Pay for performance and beyond," *American Economic Review* (107:7), pp. 1753–1777.

Huang, F., and Cappelli, P. 2010. "Applicant screening and performance-related outcomes," *American Economic Review* (100:2), pp. 214–218.

Hurst, L. 2023. "Robot reporters? Here's how news organisations are using AI in journalism," .
URL https://www.euronews.com/next/2023/08/24/robot-reporters-heres-how-news-organisations-are-using-ai-in-journalism

ICIJ 2023. "ICIJ Joins Global Initiative to Set AI Guidelines for Journalists - ICIJ," .
URL https://www.icij.org/inside-icij/2023/08/icij-joins-global-initiative-to-set-ai-guidelines-for-journalists/

Jacquemet, N., Rullière, J.-L., and Vialle, I. 2008. "Monitoring optimistic agents," *Journal of Economic Psychology* (29:5), pp. 698–714.

Jamil, S. 2021. "Artificial intelligence and journalistic practice: The crossroads of obstacles and opportunities for the Pakistani journalists," *Journalism Practice* (15:10), pp. 1400–1422.

Laffont, J.-J. 2003. *The principal agent model*, Edward Elgar Publishing.

Laffont, J.-J., and Maskin, E. 1981. *The theory of incentives: An overview*, Université des sciences sociales, Faculté des sciences économiques.

Lee, J., and Taylor, M. S. 2014. "Dual roles in psychological contracts: When managers take both agent and principal roles," *Human Resource Management Review* (24:1), pp. 95–107.

Lewicki, R. J., and Brinsfield, C. 2015. "Trust research: measuring trust beliefs and behaviours," *Handbook of research methods on trust* pp. 46–64.

Longoni, C., Fradkin, A., Cian, L., and Pennycook, G. 2022. "News from generative artificial intelligence is believed less," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, .

Martimort, D., and Laffont, J.-J. 2009. *The theory of incentives: the principal-agent model*, Princeton University Press.

Merrefield, C. 2023. "Researchers compare AI policies and guidelines at 52 news organizations," .
URL https://journalistsresource.org/home/generative-ai-policies-newsrooms/

Miller, G. J. 2005. "The political evolution of principal-agent models," *Annu Rev Polit Sci* (8), pp. 203–225.

Mills, R. W., and Reiss, D. R. 2017. "The role of trust in the regulation of complex and high-risk industries: The case of the US Federal Aviation administration's voluntary disclosure programs," in *Trust in regulatory regimes*, , Edward Elgar Publishing, pp. 37–59.

Moran, R. E., and Shaikh, S. J. 2022. "Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism," *Digital journalism* (10:10), pp. 1756–1774.

Myllylahti, M. 2020. "Paying attention to attention: A conceptual framework for studying news reader revenue models related to platforms," *Digital Journalism* (8:5), pp. 567–575.

Nahmias, Y., and Perel, M. 2021. "The oversight of content moderation by AI: Impact assessments and their limitations," *Harv J on Legis* (58), p. 145.

Napoli, P. M. 1997. "A principal-agent approach to the study of media organizations: Toward a theory of the media firm," *Political Communication* (14:2), pp. 207–219.

Nel, F. 2023. "Should the media tell you when they use AI to report the news? What consumers should know," .
URL https://theconversation.com/should-the-media-tell-you-when-they-use-ai-to-report-the-news-what-consumers-should-know-213693

Nelson, J. 2023. "Associated Press Limits How Journalists Can Use Generative AI," .
URL https://decrypt.co/152873/associated-press-limits-how-journalists-use-generative-ai

Newsquest 2024. "AI-assisted reporter," .
URL https://www.journalism.co.uk/media-jobs/ai-assisted-reporter/s75/a1107718/

North, D. C. 1986. "The new institutional economics," *Journal of Institutional and Theoretical Economics (jite)/Zeitschrift für die gesamte Staatswissenschaft* (142:1), pp. 230–237.

Oh, S. 2018. "Here's how newsrooms pay journalists who code, design, and analyze data," *Medium* .
URL https://medium.com/jsk-class-of-2018/news-nerd-salaries-2017-2c83466b994e

Opdahl, A. L., Tessem, B., Dang-Nguyen, D.-T., Motta, E., Setty, V., Throndsen, E., Tverberg, A., and Trattner, C. 2023. "Trustworthy journalism through AI," *Data & Knowledge Engineering* (146), p. 102,182.

Ostrom, E. 2010. "Beyond markets and states: polycentric governance of complex economic systems," *American economic review* (100:3), pp. 641–672.

Pavlik, J. V. 2023. "Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education," *Journalism & Mass Communication Educator* (78:1), pp. 84–93.

Reuters, T. 2023. "AI principles," .
URL https://www.thomsonreuters.com/en/artificial-intelligence/ai-principles.html

Schori, M. 2023. "Aftonbladets AI-policy – så förhåller vi oss till den nya tekniken," *Aftonbladet* .
URL https://www.aftonbladet.se/omaftonbladet/a/76ydy9/aftonbladets-ai-policy-sa-forhaller-vi-oss-till-den-nya-tekniken

Scire, S. 2023. "Most readers want publishers to label AI-generated articles — but trust outlets less when they do," .
URL https://www.niemanlab.org/2023/12/most-readers-want-publishers-to-label-ai-generated-articles-but-trust-outlets-less-when-they-do/

Serdouk, A., and Bessam, A. C. 2023. "Bots in Newsrooms: What Future for Human Journalists?" *Media Watch* (14:1), pp. 100–115.

Shapiro, T. 2023. "The risks of AI-Generated content," .
URL https://stratabeat.com/ai-generated-content-risks/

Subramaniam, V. 2023. "How must journalists and journalism view generative AI?" .
URL https://www.scu.edu/ethics-spotlight/generative-ai-ethics/how-must-journalists-and-journalism-view-generative-ai-/

Tobitt, C. 2023. "The ethics of using generative AI to create journalism: What we know so far," *Press Gazette, UK* .
URL https://pressgazette.co.uk/publishers/digital-journalism/ai-news-journalism-ethics/

Túñez-López, J.-M., Fieiras-Ceide, C., and Vaz-Álvarez, M. 2021. "Impact of Artificial Intelligence on Journalism: transformations in the company, products, contents and professional profile," *Communication & society* (34:1), pp. 177–193.

Vanacker, B., and Belmas, G. 2009. "Trust and the economics of news," *Journal of Mass Media Ethics* (24:2-3), pp. 110–126.

Vargas, A. 2023. "Traffic Is Down, Generative AI Will Make It Worse – And Publishers Must Look Inward For Answers," .
URL https://www.adexchanger.com/the-sell-sider/traffic-is-down-gen-ai-will-make-it-worse-and-publishers-must-look-inward-for-answers/

VG 2023. "Editorial guidelines for AI in VG," *VG* .
URL https://www.vg.no/informasjon/redaksjonelle-avgjorelser/188

Viner, K., and Bateson, A. 2023. "The Guardian's approach to generative AI," *The Guardian* .
URL https://www.theguardian.com/help/insideguardian/2023/jun/16/the-guardians-approach-to-generative-ai

Warzel, C. 2019. “Opinion | Congress wants data transparency, but it still doesn’t understand the internet,” *New York Times Opinion* .
URL https://www.nytimes.com/2019/06/25/opinion/congress-privacy-bill-internet.html

Yaseen, M. 2023. “Nearly Three Quarters of News Organisations Believe Generative AI Presents New Opportunities for Journalism,” .
URL https://www.lse.ac.uk/News/Latest-news-from-LSE/2023/i-September-2023/Nearly-three-quarters-of-news-organisations-believe-generative-AI-presents-new-opportunities-for-journalism

Online Appendix

A AI use in Journalism – Evidence

AI Use Type	*Example*	*Source*
Generative AI	Journalists professional use of generative AI: 1 in 4 journalists utilize generative AI. Podcasters reported the most usage, 38%. Television journalists used AI the least, at 12%. 16% of adopters employed AI for research, 13% for headline ideas, and 10% for content summarization or editing. Just 2% of newsrooms used LLMs to write whole stories.	Medianet 2024 Media Landscape Report.
	The Associated Press used AI to write quarterly earnings reports on publicly traded companies in 2015. Recently, the organization's August 2023 AI standards state that AI "cannot be used to create publishable content and images for the news service." The Verge revealed in January 2015 that the AP produced earnings stories with AI-content production business Automated Insights. Automated Insights helped the AP create almost 5,000 NCAA Division I men's basketball game previews in 2018.	Merrefield (2023)
	Gannett plans to employ generative artificial intelligence to publish stories as it and other news organizations roll out the popular technology to save money and enhance efficiency. Yet, the largest U.S. newspaper publisher with over 200 daily editions stated it would involve humans to prevent the technology from being implemented automatically without control. In a recent Reuters interview, Gannett senior vice president and head of product Renn Turiano said generative AI could help journalists save time.	Newsquest (2024)
	Futurism staff writer Frank Landymore reported that CNET published AI-generated content. Axios' Tyler Buchanan revealed over the summer that USA Today stopped using AI to write high school sports stories after numerous Columbus Dispatch articles went viral for employing odd language like "a close encounter of the athletic kind."	Merrefield (2023)
	A winter report claimed that CNET employed AI to write explanatory financial service news items ascribed to "CNET Money Staff." Only clicking on the author attribution revealed that technology was used in the text. CNET informed readers once the experiment was uncovered and publicized by other sites. After posting 77 machine-generated stories, then-editor Connie Guglielmo noted that several needed revisions. The site later clarified when AI is employed in story creation.	Merrefield (2023)

B Proof of Lemma 1

Proof. In Model 1's full information benchmark the agent's type is perfectly known to the principal. So, the principal essentially sets wage rates for each agent type. That is, for agent type $i \in \{S, NS\}$, in Stage 4, for a given wage rate $\tilde{p}_i, i \in \{S, NS\}$, the agent sets the effort such that its payoff is maximized.

$$\begin{aligned} \frac{\partial \pi_i}{\partial p_i} &\Rightarrow \tilde{p}_i - 2c_i\theta_i^2 e_i = 0, \\ &\Rightarrow e_i = \frac{\tilde{p}_i}{2c_i\theta_i^2}. \end{aligned}$$

This effort, when accounted for by the principal, results in a problem of maximizing returns given agent type. Since type is perfectly known, IC constraints do not apply. Further, given the agent has concave payoffs, a positive wage rate ensures IR. So, for agent θ_S, with oversight the principal maximizes $q_S + (1-\theta_S)e_S - p_S e_S$ while without oversight, the principal maximizes $q_S e_S - p_S e_S$. Similarly, for agent θ_{NS}, with and without oversight, the principal maximizes $q_{NS} - p_{NS} e_{NS}$. Substituting the efforts derived above into the returns, and solving for $\tilde{p}_i$ yields

$$\tilde{p}^*_{NS} = \frac{1}{2},$$

$$\tilde{p}_S^* = \frac{\theta_S}{2}, \text{ with oversight,}$$
$$\tilde{p}_S^* = \theta_S - \frac{1}{2}, \text{ without oversight,}$$

where note that without oversight, $\tilde{p}_S^* > 0$ iff $\theta_S > \frac{1}{2}$ but $\theta_S < \theta_{NS}$. □

C Proof of Proposition 1

Proof. We begin by recognizing that given the extensive game, agents set efforts after knowing their types privately. So, the effort levels derived in the proof of Lemma 1 (Appendix B), still hold. However, now the principal faces adverse selection issues and has to maximize expected payoffs. Consequently, we setup the following Lagrangian, with IC constraints. We first analyze the case where principal uses oversight. Again, we note that given payoff structure of agents, IR is satisfied as long as the principal pays a positive wage to the agent given type.

$$\begin{aligned}\mathcal{L} = \mu\left(\frac{p_S}{2c_S\theta_S} - \frac{p_S^2}{2c_S\theta_S^2}\right) + (1-\mu)\left(\frac{p_{NS}}{2c_{NS}\theta_{NS}} - \frac{p_{NS}^2}{2c_{NS}\theta_{NS}^2} + \frac{(1-\theta_{NS})p_{NS}}{2c_{NS}\theta_{NS}^2}\right) \\ + \lambda_1\left(\frac{\theta_{NS}^2 p_S^2}{4\theta_S^4 c_S} - \frac{p_{NS}p_S}{2\theta_S^2 c_S} + \frac{p_S^2}{4\theta_S^2 c_S}\right) + \lambda_2\left(\frac{\theta_S^2 p_{NS}^2}{4\theta_{NS}^4 c_{NS}} + \frac{p_{NS}^2}{4\theta_{NS}^2 c_{NS}} - \frac{p_{NS}p_S}{2\theta_{NS}^2 c_{NS}}\right)\end{aligned} \tag{5}$$

Case 1: Assuming interior solution exists, and substituting $\lambda_1 = \lambda_2 = 0$, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial p_i} = 0$ to obtain $p_S^* = \frac{\theta_S}{2}, p_{NS}^* = \frac{1}{2}$. Further, checking if this solution implies both IC constraints hold with inequality, we get the following expressions for the IC constraints:

$$\frac{\theta_{NS}^2 - (2-\theta_S)\theta_S}{16\theta_S^2 c_S},$$
$$\frac{(1-2\theta_S)\theta_{NS}^2 + \theta_S^2}{16\theta_{NS}^4 c_{NS}}.$$

For both these equations to simultaneously hold with inequality, we must have that $\theta_{NS} > \sqrt{\theta_S(2-\theta_S)}$ along with $\theta_{NS} > \theta_S$ but $\theta_{NS} < 1$. Consequently, we conclude that the interior is possible in a non-empty region as shown in Figure 4a.

Case 2: Now assuming the corner solutions exist such that $\lambda_1 > 0, \lambda_2 = 0$. Substituting these values in Equation 5, and solving for the FOCs $\frac{\partial \mathcal{L}}{\partial p_i} = 0$ along with the equation $\frac{\theta_{NS}^2 p_S^2}{4\theta_S^4 c_S} - \frac{p_{NS}p_S}{2\theta_S^2 c_S} + \frac{p_S^2}{4\theta_S^2 c_S} = 0$, for p_i, λ_1 we get

$$p_S^* = \frac{\theta_{NS}^2\theta_S^2(c_S(1-\mu) + 2\theta_S c_{NS}\mu) + \theta_S^4 c_S(1-\mu)}{\mu\left(4\theta_{NS}^2\theta_S^2 c_{NS} - c_S\left(\theta_{NS}^2 + \theta_S^2\right)^2\right) + c_S\left(\theta_{NS}^2 + \theta_S^2\right)^2},$$

$$p^*_{NS} = \frac{\left(\theta^2_{NS}+\theta^2_S\right)\left(\theta^2_{NS}(c_S(1-\mu)+2\theta_S c_{NS}\mu)+\theta^2_S c_S(1-\mu)\right)}{2\mu\left(4\theta^2_{NS}\theta^2_S c_{NS} - c_S\left(\theta^2_{NS}+\theta^2_S\right)^2\right)+2c_S\left(\theta^2_{NS}+\theta^2_S\right)^2},$$
$$\lambda_1 = \frac{2\theta_S c_S(1-\mu)\mu\left((2-\theta_S)\theta_S-\theta^2_{NS}\right)}{\theta^2_{NS}(c_S(1-\mu)+2\theta_S c_{NS}\mu)+\theta^2_S c_S(1-\mu)}.$$

Under standard assumptions of Model 1, it follows that $p_S, p_{NS} > 0$ as derived. However, for this corner solution to hold, we need that $\lambda_1 > 0$ while the second IC constraint (since $\lambda_2 = 0$) should also hold with inequality. That is possible iff $\theta_S < \theta_{NS} < \sqrt{\theta_S(2-\theta_S)}$. This region of θ_S, θ_{NS} is such that the interior solution does not hold. Hence, as shown in Figure 4a, this region is non-empty too, and so, this corner solution is valid.

Case 3: Now assuming the other corner solution exists such that $\lambda_1 = 0, \lambda_2 > 0$. Following the same procedure and solving the FOCs, we get

$$p_S = \frac{\theta_S\left(\theta^2_{NS}+\theta^2_S\right)\left(\theta^2_{NS}c_{NS}\mu+\theta^2_S c_{NS}\mu+2\theta_S c_S(1-\mu)\right)}{4\theta^2_{NS}\theta^2_S(c_{NS}\mu+2c_S(1-\mu))+2\theta^4_{NS}c_{NS}\mu+2\theta^4_S c_{NS}\mu},$$
$$p_{NS} = \frac{\theta^2_{NS}\theta_S\left(\theta^2_{NS}c_{NS}\mu+\theta^2_S c_{NS}\mu+2\theta_S c_S(1-\mu)\right)}{2\theta^2_{NS}\theta^2_S(c_{NS}\mu+2c_S(1-\mu))+\theta^4_{NS}c_{NS}\mu+\theta^4_S c_{NS}\mu},$$
$$\lambda_2 = -\frac{2c_{NS}(1-\mu)\mu\left(\theta^2_{NS}(1-2\theta_S)+\theta^2_S\right)}{\theta_S\left(\theta^2_{NS}c_{NS}\mu+\theta^2_S c_{NS}\mu+2\theta_S c_S(1-\mu)\right)}.$$

Under Model 1 assumptions, it follows that $\lambda_2 < 0$ implying that this corner is not valid. We therefore have shown that with oversight, the principal can offer first-best and second-best contracts, under different non-empty regions of non-AI work proportions for the agent.

We now consider the case where the principal does not use oversight. Noting that without oversight, the returns from type S agent changes, we setup the following Lagrangian with the same IC constraints as in Equation 5.

$$\begin{aligned}\mathcal{L} = \mu\left(\frac{p_S}{2c_S\theta_S}-\frac{p^2_S}{2c_S\theta^2_S}-\frac{(1-\theta_S)p_S}{2c_S\theta^2_S}\right)+(1-\mu)\left(\frac{p_{NS}}{2c_{NS}\theta_{NS}}-\frac{p^2_{NS}}{2c_{NS}\theta^2_{NS}}+\frac{(1-\theta_{NS})p_{NS}}{2c_{NS}\theta^2_{NS}}\right)\\ +\lambda_1\left(\frac{\theta^2_{NS}p^2_S}{4\theta^4_S c_S}-\frac{p_{NS}p_S}{2\theta^2_S c_S}+\frac{p^2_S}{4\theta^2_S c_S}\right)+\lambda_2\left(\frac{\theta^2_S p^2_{NS}}{4\theta^4_{NS}c_{NS}}+\frac{p^2_{NS}}{4\theta^2_{NS}c_{NS}}-\frac{p_{NS}p_S}{2\theta^2_{NS}c_{NS}}\right)\end{aligned} \tag{6}$$

Case 1: Assuming interior solution exists, and substituting $\lambda_1 = \lambda_2 = 0$, we solve the FOCs $\frac{\partial\mathcal{L}}{\partial p_i} = 0$ to obtain $p^*_S = \theta_S - \frac{1}{2}, p^*_{NS} = \frac{1}{2}$. Yet, when checking if this solution implies the IC constraints hold with inequality, we get

$$\frac{(1-2\theta_S)\left(\theta^2_{NS}(1-2\theta_S)+\theta^2_S(3-2\theta_S)\right)}{16\theta^4_S c_S},\ \text{IC1}$$

$$\frac{\theta_{NS}^2(3-4\theta_S)+\theta_S^2}{16\theta_{NS}^4 c_{NS}},\ \text{IC2}.$$

Observe that for wages to be positive, we must have $\frac{1}{2} < \theta_S < \theta_{NS}$. However, if this holds, then it follows that IC1 < 0. Consequently, it follows that the interior is not valid.

Case 2: Now assuming the corner solutions exist such that $\lambda_1 > 0, \lambda_2 = 0$. Substituting these values in Equation 5, and solving for the FOCs $\frac{\partial \mathcal{L}}{\partial p_i} = 0$ along with the equation $\frac{\theta_{NS}^2 p_S^2}{4\theta_S^4 c_S} - \frac{p_{NS} p_S}{2\theta_S^2 c_S} + \frac{p_S^2}{4\theta_S^2 c_S} = 0$, for p_i, λ_1 we get

$$p_S^* = \frac{\theta_{NS}^2\theta_S^2(c_S(1-\mu)+2(1-2\theta_S)c_{NS}\mu)+\theta_S^4 c_S(1-\mu)}{\mu\left(4\theta_{NS}^2\theta_S^2 c_{NS} - c_S\left(\theta_{NS}^2+\theta_S^2\right)^2\right)+c_S\left(\theta_{NS}^2+\theta_S^2\right)^2},$$
$$p_{NS}^* = \frac{\left(\theta_{NS}^2+\theta_S^2\right)\left(\theta_{NS}^2(c_S(1-\mu)+2(1-2\theta_S)c_{NS}\mu)+\theta_S^2 c_S(1-\mu)\right)}{2\mu\left(4\theta_{NS}^2\theta_S^2 c_{NS} - c_S\left(\theta_{NS}^2+\theta_S^2\right)^2\right)+2c_S\left(\theta_{NS}^2+\theta_S^2\right)^2},$$
$$\lambda_1 = \frac{2c_S(1-\mu)\mu\left(\theta_{NS}^2(1-2\theta_S)+\theta_S^2(3-2\theta_S)\right)}{\theta_{NS}^2(c_S(1-\mu)-2(1-2\theta_S)c_{NS}\mu)+\theta_S^2 c_S(1-\mu)}.$$

For this corner to be valid, $\lambda_1 > 0$ and the IC2 > 0. It follows that for $\frac{1}{2} < \theta_S < \theta_{NS}$, both $\lambda_1 > 0$ and the IC2 > 0. Hence, this corner is valid.

Case 3: Now assuming the other corner solution exists such that $\lambda_1 = 0, \lambda_2 > 0$. Following the same procedure and solving the FOCs, we get

$$p_S = \frac{\left(\theta_{NS}^2+\theta_S^2\right)\left((2\theta_S-1)c_{NS}\mu\left(\theta_{NS}^2+\theta_S^2\right)+2\theta_S^2 c_S(1-\mu)\right)}{4\theta_{NS}^2\theta_S^2(c_{NS}\mu+2c_S(1-\mu))+2c_{NS}\mu(\theta_{NS}^4+\theta_S^4)},$$
$$p_{NS} = \frac{\theta_{NS}^2\left((2\theta_S-1)c_{NS}\mu\left(\theta_{NS}^2+\theta_S^2\right)+2\theta_S^2 c_S(1-\mu)\right)}{2\theta_{NS}^2\theta_S^2(c_{NS}\mu+2c_S(1-\mu))+\theta_{NS}^4 c_{NS}\mu+\theta_S^4 c_{NS}\mu},$$
$$\lambda_2 = -\frac{2c_{NS}(1-\mu)\mu\left(\theta_{NS}^2(3-4\theta_S)+\theta_S^2\right)}{(2\theta_S-1)c_{NS}\mu\left(\theta_{NS}^2+\theta_S^2\right)+2\theta_S^2 c_S(1-\mu)},$$

where under Model 1 assumptions, $\lambda_2 < 0$. It therefore follows that this corner is invalid.

We have therefore shown that only possible solution for setting wage rates without oversight is when $\lambda_1 > 0, \lambda_2 = 0$, with wage rates similar to the case where principal uses oversight. □

D Proof of Proposition 2

Proof. Without screening, the principal has no choice but to offer all journalists the same wage rate. Still, the wages should be such that it is individually rational for journalists to participate. Since wage rates are the same, note that IC constraints do not apply. Given this, the efforts exerted by each agent

type is

$$e_i = \frac{p}{2\theta_i c_i}.$$

Accounting for this effort for each agent type, the principal sets the wage p, by maximizing expected payoff. Now, if the principal uses oversight, then objective to maximize is

$$\pi = \mu\left(\frac{p}{2c_S} - \frac{p^2}{2c_S\theta_S}\right) + (1-\mu)\left(\frac{p}{2c_{NS}} - \frac{p^2}{2c_{NS}\theta_{NS}} + \frac{(1-\theta_{NS})p}{2c_{NS}\theta_{NS}}\right).$$

Differentiating w.r.t p, we have that the optimal wage is

$$p^* = \frac{\theta_S(\theta_{NS}c_{NS}\mu + c_S(1-\mu))}{2(\theta_{NS}c_{NS}\mu + \theta_S c_S(1-\mu))}.$$

Under Model 1 assumptions, this is a strictly positive wage rate, and so the principal offers the agent, regardless of type, wages that are $> \frac{\theta}{2}$. This is because in the derived expression,

$$\begin{aligned}
&\Rightarrow \frac{\theta_S(\theta_{NS}c_{NS}\mu + c_S(1-\mu))}{2(\theta_{NS}c_{NS}\mu + \theta_S c_S(1-\mu))},\\
&\Rightarrow \frac{\theta_S}{2}\left(\frac{(\theta_{NS}c_{NS}\mu + c_S(1-\mu))}{(\theta_{NS}c_{NS}\mu + \theta_S c_S(1-\mu))}\right),\\
&\Rightarrow \frac{\theta_S}{2}\left(\frac{(\theta_{NS}c_{NS}\mu + c_S(1-\mu))}{(\theta_{NS}c_{NS}\mu + (\theta_S+1-1)c_S(1-\mu))}\right),\\
&\Rightarrow \frac{\theta_S}{2}\left(\frac{(\theta_{NS}c_{NS}\mu + c_S(1-\mu))}{(\theta_{NS}c_{NS}\mu + c_S(1-\mu) - (1-\theta_S)c_S(1-\mu))}\right) > \frac{\theta_S}{2}
\end{aligned}$$

because the expression within brackets is greater than 1.

Now consider the case where the principal does not use oversight. Then the objective to maximize is,

$$\pi = \mu\left(\frac{p}{2c_S} - \frac{p(1-\theta_S)}{2\theta_S c_S} - \frac{p^2}{2c_S\theta_S}\right) + (1-\mu)\left(\frac{p}{2c_{NS}} - \frac{p^2}{2c_{NS}\theta_{NS}} + \frac{(1-\theta_{NS})p}{2c_{NS}\theta_{NS}}\right).$$

Differentiating w.r.t p we have that the optimal wage is

$$p^* = \frac{\theta_S c_S(1-\mu) - \theta_{NS}(1-2\theta_S)c_{NS}\mu}{2(\theta_{NS}c_{NS}\mu + \theta_S c_S(1-\mu))}.$$

For this wage rate to be strictly positive, it follows that the numerator of the above expression should be

positive. That is, solving the expression for θ_S

$$\begin{aligned}
&\Rightarrow \theta_S c_S(1-\mu) - \theta_{NS}(1-2\theta_S)c_{NS}\mu > 0, \\
&\Rightarrow \theta_S > \frac{\theta_{NS}c_{NS}\mu}{2\theta_{NS}c_{NS}\mu + c_S(1-\mu)}
\end{aligned}$$

Observe that while this is a valid lower bound of θ_S, under Model 1 assumptions, it should follow that $\frac{\theta_{NS}c_{NS}\mu}{2\theta_{NS}c_{NS}\mu+c_S(1-\mu)} < \theta_{NS}$. That is,

$$\begin{aligned}
&\Rightarrow \frac{\theta_{NS}c_{NS}\mu}{2\theta_{NS}c_{NS}\mu + c_S(1-\mu)} < \theta_{NS}, \\
&\Rightarrow 2\theta_{NS}c_{NS}\mu > \mu c_{NS} - c_S(1-\mu), \\
&\Rightarrow \theta_{NS} > \frac{1}{2} - \frac{1}{2}\left(\frac{c_S(1-\mu)}{c_{NS}\mu}\right),
\end{aligned}$$

where if $\mu < \frac{c_S}{c_S+c_{NS}}$, then the inequality is always satisfied. Hence, for the derived ranges of θ_S, μ, this wage rate is valid. □

E Proof of Proposition 3

Proof. The basic strategy of the proof is to show that with or without screening, oversight is always the dominant action for the principal. This can be done by comparing the payoffs for each of these conditions.

Case 1: Consider the case with screening. Then with and without oversight, the following are the payoffs.

$$\begin{aligned}
\pi^*_{P,O} &= \frac{\theta^2_{NS}c_{NS}\mu + c_S(1-\mu)}{8\theta^2_{NS}c_{NS}c_S}, \\
\pi^*_{P,WO} &= \frac{\left(\theta^2_{NS}(2(1-2\theta_S)c_{NS}\mu - c_S(1-\mu)) - \theta^2_S c_S(1-\mu)\right)^2}{8\theta^2_{NS}c_{NS}c_S\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - \mu\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - 4\theta^2_{NS}\theta^2_S c_{NS}\right)\right)}.
\end{aligned}$$

Both these profits are associated with existence of two contracts that are possible given ranges of θ_S, θ_{NS}. From Figure 4a, we compare these two for the ranges in the intersection of $\theta_{NS} \in (\sqrt{(2-\theta_S)\theta_S}, 1)$ and $\theta_S \in \left(\frac{1}{2}, \theta_{NS}\right)$. In order to establish whether $\pi^*_{P,O} - \pi^*_{P,WO} > 0$, observe that the expression that needs to be strictly positive is obtained by multiplying the numerator of $\pi^*_{P,O}$ with the denominator of $\pi^*_{P,WO}$ and subtracting the numerator of $\pi^*_{P,WO}$. The expression is as follows:

$$\begin{aligned}
&\left(\theta^2_{NS}c_{NS}\mu + c_S(1-\mu)\right)\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - \mu\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - 4\theta^2_{NS}\theta^2_S c_{NS}\right)\right) \\
&- \left(\theta^2_{NS}(2(1-2\theta_S)c_{NS}\mu - c_S(1-\mu)) - \theta^2_S c_S(1-\mu)\right)^2
\end{aligned} \tag{7}$$

Expanding and simplifying, we get the following expression

$$\Rightarrow \theta_{NS}^2 c_{NS}\mu \left(2\theta_{NS}^2(2(1-\theta_S)(3\theta_S-1)c_{NS}\mu + ((2-(4-\theta_S)\theta_S)c_S(1-\mu))) + \theta_{NS}^4(c_S(1-\mu)) + \theta_S^2(8-(8-\theta_S)\theta_S)c_S(1-\mu)\right) \tag{8}$$

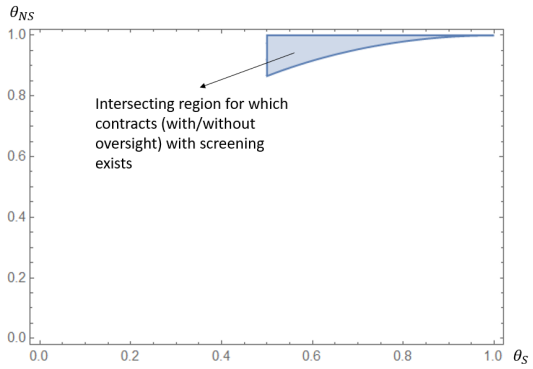


Figure E.1: Contracts with screening but with or without oversight. The highlighted region is the intersection of contracts with screening and without screening (Figs. 4a & 4b).

It follows that for the given ranges of θ_S, θ_{NS} the above expression is strictly positive. Hence, $\pi^*_{P,O} - \pi^*_{P,WO} > 0$ holds for the regions of parameters that are non-empty and for which contracts are possible. Figure E.1 demonstrates the non-empty region for which this holds.

Case 2: Consider the case without screening. Then with and without oversight, following are the payoffs.

$$\pi^*_{P,O} = \frac{\theta_S(1-(1-\theta_{NS})\mu)^2}{8\theta_{NS}(\theta_{NS}\mu+\theta_S(1-\mu))},$$
$$\pi^*_{P,WO} = \frac{(\theta_S(1-(1-2\theta_{NS})\mu)-\theta_{NS}\mu)^2}{8\theta_{NS}\theta_S(\theta_{NS}\mu+\theta_S(1-\mu))}.$$

In order to derive whether $\pi^*_{P,O} - \pi^*_{P,WO} > 0$, observe that the expression to analyze is the numerators of the two payoffs. That is,

$$\Rightarrow \theta_S(1-(1-\theta_{NS})\mu) - (\theta_S(1-(1-2\theta_{NS})\mu) - \theta_{NS}\mu) > 0,$$
$$\Rightarrow \theta_{NS}(1-\theta_S)\mu > 0,$$

which follows under the assumptions for parameters $\theta_S, \theta_{NS}, \mu$ in Model 1. Hence, $\pi^*_{P,O} - \pi^*_{P,WO} > 0$ always holds. □

F Proof of Lemma 2

Proof. To establish the full information benchmark, we ignore IC constraints and consider the principal knows each agent type perfectly. Further, note that the agent's payoff is linear in effort. Hence, the principal, deciding on costly attention allocation, pays as much attention as possible that makes the agent indifferent to participation (IR binding).

Case 1: Consider the scenario with no transparency. In that case, the agent's effort is

$$e_S = \frac{r_S x_S(a_H\gamma_S + a_F(1-\gamma_S))}{\theta_S k_S},$$
$$e_{NS} = \frac{r_{NS} x_{NS}(b_H\gamma_{NS} + b_F(1-\gamma_{NS}))}{\theta_{NS} k_{NS}}.$$

Using these effort levels and substituting into the payoff for the principal, we simply differentiate and obtain the first-best attention levels as follows.

$$x_S^* = \frac{r_S(a_F(1-\gamma_S) + a_H\gamma_S)(v + a_H\gamma_S - a_F(1-\gamma_S))}{2\theta_S k_S},$$
$$x_{NS}^* = \frac{r_{NS}(b_F(1-\gamma_{NS}) + b_H\gamma_{NS})(v + b_H\gamma_{NS} - b_F(1-\gamma_{NS}))}{2\theta_{NS} k_{NS}},$$

which under the general assumptions of Model 2 are strictly positive.

Case 2: Consider the scenario with transparency. In that case, the agent's effort is

$$e_S = \frac{r_S x_S g_S \gamma_S}{\theta_S k_S},$$
$$e_{NS} = \frac{r_{NS} x_{NS} g_{NS} \gamma_{NS}}{\theta_{NS} k_{NS}}.$$

Using these effort levels and substituting into the payoff for the principal, we simply differentiate to obtain the first-best attention levels as follows.

$$x_S^* = \frac{r_S g_S \gamma_S (v + g_S\gamma_S)}{2\theta_S k_S},$$
$$x_{NS}^* = \frac{r_{NS} g_{NS} \gamma_{NS} (v + g_{NS}\gamma_{NS})}{2\theta_{NS} k_{NS}},$$

which under the assumptions of Model 2 is strictly positive. □

G Proof of Proposition 4

Proof. We begin by recognizing that given the extensive game, the agent efforts derived in the proof of Lemma 2 (Appendix F) still hold. Further, under asymmetric information about agent types, the principal maximizes the expected payoff as shown in Section 3.2. Besides, since the principal is risk-averse while the agent is risk-neutral, the IR for the agent always binds. Hence, the Lagrangian is set up with the IC constraints.

We begin by considering the no transparency case, with the following Lagrangian.

$$\begin{aligned}\mathcal{L} = \mu &\left(\frac{r_S x_S(a_H\gamma_S + a_F(1-\gamma_S))(\gamma_S(a_H+v)+(1-\gamma_S)(v-a_F))}{\theta_S k_S} - x_S^2\right)\\ &+ (1-\mu)\left(\frac{r_{NS}x_{NS}(b_H\gamma_{NS}+b_F(1-\gamma_{NS}))(\gamma_{NS}(b_H+v)+(1-\gamma_{NS})(v-b_F))}{\theta_{NS}k_{NS}} - x_{NS}^2\right)\\ &+ \lambda_1\left(\frac{r_S x_S(\theta_{NS}-\theta_S)(a_H\gamma_S+a_F(1-\gamma_S))}{\theta_S} - x_S(r_{NS}-r_S)(a_F(1-\gamma_S)+a_H\gamma_S)\right)\\ &+ \lambda_2\left(\frac{r_{NS}x_{NS}(\theta_S-\theta_{NS})(b_H\gamma_{NS}+b_F(1-\gamma_{NS}))}{\theta_{NS}} + x_{NS}(r_{NS}-r_S)(b_F(1-\gamma_{NS})+b_H\gamma_{NS})\right)\end{aligned} \tag{9}$$

Case 1: Assuming interior solution exists, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_1 = \lambda_2 = 0$ we get

$$\begin{aligned} x_S &= \frac{r_S(a_F(1-\gamma_S)+a_H\gamma_S)(v+a_H\gamma_S-a_F(1-\gamma_S))}{2\theta_S k_S},\\ x_{NS} &= \frac{r_{NS}(b_F(1-\gamma_{NS})+b_H\gamma_{NS})(v+b_H\gamma_{NS}-b_F(1-\gamma_{NS}))}{2\theta_{NS}k_{NS}},\end{aligned}$$

which is the same as the first-best case. Yet, on substituting these values in the IC expressions to check if they hold with inequality, we get the following

$$\begin{aligned} &-\frac{r_S(a_H\gamma_S+a_F(1-\gamma_S))^2(v+a_H\gamma_S-a_F(1-\gamma_S))(\theta_S r_{NS}-\theta_{NS}r_S)}{2\theta_S^2 k_S},\ \text{IC 1}\\ &\frac{r_{NS}(b_H\gamma_{NS}+b_F(1-\gamma_{NS}))^2(v+b_H\gamma_{NS}-b_F(1-\gamma_{NS}))(\theta_S r_{NS}-\theta_{NS}r_S)}{2\theta_{NS}^2 k_{NS}},\ \text{IC 2.}\end{aligned}$$

Observe that all the terms in both the expressions are strictly positive, under Model 2 assumptions, except $(\theta_S r_{NS} - \theta_{NS} r_S)$. If $(\theta_S r_{NS} - \theta_{NS} r_S) > 0$ then IC1 < 0 but IC2 > 0; conversely if $(\theta_S r_{NS} - \theta_{NS} r_S) < 0$ then IC1 > 0 but IC2 < 0. Consequently, the interior cannot hold.

Case 2: Assuming IC1 $= 0$, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_2 = 0$, for

x_i, λ_1. We get the following.

$$
\begin{aligned}
x_S^* &= 0, \\
x_{NS}^* &= \frac{r_{NS}(b_F(1-\gamma_{NS}) + b_H\gamma_{NS})(v + b_H\gamma_{NS} - b_F(1-\gamma_{NS}))}{2\theta_{NS}k_{NS}}, \\
\lambda_1 &= \frac{\mu r_S(v + a_H\gamma_S - a_F(1-\gamma_S))}{k_S(\theta_S r_{NS} - \theta_{NS} r_S)}.
\end{aligned}
$$

Observe that for this corner solution to be valid, both $\lambda_1 > 0$ and $x_{NS}^* \geq 0$. We should also have IC2 > 0 holding, given this solution. Under Model 2 assumptions, $x_{NS}^* > 0$. However, iff $(\theta_S r_{NS} - \theta_{NS} r_S) > 0$ then $\lambda_1 > 0$. Consequently, solving the inequality $(\theta_S r_{NS} - \theta_{NS} r_S) > 0$ we get that $\theta_S > \theta_{NS}\left(\frac{r_S}{r_{NS}}\right)$ but Model 2 assumes $\theta_S < \theta_{NS}$. Hence, when $r_S < r_{NS}$, we get a non-empty region $\theta_S \in \left(\theta_{NS}\left(\frac{r_S}{r_{NS}}\right), \theta_{NS}\right)$, in which $x_{NS}^* > x_S^* = 0$. We further verify this by substituting the obtained solution in IC2 and obtaining the following expression.

$$
\frac{r_{NS}(b_H\gamma_{NS} + b_F(1-\gamma_{NS}))^2(v + b_H\gamma_{NS} - b_F(1-\gamma_{NS}))(\theta_S r_{NS} - \theta_{NS} r_S)}{2\theta_{NS}^2 k_{NS}}
$$

Observe that if $(\theta_S r_{NS} - \theta_{NS} r_S) > 0$, then the above expression is > 0. Therefore, this corner solution is valid.

Case 3: Assuming IC2 $= 0$, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_1 = 0$, for x_i, λ_2. We get the following.

$$
\begin{aligned}
x_S^* &= \frac{r_S(a_F(1-\gamma_S) + a_H\gamma_S)(v + a_H\gamma_S - a_F(1-\gamma_S))}{2\theta_S k_S}, \\
x_{NS}^* &= 0, \\
\lambda_2 &= -\frac{(1-\mu) r_{NS}(v + b_H\gamma_{NS} - b_F(1-\gamma_{NS}))}{k_{NS}(\theta_S r_{NS} - \theta_{NS} r_S)}.
\end{aligned}
$$

Observe that for this corner solution to be valid, both $\lambda_2 > 0$ and $x_S^* \geq 0$. We should also have IC1 > 0 holding, given this solution. Under Model 2 assumptions, $x_S^* > 0$. However, iff $(\theta_S r_{NS} - \theta_{NS} r_S) < 0$ then $\lambda_2 > 0$. Consequently, solving the inequality $(\theta_S r_{NS} - \theta_{NS} r_S) < 0$ we get that $\theta_S < \theta_{NS}\left(\frac{r_S}{r_{NS}}\right)$ but Model 2 assumes $0 < \theta_S < \theta_{NS}$. Hence, we get a non-empty region $\theta_S \in \left(0, \theta_{NS}\left(\frac{r_S}{r_{NS}}\right)\right)$, in which $x_S^* > x_{NS}^* = 0$. We further verify this by substituting the obtained solution in IC1 and obtaining

the following expression.

$$-\frac{r_S(a_H\gamma_S + a_F(1-\gamma_S))^2(v + a_H\gamma_S - a_F(1-\gamma_S))(\theta_S r_{NS} - \theta_{NS} r_S)}{2\theta_S^2 k_S}.$$

Observe that if $(\theta_S r_{NS} - \theta_{NS} r_S) < 0$, then the above expression is > 0. Therefore, this corner solution is valid. We therefore have two corner solutions, holding for different regions of θ_S, as shown in the Proposition 4.

We now consider the transparency scenario. The following Lagrangian is used with IC constraints.

$$\begin{aligned}
&(1-\mu)\left(\frac{g_{NS}\gamma_{NS} r_{NS} x_{NS}(\gamma_{NS}(g_{NS}+v)+(1-\gamma_{NS})v)}{\theta_{NS}k_{NS}} - x_{NS}^2\right)\\
&+\mu\left(\frac{g_S\gamma_S r_S x_S(\gamma_S(g_S+v)+(1-\gamma_S)v)}{\theta_S k_S} - x_S^2\right)\\
&+\lambda_1\left(\frac{g_S\gamma_S r_S x_S(\theta_{NS}-\theta_S)}{\theta_S} + g_S\gamma_S x_S(r_S - r_{NS})\right)\\
&+\lambda_2\left(\frac{g_{NS}\gamma_{NS} r_{NS} x_{NS}(\theta_S-\theta_{NS})}{\theta_{NS}} + g_{NS}\gamma_{NS} x_{NS}(r_{NS} - r_S)\right)
\end{aligned} \tag{10}$$

Case 1: Assuming interior solution exists, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_1 = \lambda_2 = 0$ we get

$$\begin{aligned}
x_S &= \frac{r_S g_S \gamma_S (v + g_S\gamma_S)}{2\theta_S k_S},\\
x_{NS} &= \frac{r_{NS} g_{NS} \gamma_{NS} (v + g_{NS}\gamma_{NS})}{2\theta_{NS} k_{NS}},
\end{aligned}$$

which is the same as the first-best case. Yet, on substituting these values in the IC expressions to check if they hold with inequality, we get the following

$$\begin{aligned}
&-\frac{g_S^2\gamma_S^2 r_S(g_S\gamma_S + v)(\theta_S r_{NS} - \theta_{NS} r_S)}{2\theta_S^2 k_S},\ \text{IC 1}\\
&\frac{g_{NS}^2\gamma_{NS}^2 r_{NS}(g_{NS}\gamma_{NS} + v)(\theta_S r_{NS} - \theta_{NS} r_S)}{2\theta_{NS}^2 k_{NS}},\ \text{IC 2}.
\end{aligned}$$

Observe that all the terms in both the expressions are strictly positive, under Model 2 assumptions, except $(\theta_S r_{NS} - \theta_{NS} r_S)$. If $(\theta_S r_{NS} - \theta_{NS} r_S) > 0$ then IC1 < 0 but IC2 > 0; conversely if $(\theta_S r_{NS} - \theta_{NS} r_S) < 0$ then IC1 > 0 but IC2 < 0. Consequently, the interior cannot hold.

Case 2: Assuming IC1 $= 0$, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_2 = 0$, for

x_i, λ_1. We get the following.

$$
\begin{aligned}
x_S^* &= 0, \\
x_{NS}^* &= \frac{r_{NS} g_{NS} \gamma_{NS} (v + g_{NS} \gamma_{NS})}{2\theta_{NS} k_{NS}}, \\
\lambda_1 &= \frac{\mu r_S (g_S \gamma_S + v)}{k_S (\theta_S r_{NS} - \theta_{NS} r_S)}.
\end{aligned}
$$

Again, similar to the non-transparency case, this corner solution is valid only if $\lambda_1 > 0$. We do not repeat the same analysis here, but conclude that when $r_S < r_{NS}$, we get a non-empty region $\theta_S \in \left(\theta_{NS}\left(\frac{r_S}{r_{NS}}\right), \theta_{NS}\right)$, in which $x_{NS}^* > x_S^* = 0$. We further verify this by substituting the obtained solution in IC2 and obtaining the following expression.

$$
\frac{g_{NS}^2 \gamma_{NS}^2 r_{NS} (g_{NS}\gamma_{NS} + v)(\theta_S r_{NS} - \theta_{NS} r_S)}{2\theta_{NS}^2 k_{NS}}
$$

Observe that if $(\theta_S r_{NS} - \theta_{NS} r_S) > 0$, then the above expression is > 0. Therefore, this corner solution is valid.

Case 3: Assuming IC2 $= 0$, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_1 = 0$, for x_i, λ_2. We get the following.

$$
\begin{aligned}
x_S^* &= \frac{r_S g_S \gamma_S (v + g_S \gamma_S)}{2\theta_S k_S}, \\
x_{NS}^* &= 0, \\
\lambda_2 &= -\frac{(1-\mu) r_{NS} (g_{NS} \gamma_{NS} + v)}{k_{NS} (\theta_S r_{NS} - \theta_{NS} r_S)}.
\end{aligned}
$$

Observe that for this corner solution to be valid, both $\lambda_2 > 0$ and $x_S^* \geq 0$. We should also have IC1 > 0 holding, given this solution. Under Model 2 assumptions, $x_S^* > 0$. However, iff $(\theta_S r_{NS} - \theta_{NS} r_S) < 0$ then $\lambda_2 > 0$. Consequently, solving the inequality $(\theta_S r_{NS} - \theta_{NS} r_S) < 0$ we get that $\theta_S < \theta_{NS}\left(\frac{r_S}{r_{NS}}\right)$ but Model 2 assumes $0 < \theta_S < \theta_{NS}$. Hence, we get a non-empty region $\theta_S \in \left(0, \theta_{NS}\left(\frac{r_S}{r_{NS}}\right)\right)$, in which $x_S^* > x_{NS}^* = 0$. We further verify this by substituting the obtained solution in IC1 and obtaining the following expression.

$$
-\frac{g_S^2 \gamma_S^2 r_S (g_S \gamma_S + v)(\theta_S r_{NS} - \theta_{NS} r_S)}{2\theta_S^2 k_S}.
$$

Observe that if $(\theta_S r_{NS} - \theta_{NS} r_S) < 0$, then the above expression is > 0. Therefore, this corner solution

is valid. We therefore have two corner solutions, holding for different regions of θ_S, as shown in the Proposition 4. □

H Extension Proofs

H.1 Proof of Lemma 3

Proof. This extension considers costly transaction costs for oversight. For brevity, we show the Lagrangian used for this case and the expressions derived when considering the interior and corner. We further demonstrate why these solutions are valid in each case, using the expressions shown.

$$\mathcal{L} = (1-\mu)\left(-\frac{p_{NS}^2}{2\theta_{NS}^2 c_{NS}} + \frac{(1-\theta_{NS})p_{NS}}{2\theta_{NS}^2 c_{NS}} + \frac{p_{NS}}{2\theta_{NS} c_{NS}}\right) + \mu\left(\frac{p_S}{2\theta_S c_S} - \frac{p_S^2}{2\theta_S^2 c_S}\right) - \frac{(1-\theta_S)^2 k p_S^2}{4\theta_S^4 c_S^2} + \lambda_1\left(\frac{\theta_{NS}^2 p_S^2}{4\theta_S^4 c_S} - \frac{p_{NS}p_S}{2\theta_S^2 c_S} + \frac{p_S^2}{4\theta_S^2 c_S}\right) + \lambda_2\left(\frac{\theta_S^2 p_{NS}^2}{4\theta_{NS}^4 c_{NS}} + \frac{p_{NS}^2}{4\theta_{NS}^2 c_{NS}} - \frac{p_{NS}p_S}{2\theta_{NS}^2 c_{NS}}\right) \tag{11}$$

Case 1: Assuming interior solution exists, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_1 = \lambda_2 = 0$ we get

$$p_S = \frac{\theta_S^3 c_S \mu}{2\theta_S^2 c_S \mu + (1-\theta_S)^2 k},$$
$$p_{NS} = \frac{1}{2},$$

which is the same as the first-best case. On substituting these values in the IC expressions to check if they hold with inequality, we get the following

$$\frac{\theta_S \mu\left(\theta_S c_S \mu\left(\theta_{NS}^2 - (2-\theta_S)\theta_S\right) - (1-\theta_S)^2 k\right)}{4\left(2\theta_S^2 c_S \mu + (1-\theta_S)^2 k\right)^2}, \text{ IC 1}$$
$$\frac{\theta_{NS}^2\left(1 - \frac{4\theta_S^3 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k}\right) + \theta_S^2}{16\theta_{NS}^4 c_{NS}}, \text{ IC 2.}$$

In order to establish regions for which the interior holds, observe that in each IC expression, we need to establish whether the following is satisfied. For IC1 we need

$$\theta_S c_S \mu\left(\theta_{NS}^2 - (2-\theta_S)\theta_S\right) - (1-\theta_S)^2 k > 0,$$

and for IC2, we need

$$\theta_{NS}^2\left(1 - \frac{4\theta_S^3 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k}\right) + \theta_S^2 > 0.$$

From the first expression for IC1, we get

$$\begin{aligned}
&\Rightarrow \theta_S c_S \mu \left(\theta_{NS}^2 - (2-\theta_S)\theta_S\right) - (1-\theta_S)^2 k > 0, \\
&\Rightarrow \theta_S c_S \mu \left(\theta_{NS}^2 - (2-\theta_S)\theta_S\right) > (1-\theta_S)^2 k, \\
&\Rightarrow k < \frac{\theta_S c_S \mu \left(\theta_{NS}^2 - (2-\theta_S)\theta_S\right)}{(1-\theta_S)^2}
\end{aligned}$$

From the second expression for IC2, under the assumption $k > 0$ we get

$$\begin{aligned}
&\Rightarrow \theta_{NS}^2 \left(1 - \frac{4\theta_S^3 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k}\right) + \theta_S^2 > 0, \\
&\Rightarrow -\theta_{NS}^2 \left(\frac{4\theta_S^3 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k} - 1\right) + \theta_S^2 > 0, \\
&\Rightarrow \theta_S^2 > \theta_{NS}^2 \left(\frac{4\theta_S^3 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k} - 1\right), \\
&\Rightarrow \theta_S^2 > \theta_{NS}^2 \left(2\theta_S \left(\frac{2\theta_S^2 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k}\right) - 1\right).
\end{aligned}$$

The term within brackets $0 < \frac{2\theta_S^2 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k} < 1$ for $k > 0$. Additionally, the above inequality can be reduced to the following for a tighter bound.

$$\begin{aligned}
&\Rightarrow \theta_S^2 > \left(2\theta_{NS}^2 \theta_S \left(\frac{2\theta_S^2 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k}\right) - \theta_{NS}^2\right), \\
&\Rightarrow \theta_S^2 + \theta_{NS}^2 > 2\theta_S \left(\frac{2\theta_S^2 \theta_{NS}^2 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k}\right), \\
&\Rightarrow \theta_S^2 + \theta_{NS}^2 > 2\theta_S > 2\theta_S \left(\frac{2\theta_S^2 \theta_{NS}^2 c_S \mu}{2\theta_S^2 c_S \mu + (\theta_S - 1)^2 k}\right), \\
&\Rightarrow \theta_S^2 + \theta_{NS}^2 > 2\theta_S, \\
&\Rightarrow \theta_S^2 - 2\theta_S > -\theta_{NS}^2, \\
&\Rightarrow \theta_S^2 - 2\theta_S + 1 > 1 - \theta_{NS}^2, \\
&\Rightarrow (1-\theta_S)^2 > (1 - \theta_{NS}^2), \\
&\Rightarrow \theta_{NS} > \sqrt{1 - (1-\theta_S)^2},
\end{aligned}$$

which is as per the Lemma 3.

Case 2: Assuming IC1 $= 0$, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_2 = 0$, for

p_i, λ_1. We get the following.

$$p_S^* = \frac{\theta_S^2 c_S \left(\theta_{NS}^2(c_S(1-\mu)+2\theta_S c_{NS}\mu)+\theta_S^2 c_S(1-\mu)\right)}{2\theta_{NS}^2\left(\theta_S\left(\theta_S\left(c_{NS}k+c_S^2\right)+\theta_S c_S\mu(2c_{NS}-c_S)-2c_{NS}k\right)+c_{NS}k\right)+\theta_{NS}^4 c_S^2(1-\mu)+\theta_S^4 c_S^2(1-\mu)},$$

$$p_{NS}^* = \frac{c_S\left(\theta_{NS}^2+\theta_S^2\right)\left(\theta_{NS}^2(c_S(1-\mu)+2\theta_S c_{NS}\mu)+\theta_S^2 c_S(1-\mu)\right)}{2c_S^2\left(\theta_{NS}^2+\theta_S^2\right)^2+4\theta_{NS}^2(1-\theta_S)^2 c_{NS}k-2c_S\mu\left(c_S\left(\theta_{NS}^2+\theta_S^2\right)^2-4\theta_{NS}^2\theta_S^2 c_{NS}\right)},$$

$$\lambda_1 = \frac{2(1-\mu)\left((1-\theta_S)^2 k-\theta_S c_S\mu\left(\theta_{NS}^2+(\theta_S-2)\theta_S\right)\right)}{\theta_{NS}^2(c_S(1-\mu)+2\theta_S c_{NS}\mu)+\theta_S^2 c_S(1-\mu)}.$$

Observe that for this corner to be a valid solution, both $\lambda_1 > 0$ and $p_i^* > 0$. Under Model 1 assumptions, the expressions for p_i^* show under inspection that they are positive. However, for $\lambda_1 > 0$ we need that $(1-\theta_S)^2 k - \theta_S c_S\mu\left(\theta_{NS}^2+(\theta_S-2)\theta_S\right) > 0$. That is,

$$\begin{aligned}
&\Rightarrow (1-\theta_S)^2 k - \theta_S c_S\mu\left(\theta_{NS}^2+(\theta_S-2)\theta_S\right) > 0,\\
&\Rightarrow k > \frac{\theta_S c_S\mu\left(\theta_{NS}^2-(2-\theta_S)\theta_S\right)}{(1-\theta_S)^2}.
\end{aligned}$$

Further, substituting this solution into the IC2, we get

$$\frac{c_S^2(\theta_{NS}-\theta_S)^2(\theta_{NS}+\theta_S)^2\left(\theta_{NS}^2+\theta_S^2\right)\left(\theta_{NS}^2(c_S(\mu-1)-2\theta_S c_{NS}\mu)+\theta_S^2 c_S(\mu-1)\right)^2}{16\theta_{NS}^4 c_{NS}\left(-2\theta_{NS}^2(\theta_S(\theta_S c_S(2c_{NS}\mu+c_S(1-\mu))+\theta_S c_{NS}k-2c_{NS}k)+c_{NS}k)+\theta_{NS}^4 c_S^2(\mu-1)+\theta_S^4 c_S^2(\mu-1)\right)^2} > 0$$

Hence, it follows that this corner is a valid solution but for a different marginal transaction cost range.

Case 3: Assuming IC2 $= 0$, we solve the FOCs $\frac{\partial \mathcal{L}}{\partial x_i} = 0, \forall\, i \in \{S, NS\}$ after substituting $\lambda_1 = 0$, for p_i, λ_2. We get the following.

$$p_S = \frac{\theta_S^3 c_S\left(\theta_{NS}^2+\theta_S^2\right)\left(\theta_{NS}^2 c_{NS}\mu+\theta_S^2 c_{NS}\mu-2\theta_S c_S(\mu-1)\right)}{2\theta_{NS}^2\theta_S^2(\theta_S(2\theta_S c_S(c_{NS}\mu-2c_S(\mu-1))+\theta_S c_{NS}k-2c_{NS}k)+c_{NS}k)+\theta_{NS}^4 c_{NS}\left(2\theta_S^2 c_S\mu+(\theta_S-1)^2 k\right)+2\theta_S^6 c_{NS}c_S\mu+(\theta_S-1)^2\theta_S^4 c_{NS}k},$$

$$p_{NS} = \frac{2\theta_{NS}^2\theta_S^3 c_S\left(\theta_{NS}^2 c_{NS}\mu+\theta_S^2 c_{NS}\mu-2\theta_S c_S(\mu-1)\right)}{2\theta_{NS}^2\theta_S^2(\theta_S(2\theta_S c_S(c_{NS}\mu-2c_S(\mu-1))+\theta_S c_{NS}k-2c_{NS}k)+c_{NS}k)+\theta_{NS}^4 c_{NS}\left(2\theta_S^2 c_S\mu+(\theta_S-1)^2 k\right)+2\theta_S^6 c_{NS}c_S\mu+(\theta_S-1)^2\theta_S^4 c_{NS}k},$$

$$\lambda_2 = -\frac{c_{NS}(1-\mu)\left(2\theta_S^2 c_S\mu\left(\theta_{NS}^2(1-2\theta_S)+\theta_S^2\right)+(\theta_S-1)^2 k\left(\theta_{NS}^2+\theta_S^2\right)\right)}{\theta_S^3 c_S\left(\theta_{NS}^2 c_{NS}\mu+\theta_S^2 c_{NS}\mu+2\theta_S c_S(1-\mu)\right)} < 0$$

where $\lambda_2 < 0$ under Model 1 assumptions. Hence, this corner is invalid. We therefore have only one interior and one corner solution. □

H.2 Proof of Proposition 5

Proof. For this proof, we compare the payoffs for payoffs for screening and oversight vs. screening without oversight. As showed in the proof for Lemma 3, there are two possible contracts when screening

and oversight is employed. When screening without oversight, the payoff is as follows (mentioned here for brevity) with $\theta_S \in \left(\frac{1}{2}, \theta_{NS}\right)$:

$$\pi^*_{P,WO} = \frac{\left(\theta^2_{NS}(2(1-2\theta_S)c_{NS}\mu + c_S(\mu-1)) + \theta^2_S c_S(\mu-1)\right)^2}{8\theta^2_{NS}c_{NS}c_S\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - \mu\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - 4\theta^2_{NS}\theta^2_S c_{NS}\right)\right)}.$$

First, comparing the interior equilibrium payoff with $\pi^*_{P,WO}$, we get that under the region $\theta_{NS} > \sqrt{1-(1-\theta_S)^2}$ and when $k < \frac{\theta_S c_S \mu\left(\theta^2_{NS}-(2-\theta_S)\theta_S\right)}{(1-\theta_S)^2}$, $\pi^*_{P,O} > \pi^*_{P,WO}$. Further, when comparing the second corner equilibrium contract payoff, we have that in order to establish when the payoff from oversight dominates the payoff from without oversight, we have to compare the following payoff with $\pi^*_{P,WO}$.

$$\pi^*_{P,O} = \frac{\left(\theta^2_{NS}(c_S(\mu-1)-2\theta_S c_{NS}\mu)+\theta^2_S c_S(\mu-1)\right)^2}{8\theta^2_{NS}c_{NS}\left(2\theta^2_{NS}\left(\theta_S\left(\theta_S\left(c_{NS}k+c^2_S\right)+\theta_S c_S\mu(2c_{NS}-c_S)-2c_{NS}k\right)+c_{NS}k\right)+\theta^4_{NS}c^2_S(1-\mu)+\theta^4_S c^2_S(1-\mu)\right)},$$

under the region $k > \frac{\theta_S c_S \mu\left(\theta^2_{NS}-(2-\theta_S)\theta_S\right)}{(1-\theta_S)^2}$, $\theta_{NS} > \sqrt{1-(1-\theta_S)^2}$ and $\theta_S \in \left(\frac{1}{2}, \theta_{NS}\right)$. In order to establish this dominance analysis, we first observe that under $\theta_S \in \left(\frac{1}{2}, \theta_{NS}\right)$, the following holds for the numerators

$$\left(\theta^2_{NS}(2(1-2\theta_S)c_{NS}\mu + c_S(\mu-1)) + \theta^2_S c_S(\mu-1)\right)^2 > \left(\theta^2_{NS}(c_S(\mu-1)-2\theta_S c_{NS}\mu) + \theta^2_S c_S(\mu-1)\right)^2.$$

Second, for denominator of $\pi^*_{P,O}$ to be greater than denominator of $\pi^*_{P,WO}$, the following must hold,

$$\left(2\theta^2_{NS}\left(\theta_S\left(\theta_S\left(c_{NS}k+c^2_S\right)+\theta_S c_S\mu(2c_{NS}-c_S)-2c_{NS}k\right)+c_{NS}k\right)+\theta^4_{NS}c^2_S(1-\mu)+\theta^4_S c^2_S(1-\mu)\right) > \left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - \mu\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - 4\theta^2_{NS}\theta^2_S c_{NS}\right)\right)$$

Under the region for both equilibria holding, $k > \frac{\theta_S c_S \mu\left(\theta^2_{NS}-(2-\theta_S)\theta_S\right)}{(1-\theta_S)^2}$, $\theta_{NS} > \sqrt{1-(1-\theta_S)^2}$ and $\theta_S \in \left(\frac{1}{2}, \theta_{NS}\right)$, it follows that when $k < \frac{(1-c_S)\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - \mu\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - 4\theta^2_{NS}\theta^2_S c_{NS}\right)\right)}{2\theta^2_{NS}(1-\theta_S)^2 c_{NS}}$ and $c_S < \frac{2\theta_S}{\theta^2_{NS}+\theta^2_S}$, the above inequality is true. Putting this and the previous inequality, it follows that under $k < \frac{(1-c_S)\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - \mu\left(c_S\left(\theta^2_{NS}+\theta^2_S\right)^2 - 4\theta^2_{NS}\theta^2_S c_{NS}\right)\right)}{2\theta^2_{NS}(1-\theta_S)^2 c_{NS}}$ and $c_S < \frac{2\theta_S}{\theta^2_{NS}+\theta^2_S}$, $\pi^*_{P,O} > \pi^*_{P,WO}$. □

H.3 Proof of Proposition 6

Proof. For this proof we first show the objective functions of the agents, with revised marginal costs, to demonstrate the difference with Model 1.

$$\pi_i = p_i e_i - c_i e_i^2, i \in \{S, NS\}$$

Observe that the cost of efforts depend not just on non-AI work proportion but on the entire effort levels. Additionally, note that the IC constraints are now simply

$$\text{IC1: } e_S(p_S - p_{NS}),$$
$$\text{IC2: } e_{NS}(p_{NS} - p_S).$$

This implies that when either constraint binds, we should have wage rates equal for either agent types.

Case 1: The interior solution is simply the first best which is

$$p_S = \frac{\theta}{2},$$
$$p_{NS} = \frac{1}{2}.$$

Now, substituting these into the IC constraints and checking if they hold with inequality, it is clear that

$$\text{IC1: } -\frac{(1-\theta_S)\theta_S}{8c_S} < 0,$$
$$\text{IC2: } \frac{(1-\theta_S)}{8c_{NS}} > 0.$$

Consequently, the interior solution is invalid.

Case 2: Checking the corners, it is easy to see that when setting $p_S = p_{NS}$, one of the IC constraints that is supposed to hold with inequality will fail to hold as such. Consequently corner solutions are not possible too. Hence, contracts using screening are impossible.

Now, considering contracts without screening, both agents are paid the same. Consequently, we solve the problem assuming same wage rate for both types and obtain

$$p_S^* = p_{NS}^* = \frac{c_S - \mu((1-2\theta_S)c_{NS} + c_S)}{2(c_{NS}\mu + c_S(1-\mu))}, \text{ without oversight,}$$
$$p_S^* = p_{NS}^* = \frac{\theta_S c_{NS}\mu + c_S(1-\mu)}{2c_{NS}\mu + 2c_S(1-\mu)}, \text{ with oversight.}$$

Under Model 1 assumptions, both these wage rates are positive. Further, comparing payoffs for principal for these wage rates, we have to compare the following expressions

$$\pi_{P,O}^* = \frac{(\theta_S c_{NS}\mu + c_S(1-\mu))^2}{8c_{NS}c_S(c_{NS}\mu + c_S(1-\mu))},$$

$$\pi^*_{P,WO} = \frac{(c_S - \mu((1-2\theta_S)c_{NS} + c_S))^2}{8c_{NS}c_S(c_{NS}\mu + c_S(1-\mu))}.$$

In order to derive whether $\pi^*_{P,O} - \pi^*_{P,WO} > 0$, observe that the expression to analyze is the numerators of the two payoffs. That is,

$$\begin{aligned}
&\Rightarrow (\theta_S c_{NS}\mu + c_S(1-\mu)) - (c_S - \mu((1-2\theta_S)c_{NS} + c_S)) > 0,\\
&\Rightarrow c_{NS}(1-\theta_S)\mu > 0,
\end{aligned}$$

which follows under the assumptions for parameters $\theta_S, \theta_{NS}, \mu$ in Model 1. Hence, $\pi^*_{P,O} - \pi^*_{P,WO} > 0$ always holds. □

H.4 Proof of Proposition 7

Proof. We begin by noting the expression for the principal's payoff maximization when screening, in both the cases: with and without oversight.

$$\begin{aligned}
\pi_{P,O} &= \mu(\lambda(R_S + (1-\theta_S)e_S - e_S p_S) + (1-\lambda)(R_S - e_S p_S)) + (1-\mu)(R_{NS} - e_{NS}p_{NS}),\\
\pi_{P,WO} &= \mu(\lambda(R_S - e_S p_S) + (1-\lambda)(R_S - e_S p_S)) + (1-\mu)(R_{NS} - e_{NS}p_{NS})
\end{aligned}$$

We now solve the case with oversight starting with the agent. Note that since now the agent's marginal costs are the same, the agent essentially sets for efforts such that

$$\begin{aligned}
e_S &= \frac{p_S}{2c_{NS}(\theta_{NS}(1-\lambda) + \theta_S\lambda)^2},\\
e_{NS} &= \frac{p_{NS}}{2c_{NS}\theta^2_{NS}}
\end{aligned}$$

The first stage is the following Lagrangian, after substituting for the two effort levels derived in $\pi_{P,O}$

$$\begin{aligned}
\mathcal{L} = (1-\mu)&\left(-\frac{p^2_{NS}}{2\theta^2_{NS}c_{NS}} + \frac{(1-\theta_{NS})p_{NS}}{2\theta^2_{NS}c_{NS}} + \frac{p_{NS}}{2\theta_{NS}c_{NS}}\right)\\
+\mu&\left(\lambda\left(\frac{\theta_S p_S}{2c_{NS}(\theta_{NS}(-\lambda) + \theta_{NS} + \theta_S\lambda)^2} - \frac{p^2_S}{2c_{NS}(\theta_{NS}(-\lambda) + \theta_{NS} + \theta_S\lambda)^2}\right)\right.\\
&+(1-\lambda)\left(-\frac{p^2_S}{2c_{NS}(\theta_{NS}(-\lambda) + \theta_{NS} + \theta_S\lambda)^2} - \frac{(1-\theta_S)p_S}{2c_{NS}(\theta_{NS}(-\lambda) + \theta_{NS} + \theta_S\lambda)^2}\right.\\
&\left.\left.+\frac{\theta_S p_S}{2c_{NS}(\theta_{NS}(-\lambda) + \theta_{NS} + \theta_S\lambda)^2}\right)\right) + \lambda_1\left(\frac{p^2_{NS}}{4\theta^2_{NS}c_{NS}} - \frac{p^2_S}{4c_{NS}(\theta_{NS}(-\lambda) + \theta_{NS} + \theta_S\lambda)^2}\right).
\end{aligned} \tag{12}$$

Case 1: Solving for the interior by substituting $\lambda_1 = 0$, we get the following

$$p_S^* = \frac{1}{2}(\theta_S - (1-\theta_S)(1-\lambda)),$$
$$p_{NS}^* = \frac{1}{2}.$$

On substituting this into the constraint and checking for inequality, we get that the interior solution is valid for $\lambda \in (0,1)$ and $\theta_S \in (\frac{1}{2}, \theta_{NS})$. This is similar to the Model 1 scenario but now, since shirking is probabilistic and not associated with the agent type, the principal contracts only when the non-AI work proportions are high across shirking or non-shirking.

Case 2: Assuming the corner holds with equality, we solve for the FOCs and the constraint to obtain the following:

$$p_S = \frac{\theta_{NS}(\mu(\theta_S(2-\lambda)+2\lambda-2)+1-\lambda)+\theta_S\lambda(1-\mu)}{2\theta_{NS}},$$
$$p_{NS} = \frac{\mu(\theta_{NS}(\theta_S(\lambda-2)-2\lambda+2)+\theta_S\lambda)}{2\theta_{NS}(\lambda-1)-2\theta_S\lambda} + \frac{1}{2},$$
$$\lambda_1 = \frac{2(1-\mu)\mu(\theta_{NS}(\theta_S(2-\lambda)+2\lambda-2)+\theta_S\lambda)}{\theta_{NS}(\mu(\theta_S(2-\lambda)+2\lambda-2)+1-\lambda)+\theta_S\lambda(1-\mu)}.$$

On analyzing these values, for the region $\theta_S < \theta_{NS}$ either $\lambda_1 > 0$ and $p_{NS} < 0$ or vice versa. Both these situations are not valid and so, the corner solution does not hold.

Turning to the case without oversight, the analysis is simpler and follows same steps as above. We do not repeat it for sake of brevity but mention the contract wage rates as follows

$$p_S^* = \theta_S - \frac{1}{2},$$
$$p_{NS}^* = \frac{1}{2},$$

which is same as the base case in Model 1. □

H.5 Proof of Proposition 8

Proof. We begin by noting that the effort by the agent, given attention and a fee, is given as follows

$$e_i = \frac{g_i\gamma_i r_i x_i + w_i}{\theta_i k_i}, i \in \{S, NS\}.$$

Further, the first stage Lagrangian is as follows.

$$\begin{aligned}\mathcal{L} = &\lambda_2 \left(\frac{(\theta_S - \theta_{NS})(g_{NS}\gamma_{NS} r_{NS} x_{NS} + w_{NS})}{\theta_{NS}} + g_{NS}\gamma_{NS} x_{NS}(r_{NS} - r_S) \right) \\ &+ \lambda_1 \left(\frac{(\theta_{NS} - \theta_S)(g_S\gamma_S r_S x_S + w_S)}{\theta_S} + g_S\gamma_S x_S(r_S - r_{NS}) \right) \\ &+ (1-\mu) \left(\frac{(\gamma_{NS}(g_{NS}+v) + (1-\gamma_{NS})v)(g_{NS}\gamma_{NS} r_{NS} x_{NS} + w_{NS})}{\theta_{NS} k_{NS}} - w_{NS} - x_{NS}^2 \right) \\ &+ \mu \left(\frac{(\gamma_S(g_S+v) + (1-\gamma_S)v)(g_S\gamma_S r_S x_S + w_S)}{\theta_S k_S} - w_S - x_S^2 \right)\end{aligned} \tag{13}$$

Case 1: Assuming interior solution exists, we solve FOC after substituting $\lambda_1 = \lambda_2 = 0$ to get

$$\begin{aligned} x_S &= \frac{r_S g_S \gamma_S (v + g_S\gamma_S)}{2\theta_S k_S}, \\ x_{NS} &= \frac{r_{NS} g_{NS} \gamma_{NS} (v + g_{NS}\gamma_{NS})}{2\theta_{NS} k_{NS}}, \end{aligned}$$

which is the same as in the base Model 2 analysis. Substituting these into the IC constraints to check if they hold with inequality, we get the following expressions

$$\begin{aligned} \text{IC 1: } &\frac{g_S^2\gamma_S^2 r_S (g_S\gamma_S + v)(\theta_{NS} r_S - \theta_S r_{NS}) + 2\theta_S k_S w_S (\theta_{NS} - \theta_S)}{2\theta_S^2 k_S}, \\ \text{IC 2: } &\frac{g_{NS}^2\gamma_{NS}^2 r_{NS} (g_{NS}\gamma_{NS} + v)(\theta_S r_{NS} - \theta_{NS} r_S) + 2\theta_{NS} k_{NS} w_{NS} (\theta_S - \theta_{NS})}{2\theta_{NS}^2 k_{NS}}. \end{aligned}$$

Observe that when $\theta_S r_{NS} - \theta_{NS} r_S < 0$ then because $\theta_S < \theta_{NS}$, IC2 < 0 and hence, the interior solution is invalid. However, if $\theta_S r_{NS} - \theta_{NS} r_S > 0$ then not only is IC2 > 0, even IC1 > 0 for regions where $k_S > \frac{g_S^2\gamma_S^2 r_S (g_S\gamma_S + 1)(\theta_S r_{NS} - \theta_{NS} r_S)}{2\theta_S w_S (\theta_{NS} - \theta_S)}$, $w_{NS} \leq \frac{g_{NS}^2\gamma_{NS}^2 r_{NS} (g_{NS}\gamma_{NS} + 1)(\theta_S r_{NS} - \theta_{NS} r_S)}{2\theta_{NS} k_{NS} (\theta_{NS} - \theta_S)}$ and $v < \frac{2\theta_S k_S w_S (\theta_S - \theta_{NS})}{g_S^2\gamma_S^2 r_S (\theta_{NS} r_S - \theta_S r_{NS})} - g_S\gamma_S$. We therefore have that under these regions of k_S, w_{NS} and v, and given $\theta_S r_{NS} - \theta_{NS} r_S > 0$, we have an interior valid solution when realizable property rights exist.

Case 2: Now assuming corners exist, we solve for when $\lambda_1 > 0$ get the following

$$\begin{aligned} x_S &= \frac{\theta_{NS} w_S - \theta_S w_S}{\theta_S g_S \gamma_S r_{NS} - \theta_{NS} g_S \gamma_S r_S}, \\ x_{NS} &= \frac{r_{NS} g_{NS} \gamma_{NS} (v + g_{NS}\gamma_{NS})}{2\theta_{NS} k_{NS}}, \\ \lambda_1 &= \frac{\mu \left(\frac{2\theta_S w_S (\theta_S - \theta_{NS})}{g_S^2\gamma_S^2} + \frac{r_S (g_S\gamma_S + v)(\theta_S r_{NS} - \theta_{NS} r_S)}{k_S} \right)}{(\theta_S r_{NS} - \theta_{NS} r_S)^2}. \end{aligned}$$

Observe that again, when $\theta_S r_{NS} - \theta_{NS} r_S < 0$ then because $\theta_S < \theta_{NS}$ $\lambda_1 < 0$ and $x_S < 0$, the corner is invalid. However, when $\theta_S r_{NS} - \theta_{NS} r_S > 0$ then regions can be established for which the corner is

valid. Specifically, substituting values of x_S, x_{NS} into IC 2 expression we get

$$\frac{g_{NS}^2\gamma_{NS}^2 r_{NS}(g_{NS}\gamma_{NS}+v)(\theta_S r_{NS}-\theta_{NS}r_S)+2\theta_{NS}k_{NS}w_{NS}(\theta_S-\theta_{NS})}{2\theta_{NS}^2 k_{NS}}.$$

Subsequently, analyzing this expression along with expressions for x_S, x_{NS} and λ_1, we get that when $k_S < \frac{g_S^2\gamma_S^2 r_S(g_S\gamma_S+1)(\theta_S r_{NS}-\theta_{NS}r_S)}{2\theta_S w_S(\theta_{NS}-\theta_S)}$, $w_{NS} \leq \frac{g_{NS}^2\gamma_{NS}^2 r_{NS}(g_{NS}\gamma_{NS}+1)(\theta_S r_{NS}-\theta_{NS}r_S)}{2\theta_{NS}k_{NS}(\theta_{NS}-\theta_S)}$ and $\frac{\gamma_{NS}}{\gamma_S} < g_S < 1$, this corner solution is valid.

Case 3: Now assuming the other corner is valid where $\lambda_2 > 0$. Then solving the FOCs we get

$$\begin{aligned}
x_S &= \frac{r_S g_S\gamma_S(v+g_S\gamma_S)}{2\theta_S k_S},\\
x_{NS} &= \frac{\theta_{NS}w_{NS}-\theta_S w_{NS}}{\theta_S g_{NS}\gamma_{NS}r_{NS}-\theta_{NS}g_{NS}\gamma_{NS}r_S},\\
\lambda_2 &= -\frac{(1-\mu)\left(g_{NS}^2\gamma_{NS}^2 r_{NS}(g_{NS}\gamma_{NS}+v)(\theta_S r_{NS}-\theta_{NS}r_S)+2\theta_{NS}k_{NS}w_{NS}(\theta_S-\theta_{NS})\right)}{g_{NS}^2\gamma_{NS}^2 k_{NS}(\theta_S r_{NS}-\theta_{NS}r_S)^2}.
\end{aligned}$$

Observe that in this case, when $\theta_S r_{NS}-\theta_{NS}r_S < 0$ then because $\theta_S < \theta_{NS}$. then $\lambda_2 > 0$. However, in this case $x_{NS} < 0$ and hence, even in this case, $\theta_S r_{NS}-\theta_{NS}r_S < 0$ results in an invalid corner solution.

Still, when $\theta_S r_{NS}-\theta_{NS}r_S > 0$ then analyzing all three expressions, along with the following IC 1 expression

$$\frac{g_S^2\gamma_S^2 r_S(g_S\gamma_S+v)(\theta_{NS}r_S-\theta_S r_{NS})+2\theta_S k_S w_S(\theta_{NS}-\theta_S)}{2\theta_S^2 k_S},$$

we get that for the following ranges of k_S, w_{NS}, v this corner is valid too.

$$\begin{aligned}
&k_S > \frac{g_S^2\gamma_S^2 r_S(g_S\gamma_S+1)(\theta_S r_{NS}-\theta_{NS}r_S)}{2\theta_S w_S(\theta_{NS}-\theta_S)},\\
&v < \frac{2\theta_{NS}k_{NS}w_{NS}(\theta_{NS}-\theta_S)}{g_{NS}^2\gamma_{NS}^2 r_{NS}(\theta_S r_{NS}-\theta_{NS}r_S)} - g_{NS}\gamma_{NS},\\
&\tfrac{g_{NS}^2\gamma_{NS}^2 r_{NS}(g_{NS}\gamma_{NS}+1)(\theta_S r_{NS}-\theta_{NS}r_S)}{2\theta_{NS}k_{NS}(\theta_{NS}-\theta_S)} < w_{NS} < \tfrac{g_{NS}^2\gamma_{NS}^2 r_{NS}\left(g_S^2\gamma_S^2 r_S(\theta_{NS}r_S-\theta_S r_{NS})(g_S\gamma_S-g_{NS}\gamma_{NS})+2\theta_S k_S w_S(\theta_{NS}-\theta_S)\right)}{2\theta_{NS}g_S^2\gamma_S^2 k_{NS}r_S(\theta_{NS}-\theta_S)},\\
&g_S < \frac{\gamma_{NS}}{\gamma_S}.
\end{aligned}$$

Notice that these regions are different from the interior and corner 1, as derived in the previous two cases. Hence, overall, $\theta_S r_{NS}-\theta_{NS}r_S > 0$ is necessary to have all three corners exist. If this condition is true, then according to each case, further regions established by k_i, w_i and v provide existence of contracts

that are first- or second-best. □