DYNAMIC INCENTIVE DESIGN FOR MANAGING STRATEGIC MANIPULATION UNDER CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

When algorithmic predictions affect users' outcomes, people may benefit from manipulating their inputs. How should an organization design incentives when these inputs are also used to train future models? We study a setting where a firm chooses costly incentives around a continually updated predictor, taking the training procedure as given. Incentives affect both current operating payoff and future learning. We characterize this trade-off and derive bounds on the dynamic value of departing from the myopic incentive. In a continual learning system, retained information also affects future training: organizations with the same predictor can share the same myopic optimum, yet different histories can make their optimal dynamic incentives lie on opposite sides of the myopic point. We also examine how incentive design interacts with strategy-aware learning. For small training updates, the two interventions are complementary when incentives increase the gain from the update, and substitutes when they reduce it. To examine the economic implications, we use a stylized model to connect our analysis to managerial agency problems and worker welfare. We then provide simulation evidence on the gains from dynamic incentive design. In our benchmark, an approximate history-aware oracle improves net discounted payoff by 6.53% over no intervention, compared with 1.38% for myopic control and 3.19% for current-information control. We further test whether policy learning can recover oracle gains using only observable information, and find that including retained-history information in fitted Q-iteration can achieve outcomes similar to the oracle benchmark.

1 INTRODUCTION

Algorithms guide real-world decisions, including credit decisions and targeted offers. When a prediction affects an individual's payoff, it also creates an incentive to change the inputs used to generate the prediction. A worker may inflate recorded activity; a customer may change behavior to obtain a more favorable offer. The resulting data therefore also reflect responses to the prediction rule, not just the relationship the algorithm was intended to learn (Hardt et al., 2016; Björkegren et al., 2026; Cong & Zhao, 2025).

The strategic-classification literature usually addresses these issues through the learning method itself (Hardt et al., 2016; Levanon & Rosenfeld, 2021; Dong et al., 2018). Similarly, Björkegren et al. (2026) directly estimate how individuals respond to an algorithm and account for those responses when choosing a policy rule. We instead study a different lever: the incentives provided to users. For example, reporting requirements and compensation rules affect the return to manipulation. Our focus is on these incentives around a continually updated system, where incentives also change the observations used for future retraining. An intervention that improves today's operations can therefore have a different effect on future payoff.

We ask two questions: (1) *How can incentive design improve operations under strategic responses?* (2) *How can such incentives be designed when the system is continually updated?* In our model, the objective is not to minimize manipulation or maximize prediction accuracy in isolation, but to align with the firm's goal of maximizing profit. The firm chooses an organizational incentive while taking the training and data-retention procedures as given. In principle, the firm can combine incentive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

design with strategy-aware learning. We examine this possibility and derive conditions under which those two methods are substitutes or complements.

We characterize two trade-offs using our framework. First, the firm balances current payoff against future learning. Myopic optimization ignores how incentive-induced changes in data distribution affect future learning. Second, the firm balances the long-run gains from steering against the lower transition costs from holding. We show that neither preserving the current learning state nor pursuing the long-run optimum necessarily maximizes the dynamic objective function.

A unique feature of our framework is that we allow retained historical data to affect future training as in a continual learning system. Retained data affect how the updated algorithm responds to new observations, so the same incentive can have different effects on future learning across histories. We characterize these history-dependent incentive choices and show that even though organizations with the same predictor share a common myopic optimum, their optimal dynamic incentives can lie on opposite sides of it.

We then use a stylized model to examine the economic implications of such incentives in an organizational setting where strategic workers can inflate their reported effort. Here, a manipulation-deterring incentive can reduce workers' current misreporting. However, as the algorithm updates, retraining can also reward the reported signals more as they become less noisy, which in turn increases the return to manipulation in future periods. We extend our analysis to managerial agency problems and worker welfare. In our contracting model, incentives that sacrifice current payoff for future learning require deferred bonuses. The cost of compensating managers for risk can also lead to weaker incentives than in the first-best. Reporting requirements also burden honest workers, so a tighter requirement that narrows allocation gaps between honest and strategic workers can also lower utility for both groups.

Finally, we provide numerical evidence on the gains from dynamic incentive design. In an illustrative environment, the history-aware oracle benchmark earns 6.53% higher payoff than no intervention, compared with 1.38% for myopic control and 3.19% for current-information control. We then test whether policy learning using only observables can recover these gains. Standard fitted Q-iteration (FQI), after incorporating retained-history summaries, learns a policy earning 6.49% higher payoff than no intervention, close to the oracle benchmark. For robustness, we also examine how optimal gains vary with incentive costs, replay, and the fraction of strategic workers.

Related work. The first strand of literature is strategic classification, which studies how to learn prediction rules that account for individuals' strategic responses (Kleinberg & Raghavan, 2019; Milli et al., 2019; Miller et al., 2020; Jagadeesan et al., 2021). Levanon & Rosenfeld (2021) incorporate these responses into classifier training and also consider jointly choosing the classifier and the costs of modifying inputs. Cong & Zhao (2025) account for strategic consumer behavior when learning targeting policies. Closely related work studies using audits to discourage misreporting (Estornell et al., 2021). We instead study incentive design as a dynamic problem when the prediction system is continually updated: incentives change not only current behavior but also the data retained for future training. Our focus is how these learning effects influence optimal incentives. We also examine when organizational incentives substitute for or complement strategy-aware learning.

Our paper also relates to performative prediction, which studies how deployed models change the data distribution, including when responses depend on the population's current state (Perdomo et al., 2020; Brown et al., 2022; Mendler-Dünner et al., 2020; Miller et al., 2021; Izzo et al., 2021; 2022). The trade-off between current performance and future learning also arises in human-AI interaction and iterative reinforcement learning from human feedback (Noti et al., 2025; Gauthier et al., 2026). We examine this trade-off through organizational incentive design while taking the retraining procedure as given. Our analysis also connects to continual learning, where past observations are commonly reused in subsequent training through replay (Chaudhry et al., 2019; van de Ven et al., 2022; Rolnick et al., 2019; Lopez-Paz & Ranzato, 2017; Aljundi et al., 2019). We show why these retained observations matter for incentive design: organizations with identical current predictors can have different optimal incentives because their training histories differ.

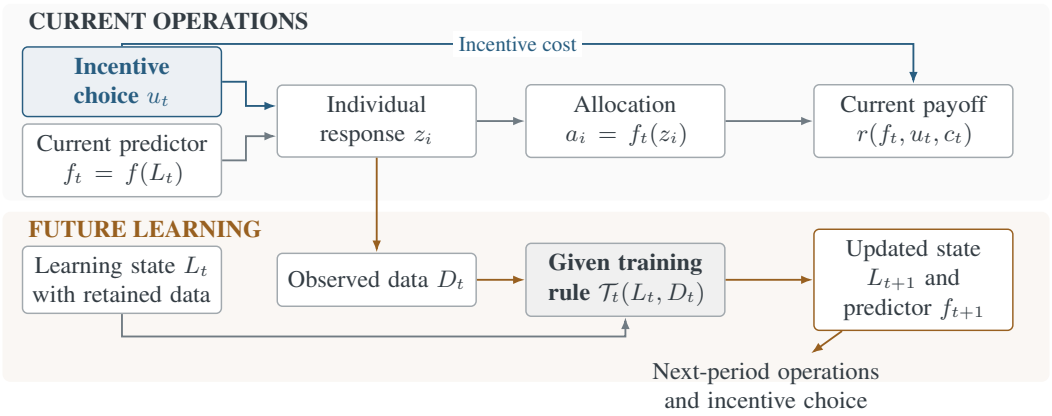

Figure 1: Incentives affect current operations and future learning. The firm chooses the incentive while taking the training procedure as given. Training uses the observed inputs and delayed individual outcomes, together with retained historical data.

2 Environment

A firm uses algorithmic predictions to make operational decisions. Each individual i has an outcome variable y_i, such as service quality, and observed features z_i that are informative about it. The individual outcome affects the firm's operating process and payoff but becomes observable only at the end of the period. The firm therefore uses an algorithm f to predict the outcome from observed features before making its allocation decision. Because predictions affect individuals' allocations and payoffs, individuals have incentives to respond strategically by changing their features z_i. In each period, the firm chooses a common incentive $u_t \in \mathcal{U}$, such as the intensity of a reporting requirement. This incentive affects both individuals' current behavior and the observations used for future training. Our framework allows general strategic responses, with manipulation as the main application.

Manipulation as an example. A worker has a productive input x_i that the firm cannot directly observe. The worker can change the reported feature z_i at a private cost, but the individual outcome y_i depends on x_i, not directly on the report. Inflating recorded activity may therefore increase the worker's predicted outcome without improving actual service quality. The firm can use reporting requirements or audit-based penalties to make manipulation more costly. Section 4 uses this setting to study the economic implications of incentive design.

The organization uses a continual learning system, so the current predictor alone does not fully describe its learning state. Let $L_t \in \mathcal{L}$ denote this state, including model parameters, retained observations, and any other information stored by the training algorithm, and let $f_t = f(L_t)$ be the deployed predictor. Observed operating conditions, such as demand or task composition, are summarized by context c_t, which follows a given Markov process. Letting $\mathcal{I}_t$ denote the set of individuals in period t, the new data available at the end of the period are $D_t = \{(z_{it}, y_{it}) : i \in \mathcal{I}_t\}$. They contain observed features and delayed individual outcomes, but not private information such as effort, types, or manipulation. The distribution of these records and the training update satisfy

$$D_t \sim P(\cdot \mid f(L_t), u_t, c_t), \qquad L_{t+1} = \mathcal{T}_t(L_t, D_t) \tag{1}$$

The firm takes the training procedure $\mathcal{T}_t$ as given, including learning rates and rules for retaining and reusing observations. Replay reuses retained observations in later training, so past data can affect future updates, introducing history dependence through the training process.

The prediction determines the allocation $a_i = f_t(z_i)$. Write $r(f, u, c)$ for the firm's expected payoff net of the cost $K(u)$ of implementing incentives. For example, an allocation problem may have quadratic payoff $r(f, u, c) = \mathbb{E}[\, y_i f(z_i) - f(z_i)^2/2 \,] - K(u)$. Here, the firm maximizes its own payoff without internalizing individuals' private costs. Because incentives are costly, maximizing this objective need not minimize prediction error or manipulation alone.

In each period, the firm chooses the incentive, individuals respond, and the firm makes allocations. Outcomes are then observed and the algorithm is retrained, as shown in Figure 1. For a problem with horizon T, discount factor $0 < \delta < 1$, and terminal value $V_T = 0$, the value function is

$$V_t(L, c) = \max_{u \in \mathcal{U}} \{r(f(L), u, c) + \delta \mathbb{E}[V_{t+1}(\mathcal{T}_t(L, D), c') \mid L, c, u]\} \tag{2}$$

Our analytical specification assumes that the firm knows the full learning state L, the training procedure $\mathcal{T}_t$, and how individuals respond to the predictor and incentives. These information requirements are unlikely to hold in practice. This specification serves as an oracle benchmark, and in Section 5, we use simulations to examine whether policy learning based on observed data and partial information can approach oracle performance.

3 Dynamic Incentive Design

3.1 Trade-offs When Designing Incentives

In our setting, incentive design faces two trade-offs. First, an incentive that improves current operations can have a different effect on future learning. Second, steering the learning system toward a better long-run outcome may lead to lower payoffs during the transition. We examine how these two trade-offs affect optimal incentives.

Static payoff versus future learning. Within a given period, changing the incentive does not affect the deployed predictor or the current aggregate state. Holding these fixed, we write the current payoff as $r(u)$. The dynamic objective is

$$C(u) = \mathbb{E}[V_{t+1}(\mathcal{T}_t(L, D), c') \mid L, c, u], \qquad J(u) = r(u) + \delta C(u) \tag{3}$$

Here $C(u)$ is the expected value of future periods when the firm chooses optimally after today's incentive, using the same fixed discount factor δ. Let u_{myop} maximize current payoff r, and let u_{dyn} maximize the dynamic objective J. At a differentiable interior optimum,

$$\underbrace{-r'(u_{\mathrm{dyn}})}_{\text{marginal current payoff sacrificed}} = \underbrace{\delta C'(u_{\mathrm{dyn}})}_{\text{marginal future payoff gained}} \tag{4}$$

Increasing the incentive may sacrifice current payoff to improve future payoff. Myopic optimization ignores these future effects and sets the marginal current payoff to zero. The following proposition characterizes how much the firm can gain by accounting for future learning.

Proposition 1 (The value of dynamic incentives)**.** *Fix $\delta \in (0, 1)$. Suppose r is strictly concave with interior maximizer u_{myop}, and the regularity conditions in Appendix A.1 hold. Define $\zeta = \delta C'(u_{\mathrm{myop}})$. If $J = r + \delta C$ is strictly concave, its unique maximizer satisfies $\mathrm{sgn}(u_{\mathrm{dyn}} - u_{\mathrm{myop}}) = \mathrm{sgn}(\zeta)$. If additionally $0 < \underline{\kappa} \le -J''(u) \le \overline{\kappa}$, the difference between incentives and the gain from the dynamic choice satisfy*

$$|\zeta|/\overline{\kappa} \le |u_{\mathrm{dyn}} - u_{\mathrm{myop}}| \le |\zeta|/\underline{\kappa}, \qquad \zeta^2/(2\overline{\kappa}) \le J(u_{\mathrm{dyn}}) - J(u_{\mathrm{myop}}) \le \zeta^2/(2\underline{\kappa}) \tag{5}$$

At the myopic optimum, a small change in the incentive has no first-order effect on current payoff, but it can still change future payoff. The dynamic incentive is therefore higher when a stronger incentive improves future payoff, and lower when it worsens it. The gain from choosing the dynamically optimal incentive depends on the curvature of the firm's discounted payoff.

To provide more insights on how incentives improve future learning, Appendices A.5.2 and A.5.3 use a stylized scalar model with exogenous replay. Replay changes how important new observations are for future learning. In our stylized example, increasing replay can change an incentive's future effect from beneficial to harmful. Thus, replay affects not only the strength of the learning effect but also whether accounting for future learning calls for stronger or weaker incentives.

Holding versus steering. When designing incentives over time, the firm also faces a choice between preserving the current learning state and steering it to improve the long-run outcome. To examine this trade-off, consider an infinite-horizon stationary environment. A holding incentive u^{H} preserves the current state as a fixed point under updating: $F(L, u^{\mathrm{H}}) = L$. Steering instead seeks to move

the system toward a desired state. For a convergent incentive sequence whose per-period payoffs converge to r_∞, the resulting value can be written as

$$V_\pi(L_0) = \underbrace{\frac{r_\infty}{1-\delta}}_{\text{value of the long-run payoff}} \underbrace{- \sum_{t\geq 0} \delta^t (r_\infty - r_t)}_{\text{transition cost}} \tag{6}$$

The transition cost measures payoff forgone relative to the long-run outcome. For sufficiently patient firms, steering outperforms holding when it achieves a higher long-run payoff. However, steering toward the state with the highest long-run payoff need not maximize the dynamic objective when the transition is costly. Therefore, neither holding nor steering to the long-run optimum is necessarily optimal. Appendix A.4 gives the conditions for these comparisons.

3.2 The Value of History

Continual learning introduces another dimension of complexity: future learning can depend on retained history. When training reuses past observations, two organizations with the same predictor can obtain different updates from the same new data. Consider learning states L_1, L_2 with different retained data but the same predictor f_0. All else being equal, the organizations share the same myopic optimum. Under the conditions of Proposition 1, if $C_1'(u_{\text{myop}}) > 0 > C_2'(u_{\text{myop}})$, their optimal incentives satisfy $u_1^* > u_{\text{myop}} > u_2^*$. Near the myopic choice, the same increase in the incentive improves future payoff in one organization but worsens it in the other. Differences in retained history can therefore place the optimal incentives on opposite sides of the myopic choice. Proposition 4 constructs such histories and identifies when replay produces opposite choices in a stylized model.

A natural follow-up question is how much the firm loses if it cannot distinguish these histories. We answer this question in the following proposition.

Proposition 2 (The value of observing retained history)**.** *Suppose two histories share the same current predictor f. If J_1, J_2 are μ-strongly concave, with maximizers u_1^*, u_2^* and $\Delta_u = |u_1^* - u_2^*|$, then*

$$\Delta_H = \frac{\max_u J_1(u) + \max_u J_2(u)}{2} - \max_u \frac{J_1(u) + J_2(u)}{2} \geq \frac{\mu \Delta_u^2}{8} \tag{7}$$

For the upper bound, if r, C_1, C_2 are continuously differentiable and $\varepsilon_C = \sup_u |C_1'(u) - C_2'(u)|$, then $\Delta_H \leq \delta^2 \varepsilon_C^2/(8\mu)$.

This proposition considers a firm facing two equally likely histories that produce the same current predictor. Without observing retained history, the firm cannot distinguish between them. The uninformed firm therefore chooses a common incentive to maximize the average payoff across the two histories, which may incur a loss from using the same incentive across different histories. The comparison assumes that history is revealed after the current period. If it remains unobserved, the firm may incur further losses.

3.3 Strategy-Aware Learning and Incentives: Substitutes or Complements?

So far, the firm has taken the training procedure as given. In principle, it can also use strategy-aware learning methods that account for individuals' responses when choosing the prediction rule, as in strategic classification (Levanon & Rosenfeld, 2021; Björkegren et al., 2026). In this section, we ask whether this approach reduces the need for organizational incentives or makes them more valuable.

To simplify the analysis, consider two periods with a given current predictor f_0 at $t = 1$. The firm chooses today's incentive u, while we fix the terminal-period incentive at u_2 under both training methods. Let $\widehat{w}_{\text{O}}(u) \in \mathbb{R}^p$ denote the updated parameters obtained by ordinary training using observations generated under incentive u at $t = 1$. Let $\Pi(w) = r(f_w, u_2)$ be the last-period payoff under the updated predictor f_w. We consider a strategy-aware update of the form

$$\widehat{w}_{\text{S},\eta}(u) = \widehat{w}_{\text{O}}(u) + \eta \nabla \Pi(\widehat{w}_{\text{O}}(u)), \qquad \eta > 0 \tag{8}$$

This update is a special case of the gradient-based approach from strategic classification: candidate predictors are evaluated after individuals respond, and the gradient accounts for those responses (Levanon & Rosenfeld, 2021). The training method affects only next-period payoff through the

updated predictor. The objectives under the two methods are $J_{\mathrm{O}}(u) = r(f_0, u) + \delta\Pi(\widehat{w}_{\mathrm{O}}(u))$ and $J_{\mathrm{S},\eta}(u) = r(f_0, u) + \delta\Pi(\widehat{w}_{\mathrm{S},\eta}(u))$. Define the gains from incentive design as $\mathcal{B}_{\mathrm{O}} = \max_u J_{\mathrm{O}}(u) - J_{\mathrm{O}}(0)$ and $\mathcal{B}_{\mathrm{S},\eta} = \max_u J_{\mathrm{S},\eta}(u) - J_{\mathrm{S},\eta}(0)$. The interventions are complements if $\mathcal{B}_{\mathrm{S},\eta} > \mathcal{B}_{\mathrm{O}}$, and substitutes if $\mathcal{B}_{\mathrm{S},\eta} < \mathcal{B}_{\mathrm{O}}$.

Proposition 3 (The role of strategy-aware learning)**.** *Suppose J_{O} is uniformly strongly concave with maximizer u_{O}, and the regularity conditions in Appendix A.3 hold. Define $\kappa_{\mathrm{O}} = -J''_{\mathrm{O}}(u_{\mathrm{O}}) > 0$ and $\Gamma(u) = \|\nabla\Pi(\widehat{w}_{\mathrm{O}}(u))\|^2$. For a sufficiently small positive η, the optimal incentives and gains under the two updating methods satisfy*

$$u_{\mathrm{S},\eta} - u_{\mathrm{O}} = (\delta\eta/\kappa_{\mathrm{O}})\Gamma'(u_{\mathrm{O}}) + O(\eta^2), \qquad \mathcal{B}_{\mathrm{S},\eta} - \mathcal{B}_{\mathrm{O}} = \delta\eta\{\Gamma(u_{\mathrm{O}}) - \Gamma(0)\} + O(\eta^2) \quad (9)$$

Thus the interventions are complements for sufficiently small steps if $\Gamma(u_{\mathrm{O}}) > \Gamma(0)$, and substitutes if $\Gamma(u_{\mathrm{O}}) < \Gamma(0)$.

The proposition gives general conditions for complementarity and substitution. Appendix A.5.5 gives an intuitive explanation of it. Incentives reduce manipulation today, but retraining on less-inflated reports raises future predictions. If the updated algorithm already overpredicts outcomes, incentives worsen this future problem. Strategy-aware learning corrects part of that adverse effect, making incentives more valuable. The interventions are therefore complements: strategy-aware learning helps more when incentives are used than when they are absent. In the example, this case can arise with greater replay or lower historical variance. In contrast, if incentives reduce underprediction, they already correct part of the problem addressed by strategy-aware learning. The benefits then overlap, making the interventions substitutes.

4 Managing Manipulation in Organizations: A Stylized Model

To illustrate the economic implications of these learning effects, we use a stylized scalar model to study managerial compensation and worker welfare when workers can inflate reported effort. Each worker has exogenous productive input $x > 0$, normalized so that $\mathbb{E}[x^2] = 1$, and produces $y = \theta x + \epsilon$, where $\theta > 0$ measures productivity and ϵ is outcome noise. A fraction α are strategic, identified by $d = 1$, while honest workers have $d = 0$. Strategic workers choose inflation g and have heterogeneous manipulation-cost coefficients $c_g > 0$. Reports are $z = (1 + dg)x$, and the allocation is $a = \beta z$. We hold productive input fixed to isolate reporting responses.

Workers receive benefit ba from their allocation, where $b > 0$ is a constant. To discourage manipulation, the firm chooses a reporting requirement $u \in \mathcal{U} = [0, \overline{u}]$ at implementation cost $ku^2/2$. Reporting imposes a burden uz on workers, which can be understood as the time and effort required to document reported activity. Manipulation has an additional private cost $c_g x g^2/2$, representing the increasing difficulty of fabricating or exaggerating activity. Where manipulation is positive, a strategic worker's optimal choice and the firm's payoff are

$$g(\beta, u; c_g) = (b\beta - u)/c_g, \qquad r(\beta, u) = \mathbb{E}[ya - a^2/2] - ku^2/2 \quad (10)$$

Write $s = b\beta - u > 0$ for the net return to reporting, and define $m_1 = \mathbb{E}[d/c_g]$ and $m_2 = \mathbb{E}[d/c_g^2] < \infty$. A worker's inflation is ds/c_g, so $m_1 s$ is average inflation and $m_2 - m_1^2 = \mathrm{Var}(d/c_g)$ measures differences in responsiveness to this return.

For simplicity, assume the firm trains on observed reports and outcomes using ordinary least squares. Write the resulting coefficient as $\widehat{\beta}(u)$. The deployed coefficient is updated to $\beta^+(u) = (1 - \eta)\beta + \eta\widehat{\beta}(u)$, a convex combination of old and new coefficients, with a given update rate $0 < \eta \leq 1$. Training combines current and retained observations as specified in Appendix B.

To isolate history's role, use the normalized two-period homogeneous specification in Appendix A.5.1. Its histories share fitted coefficient β_0 but differ in historical inflation variance $v \in [0, \overline{v}]$.

Proposition 4 (Replay and history-dependent incentives)**.** *Fix $\delta \in (0, 1)$, $\beta_0 \in (\beta_*, 1)$, and $k > \beta_0/\beta_*$, where $\beta_* = (\sqrt{5} - 1)/2$ maximizes terminal payoff. For the family of history defined in Appendix A.5.1, a unique threshold $\omega_{\mathrm{c}}(v) \in (0, 1)$, strictly increasing in v, determines any optimal first-period incentive $u^*(v, \omega)$: for $\omega \in (0, 1)$,*

$$\omega \lessgtr \omega_{\mathrm{c}}(v) \quad \Longrightarrow \quad u^*(v, \omega) \gtrless u_{\mathrm{myop}} \quad (11)$$

At equality, $u^* = u_{\mathrm{myop}}$. For $v_{\mathrm{low}} < v_{\mathrm{high}}$, replay between their thresholds implies $u^*(v_{\mathrm{low}}, \omega) < u_{\mathrm{myop}} < u^*(v_{\mathrm{high}}, \omega)$.

Appendix A.5.3 proves the result. At the myopic point, training only on new data leads to underprediction after workers respond, whereas the retained-data fit leads to overprediction. Stronger incentives raise the updated coefficient, helping in the first case but hurting in the second. Higher variance reduces retained data's weight and gives new observations more influence, which therefore can favor an incentive above the myopic.

4.1 Managerial Agency Problem

Choosing the optimal incentive and implementing it within an organization are different problems. Misaligned interests between shareholders and managers can create agency problems associated with managerial short-termism (Stein, 1989; Edmans et al., 2012). Here, we study how shareholders design managerial compensation to implement incentives. The manager chooses a reporting requirement $u \in \mathcal{U}$. Enforcing the requirement takes managerial effort, with private cost $c_M(u)$, such as time spent reviewing documentation and supervising compliance. We assume effective enforcement is not contractible.

Shareholders therefore base compensation on operating outcomes. Let Y_0 measure realized current operating payoff with mean $r(\beta, u)$, and let Y_1 measure realized discounted future payoff with mean $\delta C(u)$. Shareholders offer $w_M = w_0 + \lambda_0 Y_0 + \lambda_1 Y_1$: unrestricted fixed pay w_0 plus nonnegative bonus rates λ_0, λ_1. The manager is risk averse and therefore requires additional compensation for exposure to random shocks outside their control. Under the assumptions in Appendix B.6, the manager chooses $u_M \in \arg\max_{u\in\mathcal{U}}\{\lambda_0 r(\beta, u) + \lambda_1 \delta C(u) - c_M(u)\}$. Shareholders maximize expected payoff, subject to incentive compatibility and participation constraints. Let $\mathcal{R}_L(u)$ be the least risk premium required to implement u, setting it to $+\infty$ if implementation is infeasible.

Proposition 5 (Compensation for incentives that improve future learning)**.** *Under the above assumptions and regularity conditions, for any $u \in (0, \overline{u})$ with $r_u(\beta, u) < 0 < \delta C'(u)$, so that a stronger reporting requirement reduces current payoff but improves future payoff, the least-cost linear contract has $\lambda_0^*(u) = 0$ and $\lambda_1^*(u) = c_M'(u)/[\delta C'(u)]$. Shareholders choose a second-best requirement $u_{\mathrm{SB}} \in \arg\max_{u\in\mathcal{U}}\{J(u) - c_M(u) - \mathcal{R}_L(u)\}$. It is weakly below the first-best requirement $u_{\mathrm{FB}} = \arg\max_{u\in\mathcal{U}}\{J(u) - c_M(u)\}$, which would be chosen if enforcement were contractible.*

Here, enforcement that reduces current payoff requires deferred bonuses; a current-performance bonus discourages it. In addition, the cost of compensating the manager for risk can prevent the firm from achieving the first-best outcome. At a given u, a larger marginal learning benefit $\delta C'(u)$ allows a smaller deferred bonus and reduces the risk premium. History therefore affects both the value of enforcement and the cost of implementing it.

4.2 Allocation Gaps and Worker Welfare

Reporting requirements burden honest workers as well as strategic workers. A stricter measure to deter manipulation can therefore make both groups worse off. As the algorithm is retrained, the effect on their allocations and utilities can also change. To examine these effects, we consider a small permanent increase from u to $u + \varepsilon$, with $\varepsilon > 0$, starting in period 0. Let $\beta_t(u)$ denote the coefficient at date t under the constant requirement u, and write $\beta_{t,u} = \partial\beta_t(u)/\partial u$ for its response to this permanent change. Define $s_t = b\beta_t(u) - u$ and $m_x = \mathbb{E}[x]$. Let $\mathcal{W}_{H,t}, \mathcal{W}_{S,t}$ denote honest and strategic workers' average utilities in period t, and let $\Delta_{a,t}$ be strategic workers' average allocation minus honest workers' average allocation.

Proposition 6 (The distributional effects of a permanent incentive change)**.** *Suppose $0 < \alpha < 1$. Then, at every date,*

$$\begin{aligned} \partial_u \Delta_{a,t} &= (m_x m_1/\alpha)\{(2b\beta_t - u)\beta_{t,u} - \beta_t\} \\ \partial_u \mathcal{W}_{H,t} &= m_x(b\beta_{t,u} - 1), \qquad \partial_u \mathcal{W}_{S,t} = m_x(1 + m_1 s_t/\alpha)(b\beta_{t,u} - 1) \end{aligned} \tag{12}$$

A small permanent increase in the incentive therefore reduces manipulation and narrows allocation and utility gaps, but lowers both groups' utility on impact at $t = 0$. At a later date $t > 0$, if $\beta_{t,u} > \frac{\beta_t}{2b\beta_t - u}$, the same incentive increase instead widens the allocation gap, despite narrowing it initially. If $\beta_{t,u} < 1/b$, the increase still reduces manipulation and both groups' utility.

The direct effect of the requirement is to make reporting more costly. Initially, the predictor has not changed, so strategic workers lose some allocation advantage and both groups bear a greater burden.

Table 1: Payoff gains relative to no intervention in the benchmark history.

Oracle benchmark	Gain (%)	Learned policy	Gain (%)
Myopic oracle	1.38	Myopic regression	1.46
Current-information oracle	3.19	Current-information FQI	3.00
History-aware oracle	6.53	History-aware FQI	6.49

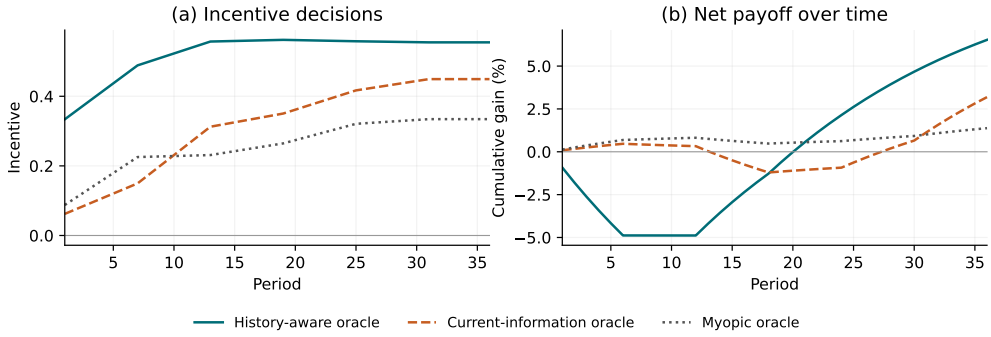

Figure 2: Incentives and payoff in the benchmark history. Left: implemented incentives over time. Right: cumulative discounted payoff gains over time.

Retraining can then raise the coefficient and reward each reported unit more. The first inequality says that the coefficient's response is strong enough to widen the allocation gap; the second says that it is not strong enough to offset the reporting burden, so manipulation and utility still fall. The eventual distributional effect therefore depends on the response generated by retraining.

5 Numerical Evidence and Policy Learning

This section uses simulations to assess the gains from incentive design. We first evaluate these effects using an oracle benchmark. We then vary model primitives to examine the sensitivity of our results. Finally, we ask whether policies learned from observed data can recover the oracle gains.

Simulation setting. Workers can choose productive effort as well as manipulation. They can inflate two recorded inputs and have heterogeneous costs. A neural network predicts individual outcomes using these inputs and additional worker characteristics. Incentives increase the private cost of manipulation and are costly for the firm to implement. The firm's payoff includes output and allocation returns, net of compensation and incentive costs. Incentives are chosen at six review dates over 36 periods and remain fixed between reviews. Appendix C specifies the detailed settings.

Oracle benchmarks. We search numerically for oracle policies from multiple starting points. Because the resulting policies are not certified global optima, we refer to them as approximate oracles. We evaluate different incentive policies for a benchmark training history. The history-aware oracle observes retained data. The current-information oracle cannot initially distinguish this history from an alternative history and maximizes the average objective across them using a common decision rule based on current information. The myopic oracle maximizes current payoff only. Table 1 reports gains from different policies relative to the firm's payoff under no intervention.

Current payoff and future gains. Figure 2 illustrates the evolution of incentives and payoffs over time under different policies. These dynamics reveal the trade-off between current payoff and future learning. History-aware control imposes strong incentives initially, sacrificing current payoff for gains that accumulate over time. By contrast, myopic control delivers immediate gains, but the long-run improvements are limited.

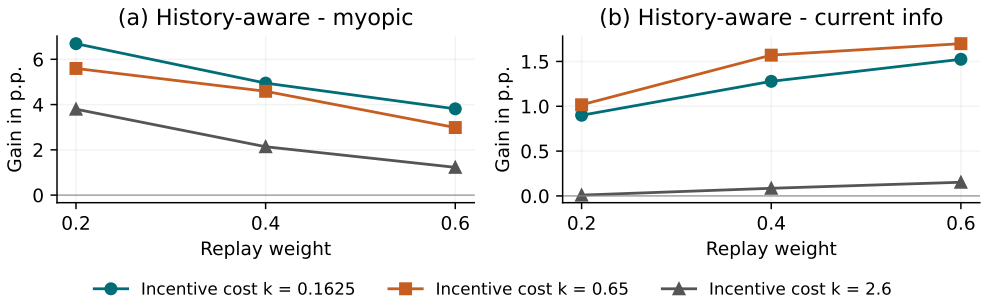

Figure 3: Oracle payoff differences across parameter settings. Left: history-aware minus myopic. Right: history-aware minus current-information. Curves use different incentive costs k. Gains are expressed in percentage points relative to the no-intervention case.

When is dynamic incentive design worthwhile? To examine how our results change with model parameters, we vary incentive costs and replay while holding other settings fixed. Figure 3 compares the history-aware oracle with myopic control on the left and with current-information control on the right. Appendix C.4 provides additional comparisons across different fractions of strategic workers.

Replay affects these two gains differently. A higher replay weight places more weight on retained data and less on new observations. It can reduce the incentive's effect on the next update while making the optimal choice depend more on past data. In the cases shown, the gain over myopic control falls with replay, whereas the gain over current-information control rises. When incentives are most costly, observing history can add little payoff in our example. This is because the strong early intervention preferred before is no longer worthwhile at higher quadratic costs, so history-aware control favors a relatively modest initial intervention, as does current-information control.

Policy learning from observables. The oracle benchmarks assume known behavioral response functions and training dynamics. We next ask whether policies learned from observable data can obtain similar gains. We use finite-horizon fitted Q-iteration (FQI) (Ernst et al., 2005), which estimates the value of each incentive by combining current payoff with estimated future value. We first estimate policies using data generated under randomized incentive sequences, then hold each learned policy fixed and evaluate the trajectory it generates. Appendix C describes the data generation and evaluation process. History-aware FQI achieves a gain of 6.49% in the benchmark history, close to the history-aware oracle's 6.53%. These results illustrate that policy learning can recover much of the oracle gain using observables only. Differences between the learned policies reflect both estimation error and information differences. This exercise demonstrates the feasibility of policy learning in the simulated setting but does not establish general or statistical guarantees for FQI.

6 Discussion and Conclusion

We study how an organization should design incentives when strategic responses affect both current operations and future training data. The optimal incentive depends not only on current payoff but also on how it changes future learning. Retained data in a continual learning system matter for this choice: organizations with the same predictor can have optimal dynamic incentives on opposite sides of the same myopic optimum. The stylized model connects these effects to managerial agency problems and worker welfare. In the simulations, dynamic incentive design improves payoff, and policy learning can recover much of the oracle gain using observed information.

An important direction for future research is to establish statistical properties of policy-learning methods in this setting. Our theory characterizes incentive choice under oracle information. The FQI results provide simulation evidence, not convergence or generalization guarantees. Further work could examine which summaries of training history are sufficient for policy learning, how much data are needed, and whether alternative policy-learning methods improve performance. Another direction is to conduct human experiments to test the framework's predictions and assess their welfare implications.

AI USE DISCLOSURE

Generative AI assistance was used to develop and revise the theoretical formulation, formulate mathematical claims, assist in the writing of proofs, design simulations and suggest simulation parameters, implement and test code, locate references, brainstorming and edit the manuscript. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. Manipulation-robust prediction. *American Economic Review*, 116(9):3263–3293, 2026. doi: 10.1257/aer.20241087.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6045–6061, 2022.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning, 2019. arXiv preprint arXiv:1902.10486.

Yahu Cong and Mingduo Zhao. Game against AI. Working paper, 2025. URL `https://drive.google.com/file/d/1YM1K-IQuVnOsshmZkJ-hJFOzMuYxRWtv/view`.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018. doi: 10.1145/3219166.3219193.

Alex Edmans, Xavier Gabaix, Tomasz Sadzik, and Yuliy Sannikov. Dynamic CEO compensation. *The Journal of Finance*, 67(5):1603–1647, 2012. doi: 10.1111/j.1540-6261.2012.01768.x.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(18):503–556, 2005.

Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. Incentivizing truthfulness through audits in strategic classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6): 5347–5354, 2021. doi: 10.1609/aaai.v35i6.16674.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. Explaining and preventing alignment collapse in iterative RLHF, 2026. arXiv preprint arXiv:2605.04266.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016. doi: 10.1145/2840728.2840730.

Bengt Holmström and Paul Milgrom. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*, 55(2):303–328, 1987. doi: 10.2307/1913238.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4641–4650, 2021.

Zachary Izzo, James Zou, and Lexing Ying. How to learn when data gradually reacts to your model. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3998–4035, 2022.

Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4687–4697, 2021.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 825–844, 2019. doi: 10.1145/3328526.3329584.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6243–6253, 2021.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Celestine Mendler-Dünner, Juan C. Perdomo, Tijana Zrnić, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4929–4939, 2020.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6917–6926, 2020.

John P. Miller, Juan C. Perdomo, and Tijana Zrnić. Outside the echo chamber: Optimizing the performative risk. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7710–7720, 2021.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 230–239, 2019. doi: 10.1145/3287560.3287576.

Gali Noti, Kate Donahue, Jon Kleinberg, and Sigal Oren. AI-assisted decision making with human learning. In *Proceedings of the 26th ACM Conference on Economics and Computation*, pp. 38–39, 2025. doi: 10.1145/3736252.3742492.

Juan C. Perdomo, Tijana Zrnić, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609, 2020.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jeremy C. Stein. Efficient capital markets, inefficient firms: A model of myopic corporate behavior. *The Quarterly Journal of Economics*, 104(4):655–669, 1989. doi: 10.2307/2937861.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022. doi: 10.1038/s42256-022-00568-3.

A Proofs of Results in Section 3

A.1 Proof of Proposition 1

Regularity conditions. Fix $\delta \in (0,1)$. The incentive is within the compact interval $\mathcal{U}$, and r, C are continuously differentiable on a neighborhood of $\mathcal{U}$. The current payoff r is strictly concave with interior maximizer u_{myop}, and $J = r + \delta C$ is strictly concave. To derive the bounds, also assume r, C are twice continuously differentiable, u_{dyn} is interior, and $0 < \underline{\kappa} \le -J'' \le \overline{\kappa}$ throughout $\mathcal{U}$.

Direction and distance. Following standard optimization results, continuity and compactness ensure existence of a maximizer, and strict concavity ensures uniqueness. At the interior myopic optimum, $r'(u_{\text{myop}}) = 0$, so $J'(u_{\text{myop}}) = \zeta = \delta C'(u_{\text{myop}})$. A positive derivative places the dynamic optimum above u_{myop}, a negative derivative places it below, and a zero derivative makes the two optima coincide. This proves the direction statement in Proposition 1.

For an interior dynamic optimum, write $h = u_{\text{dyn}} - u_{\text{myop}}$. Integrating the derivative of J' between the two optima gives the exact identity

$$\zeta = h \int_0^1 [-J''(u_{\text{myop}} + sh)]\, ds \tag{13}$$

The integral lies between $\underline{\kappa}$ and $\overline{\kappa}$, which proves the incentive-distance bounds in (5).

Bounds. For any feasible displacement v, integrating the curvature inequality gives

$$\zeta v - \frac{\overline{\kappa} v^2}{2} \le J(u_{\text{myop}} + v) - J(u_{\text{myop}}) \le \zeta v - \frac{\underline{\kappa} v^2}{2} \tag{14}$$

The right-hand side is at most $\zeta^2/(2\underline{\kappa})$, giving the upper bound on the gain. By the results above, $u_{\text{myop}} + \zeta/\overline{\kappa}$ lies between the two optima and is therefore feasible. Evaluating the left-hand side at this choice gives

$$J(u_{\text{dyn}}) - J(u_{\text{myop}}) \ge J(u_{\text{myop}} + \zeta/\overline{\kappa}) - J(u_{\text{myop}}) \ge \frac{\zeta^2}{2\overline{\kappa}} \tag{15}$$

This proves the payoff-gain bounds in (5). □

This comparison changes today's incentive while allowing optimal choices in all subsequent periods; it does not compare dynamic optimization with a policy that remains myopic throughout.

A.2 Proof of Proposition 2

In this setup, only retained data differ for two histories, while they share the same predictor and also same payoff function $r(u)$ for every incentive.

For a μ-strongly concave function J, strong concavity and optimality imply

$$J(u^*) \ge J((1-s)u^* + su) \ge (1-s)J(u^*) + sJ(u) + \frac{\mu}{2} s(1-s)(u-u^*)^2 \tag{16}$$

Rearrange and let $s \downarrow 0$:

$$J(u^*) - J(u) \ge \frac{\mu}{2}(u-u^*)^2 \tag{17}$$

Applying it to both histories shows that every common incentive satisfies

$$\frac{J_1(u_1^*) + J_2(u_2^*) - J_1(u) - J_2(u)}{2} \ge \frac{\mu}{2}\left(u - \frac{u_1^* + u_2^*}{2}\right)^2 + \frac{\mu}{8}(u_1^* - u_2^*)^2 \tag{18}$$

Minimizing the left side over u proves the lower bound.

For the upper bound, let $\overline{J} = (J_1 + J_2)/2$ and let u^*_{common} maximize it. Strong concavity gives

$$\begin{aligned} J_i(u_i^*) - J_i(u^*_{\text{common}}) &\le J_i'(u^*_{\text{common}})(u_i^* - u^*_{\text{common}}) - \frac{\mu}{2}(u_i^* - u^*_{\text{common}})^2 \\ &\le \frac{\delta \varepsilon_C}{2}|u_i^* - u^*_{\text{common}}| - \frac{\mu}{2}|u_i^* - u^*_{\text{common}}|^2 \le \frac{\delta^2 \varepsilon_C^2}{8\mu} \end{aligned} \tag{19}$$

□

The bounds are sharp: quadratic objectives with a common curvature and different maximizers attain both bounds.

The comparison assumes that history becomes observable after today's choice. Therefore, the upper bound applies only to the initial information restriction, because later information restrictions can create additional losses.

A.3 Strategy-Aware Learning: Setup and Proof

This subsection specifies the setup for strategy-aware learning in Section 3.3 and proves Proposition 3. The terminal incentive u_2 and training step size are given. Appendix A.5 applies Propositions 1–3 to a continuous family of histories with the same fitted initial predictor.

Regularity conditions for Proposition 3. The functions $\Pi, \widehat{w}_{\mathrm{O}}$, and current payoff are C^3 on neighborhoods of the compact sets. For sufficiently small training steps, the update remains feasible, and the payoff function has uniformly bounded curvature along the update: $\|\nabla^2\Pi(w)\| \le L_\Pi$ for some constant $L_\Pi > 0$, at every point w between $\widehat{w}_{\mathrm{O}}(u)$ and $\widehat{w}_{\mathrm{O}}(u) + \eta\nabla\Pi(\widehat{w}_{\mathrm{O}}(u))$. The objective J_{O} is uniformly strongly concave on $\mathcal{U}$ and has an interior maximizer. These conditions give uniform Taylor expansions and ensure that the optimum remains unique and interior for sufficiently small steps.

A.3.1 Training Objectives and the Value of Incentives

Ordinary training minimizes prediction loss using new observations D_t generated under incentive u, and retained observations in L_t:

$$\widehat{w}_{\mathrm{O}}(u) \in \arg\min_{w\in\mathcal{W}} \mathbb{E}[\ell_{\mathrm{pred}}(f_w(z), y) \mid L_t, D_t] \tag{20}$$

For least squares, $\ell_{\mathrm{pred}}(f_w(z), y) = (y - f_w(z))^2$. Ordinary training compares candidates without accounting for how individuals would change their behavior in response to a different predictor; the fitted parameters depend on u through the new data D_t.

The strategy-aware update instead uses the gradient of future payoff:

$$\Pi(w) = r(f_w, u_2) \tag{21}$$

In this two-period setting, this is simply the terminal-period payoff. It accounts for individuals' responses to the candidate predictor f_w, with the terminal incentive u_2 fixed. Its gradient therefore includes both the direct effect of predictions and the resulting behavioral responses.

Under the stated curvature bound, a feasible step with $0 < \eta < 2/L_\Pi$ satisfies

$$\Pi(\widehat{w}_{\mathrm{S},\eta}(u)) - \Pi(\widehat{w}_{\mathrm{O}}(u)) \ge \eta(1 - L_\Pi\eta/2)\Gamma(u) \tag{22}$$

Thus the update improves payoff whenever its gradient is nonzero. The first-order benefit $\Gamma(u) = \|\nabla\Pi(\widehat{w}_{\mathrm{O}}(u))\|^2$ changes with the incentive according to

$$\Gamma'(u) = 2\widehat{w}'_{\mathrm{O}}(u)^\top\nabla^2\Pi(\widehat{w}_{\mathrm{O}}(u))\nabla\Pi(\widehat{w}_{\mathrm{O}}(u)) \tag{23}$$

Retained observations affect both the fitted parameters and their response to incentives, thereby changing the benefit of the update. Proposition 3 compares this payoff-directed strategy-aware update with ordinary prediction training.

A.3.2 Proof of Proposition 3

Fix u, write $w = \widehat{w}_{\mathrm{O}}(u)$, and let $b = \nabla\Pi(w)$. The Hessian bound along the feasible update segment gives

$$\Pi(w + \eta b) \ge \Pi(w) + \eta\|b\|^2 - \frac{L_\Pi\eta^2}{2}\|b\|^2 \tag{24}$$

This proves (22). It is the standard improvement bound for a gradient step.

On the compact incentive set, the regularity assumptions also give the expansion

$$J_{\mathrm{S},\eta}(u) = J_{\mathrm{O}}(u) + \delta\eta\Gamma(u) + O(\eta^2) \tag{25}$$

Strong concavity therefore persists for sufficiently small η, and the maximizer remains interior. Differentiating its first-order condition at $\eta = 0$ yields

$$J''_{\mathrm{O}}(u_{\mathrm{O}}) \left.\frac{du_{\mathrm{S},\eta}}{d\eta}\right|_{\eta=0} + \delta\Gamma'(u_{\mathrm{O}}) = 0 \tag{26}$$

The implicit function theorem gives the first relation in (9). At the new optimal incentive, the change in J_{O} is $O(\eta^2)$ because the incentive differs by $O(\eta)$ from an interior optimum of J_{O}. Evaluating (25) there gives

$$\max_u J_{\mathrm{S},\eta}(u) - \max_u J_{\mathrm{O}}(u) = \delta\eta\Gamma(u_{\mathrm{O}}) + O(\eta^2) \tag{27}$$

Subtracting the corresponding expansion at $u = 0$ proves the second relation in (9). Differentiating the squared gradient gives (23). If the leading term is nonzero, its sign determines the comparison for sufficiently small positive steps. The change in the optimal incentive depends locally on Γ', whereas the change in the value of incentives depends on $\Gamma(u_{\mathrm{O}}) - \Gamma(0)$. □

A.4 Holding, Steering, and the Cost of Transition

We now compare preserving the current learning state with changing it to improve future payoff. Consider a stationary environment with fixed c. Write $r(L, u) = r(f(L), u, c)$, omitting the fixed context, and let $0 < \delta < 1$. A holding incentive u^{H} preserves the initial state if $F(L_0, u^{\mathrm{H}}) = L_0$. Repeating it gives payoff $r_H = r(L_0, u^{\mathrm{H}})$ in every period.

Long-run payoff and transition cost. For a feasible path with bounded payoffs r_t converging to r_∞, adding and subtracting r_∞ in the discounted sum gives

$$\sum_{t\geq 0} \delta^t r_t = \frac{r_\infty}{1-\delta} - \sum_{t\geq 0} \delta^t (r_\infty - r_t) \tag{28}$$

Both sums exist because payoffs are bounded, which gives (6).

Suppose a specified feasible steering path reaches limiting payoff $r_\infty > r_H$ and has summable deviations $\sum_{t\geq 0} |r_\infty - r_t| \leq B < \infty$. Then

$$V_{\mathrm{steer}}(L_0) - V_{\mathrm{hold}}(L_0) \geq \frac{r_\infty - r_H}{1-\delta} - B \tag{29}$$

This steering path therefore outperforms holding whenever $(r_\infty - r_H)/(1-\delta) > B$ for a sufficiently patient firm. The comparison is between two feasible policies. It does not establish convergence to the fixed point with the highest per-period payoff.

Long-run versus dynamic optimality. Suppose r, F are continuously differentiable near the stationary pairs considered, the incentive choices are interior, and $I - F_L$ and $I - \delta F_L^\top$ are nonsingular. Consider a differentiable family of fixed points satisfying $F(L^*(u), u) = L^*(u)$. The long-run optimum maximizes $r(L^*(u), u)$. Differentiating the fixed-point equation gives $dL^*/du = (I - F_L)^{-1} F_u$, so an interior long-run optimum satisfies

$$r_u + F_u^\top (I - F_L^\top)^{-1} r_L = 0 \tag{30}$$

Here F_L and F_u describe how the next state responds to the current state and incentive, and r_L is the column gradient of current payoff. Derivatives are evaluated at the stationary pair being considered.

For a stationary optimum of the dynamic problem, additionally assume that the Bellman value is continuously differentiable and that the envelope and action first-order conditions hold. The envelope equation gives $\nabla_L V_\delta = (I - \delta F_L^\top)^{-1} r_L$. Substituting this into the action condition yields

$$r_u + \delta F_u^\top (I - \delta F_L^\top)^{-1} r_L = 0 \tag{31}$$

Unlike the long-run condition, this condition discounts the effects of today's incentive on future states and payoffs.

The two conditions need not coincide. For a one-dimensional state, the comparison of two conditions leaves

$$\frac{(\delta-1)F_u r_L}{(1-\delta F_L)(1-F_L)} \tag{32}$$

This is nonzero when $0<\delta<1$ and $F_u r_L \neq 0$. Thus the fixed point with the highest long-run payoff need not be optimal for the dynamic discounted payoff.

A.5 Dynamic Incentive Design: A Worked Example

This section uses a stylized example to explain when the conditions behind the results in Section 3 hold.

A.5.1 Historical Data and the Common Initial Predictor

We use the homogeneous special case of Section 4, with two periods, $\alpha=\theta=b=c_g=1$, $\epsilon=0$, and $\mathbb{E}[x^2]=1$. Write the algorithm coefficient as β. Workers choose

$$g(\beta,u)=\max\{\beta-u,0\} \tag{33}$$

The firm chooses the incentive in period 1; the terminal incentive is fixed at $u_2=0$ under both training methods. This is a restriction of the example, not a claim that zero is optimal when terminal incentives can also be chosen. Terminal payoff is

$$\begin{aligned}\Pi(\beta)&=\frac{1}{2}-\frac{1}{2}\{\beta(1+\beta)-1\}^2\\ \Pi'(\beta)&=-\{\beta(1+\beta)-1\}(1+2\beta),\qquad \Pi''(\beta)=1-6\beta-6\beta^2\end{aligned} \tag{34}$$

The coefficient maximizing this payoff is $\beta_*=(\sqrt{5}-1)/2$.

Consider an initial coefficient $\beta_0\in(\beta_*,1)$. Let $p\in[0,\overline{p}]$ be the proportion of retained data with truthful reports, where $\overline{p}=(4\beta_0-2)/(3\beta_0-1)$. Historical reports have three fixed levels and their data distribution is summarized as:

$$\begin{array}{c|ccc}\text{Report} & x & x/\beta_0 & 2x\\ \text{Proportion} & p & 1-p/\overline{p} & p(1-\overline{p})/\overline{p}\end{array} \tag{35}$$

Increasing p replaces moderately inflated reports with a mixture of truthful and highly inflated reports. The retained moments satisfy

$$\begin{aligned}M(p)&=\mathbb{E}[z^2]=\frac{1}{\beta_0^2}\left[1-\frac{1-\beta_0}{2}p\right]\\ N(p)&=\mathbb{E}[zy]=\beta_0 M(p),\qquad M'(p)=-\frac{1-\beta_0}{2\beta_0^2}<0\end{aligned} \tag{36}$$

Every history therefore fits the same initial coefficient $N(p)/M(p)=\beta_0$ under OLS. This construction therefore gives a family of retained histories that all yield the same fitted coefficient β_0.

The firm chooses $u\in[0,\beta_0]$; a larger incentive would leave reports unchanged and increase implementation cost. Let $e_0=\beta_0(1+\beta_0)-1>0$ and $\kappa_0=\beta_0^2+k$. Current payoff and the common myopic choice are

$$r(u)=\frac{1}{2}-\frac{1}{2}(e_0-\beta_0 u)^2-\frac{k}{2}u^2,\qquad u_{\text{myop}}=\frac{\beta_0 e_0}{\kappa_0} \tag{37}$$

Write $h(u)=1+\beta_0-u$ for the current reporting multiplier. With a full coefficient update, $\eta=1$, training with replay weight $\omega\in(0,1)$ gives

$$q_p(u)\equiv\widehat{\beta}_{\mathrm{O},p}(u)=\frac{(1-\omega)h(u)+\omega\beta_0 M(p)}{(1-\omega)h(u)^2+\omega M(p)},\qquad D_p(u)=(1-\omega)h(u)^2+\omega M(p) \tag{38}$$

The algorithm coefficient at $t=2$ is a weighted average of the current-data fit $1/h(u)$ and the initial fit β_0. A larger $M(p)$ effectively gives the retained data more weight, even with the same replay ω.

A.5.2 How Do History and Replay Affect the Training Response?

The histories above can be compared by the variance of historical reporting inflation, rather than by the share p. For the reporting distribution in (35), let $v = \mathrm{Var}(z/x - 1)$. Since $y = x$ here, this variance can also be calculated from the retained ratios $z/y - 1$. Then

$$\begin{aligned} v(p) &= M(p) - \beta_0^2 M(p)^2 \\ \frac{dv}{dp} &= \frac{1-\beta_0}{2\beta_0^2}\{1 - (1-\beta_0)p\} > 0 \end{aligned} \tag{39}$$

Thus v identifies a unique history in this family. At $v = 0$, all historical reports have the same inflation; at higher v, reporting is more dispersed. Write $\overline{v} = v(\overline{p})$, and use $M(v)$ and $q(v, u)$ for the retained moment and updated coefficient of the corresponding history. Higher variance reduces $M(v)$: the mean reporting multiplier also falls, keeping the initial fitted coefficient unchanged. Here variance refers to reporting inflation, not raw reports, whose dispersion also reflects differences in productive input.

History and replay jointly determine the effective weight on the old fit. Define

$$\lambda(v, u) = \frac{\omega M(v)}{(1-\omega)h(u)^2 + \omega M(v)}, \qquad q(v, u) = \{1 - \lambda(v, u)\}\frac{1}{h(u)} + \lambda(v, u)\beta_0 \tag{40}$$

The weight λ is implied by the data and replay setting, not an additional parameter. At a given incentive, higher reporting variance reduces λ, whereas more replay increases it. When the new-data fit $1/h(u)$ is below β_0, higher variance therefore lowers the next coefficient, while more replay raises it toward β_0. If the new-data fit exceeds β_0, these effects on the coefficient are reversed.

A stronger incentive changes both the fit to new data and their weight relative to retained data. Differentiating gives

$$\begin{aligned} q_u(v, u) &= \frac{1-\lambda}{h^2} + \frac{2\lambda(1-\lambda)}{h}\left(\beta_0 - \frac{1}{h}\right) \\ &= \frac{1-\lambda}{h^2}\{1 + 2\lambda(\beta_0 h - 1)\} > 0 \end{aligned} \tag{41}$$

Here $h = h(u)$ and $\lambda = \lambda(v, u)$. The first term is the increase in the new-data coefficient when reports become less inflated. The second comes from reweighting: smaller new reports reduce their second moment, giving retained data more influence. When the old coefficient exceeds the new-data fit, this reweighting also raises the updated coefficient.

These two effects explain how history changes dq/du. Comparing histories or replay weights at a given (β_0, u),

$$\frac{\partial q_u}{\partial \lambda} = \frac{2\beta_0 h - 3 - 4\lambda(\beta_0 h - 1)}{h^2} \tag{42}$$

If $\beta_0 h \leq 3/2$, more replay or lower historical variance reduces the response to today's incentive. Higher variance then makes incentive-induced changes in new reports more influential. If $\beta_0 h > 3/2$, the old coefficient is sufficiently far above the new-data fit that reweighting can instead amplify the response: q_u increases with λ until $\lambda = (2\beta_0 h - 3)/\{4(\beta_0 h - 1)\}$, and decreases thereafter. Thus the initial coefficient determines whether retaining more old data only dampens the training response or can first strengthen it. As ω tends to one, q_u always tends to zero.

A.5.3 How Does Replay Affect Dynamic Incentives?

To apply Proposition 1, fix a retained history v and the initial coefficient β_0, and vary the replay weight ω. Write $C(v, u) = \Pi(q(v, u))$ and $J(v, u) = r(u) + \delta C(v, u)$, leaving their dependence on ω implicit. We first examine the future effect at any given incentive:

$$\begin{aligned} C_u(v, u) &= \Pi'(q)q_u = (\beta_* - q)(q + \beta_* + 1)(1 + 2q)q_u \\ J_u(v, u) &= r'(u) + \delta C_u(v, u) \end{aligned} \tag{43}$$

Here q and q_u are evaluated at (v, u). Since $q_u > 0$, a stronger incentive improves future payoff when $q < \beta_*$ and reduces it when $q > \beta_*$.

Replay at a given incentive. Suppose

$$\frac{1}{h(u)} < \beta_* < \beta_0 \tag{44}$$

The coefficient fitted to new reports alone is then below β_*, while the coefficient fitted to retained data alone is above it. Since $h(u) = 1 + \beta_0 - u$, this condition holds for any $u < \beta_0 - \beta_*$. For such an incentive, the replay weight at which the combined fit equals β_* is

$$\omega_{\mathrm{c}}(v; u) = \frac{h(u)\{\beta_* h(u) - 1\}}{h(u)\{\beta_* h(u) - 1\} + M(v)(\beta_0 - \beta_*)} \tag{45}$$

Consequently,

$$\omega \lessgtr \omega_{\mathrm{c}}(v; u) \quad \Longrightarrow \quad q(v, u) \lessgtr \beta_*, \quad C_u(v, u) \gtrless 0 \tag{46}$$

With less replay, new reports have enough influence to put q below β_*, so stronger deterrence reduces future underprediction by raising the coefficient. With more replay, retained data keep q above β_*, so stronger deterrence worsens future overprediction by raising it further. These prediction errors are evaluated after workers respond to the updated algorithm. If instead $u \geq \beta_0 - \beta_*$, the fit to new data is already at or above β_*. Then $q(v, u) > \beta_*$ and $C_u(v, u) < 0$ for every $0 < \omega < 1$.

Within (44), higher historical variance reduces $M(v)$, lowers $q(v, u)$, and increases the replay cutoff. At the same incentive and replay weight, a higher-variance history can therefore give $q < \beta_*$ and a positive future effect, while a lower-variance history gives $q > \beta_*$ and a negative one.

Proof of Proposition 4. Fix $\delta \in (0, 1)$. The current payoff is strictly concave with interior maximizer u_{myop}. At this choice,

$$J_u(v, u_{\mathrm{myop}}) = \delta C_u(v, u_{\mathrm{myop}}) = \zeta(v) \tag{47}$$

Continuity on $[0, \beta_0]$ ensures that a global maximizer exists. Since $q_u > 0$, if $q(v, u_{\mathrm{myop}}) < \beta_*$, both current and future payoffs increase below u_{myop}, and $J_u(v, u_{\mathrm{myop}}) > 0$; every global maximizer is therefore above it. If $q(v, u_{\mathrm{myop}}) > \beta_*$, both payoffs decrease above u_{myop}, and $J_u(v, u_{\mathrm{myop}}) < 0$; every global maximizer is below it. If $q(v, u_{\mathrm{myop}}) = \beta_*$, the myopic choice uniquely maximizes both components. This argument does not require concavity of the full dynamic objective.

When the myopic choice satisfies

$$\frac{1}{h(u_{\mathrm{myop}})} < \beta_* < \beta_0 \tag{48}$$

the replay comparison above applies with $u = u_{\mathrm{myop}}$. This condition is equivalent to $k > \beta_0/\beta_*$; it ensures that myopic deterrence leaves enough inflation for the new-data fit to be below β_*. Define $\omega_{\mathrm{c}}(v) = \omega_{\mathrm{c}}(v; u_{\mathrm{myop}})$ using (45). The combined fit strictly increases with replay at u_{myop}, so this threshold is unique and lies in $(0, 1)$. Equation (46) gives the incentive ordering in (11). Because $M(v)$ strictly decreases, the threshold strictly increases with variance; any two histories therefore have opposite optimal choices when replay lies between their thresholds. In particular, if

$$\omega_{\mathrm{c}}(0; u_{\mathrm{myop}}) < \omega < \omega_{\mathrm{c}}(\overline{v}; u_{\mathrm{myop}}) \tag{49}$$

then $q(0, u_{\mathrm{myop}}) > \beta_* > q(\overline{v}, u_{\mathrm{myop}})$. The zero-variance history has an optimal incentive below the myopic choice, while the highest-variance history has an optimal incentive above it. The organizations have the same current predictor and payoff, but their retained data change the future consequence of stronger deterrence. □

Under (48), replay below $\omega_{\mathrm{c}}(0; u_{\mathrm{myop}})$ gives $q < \beta_*$ for every history, so all favor stronger incentives than the myopic choice. Replay above $\omega_{\mathrm{c}}(\overline{v}; u_{\mathrm{myop}})$, but below one, gives $q > \beta_*$ for every history, so all favor weaker incentives. Opposite choices across histories therefore arise at intermediate replay. If (48) fails, all histories give $q(v, u_{\mathrm{myop}}) > \beta_*$ and favor weaker incentives for $0 < \omega < 1$.

History also enters the curvature relevant for the gain from dynamic choice:

$$-J_{uu}(v, u) = \beta_0^2 + k - \delta\{\Pi''(q) q_u^2 + \Pi'(q) q_{uu}\} \tag{50}$$

The current-payoff term is common across histories. The remaining terms differ because retained data affect q, q_u, and q_{uu}. Whenever $q = \beta_*$, this curvature is $\beta_0^2 + k + 5\delta q_u^2$: a larger training response makes future payoff fall faster when the incentive moves away from that point. Away from it, the term involving q_{uu} can have either sign. Thus neither a larger training response nor higher historical variance implies a uniformly larger dynamic gain.

A.5.4 When Does Historical Variance Raise or Lower Incentives?

Higher historical variance can increase or decrease the optimal incentive. We show how the direction of this comparison changes with replay. Assume $k > \beta_0/\beta_*$, as in (48).

Two effects of historical variance. Higher variance reduces the influence of retained data. Whenever $1/h(u) < \beta_0$,

$$\begin{aligned} M_v(v) &= \frac{1}{1 - 2\beta_0^2 M(v)} < 0 \\ q_v(v,u) &= \left(\beta_0 - \frac{1}{h}\right) \lambda(1-\lambda) \frac{M_v(v)}{M(v)} < 0 \end{aligned} \tag{51}$$

But variance also changes the response q_u to the incentive. At a differentiable interior optimum with $J_{uu} < 0$, these two effects give

$$\frac{\partial u_{\mathrm{O}}}{\partial v} = \frac{\delta\{\Pi''(q) q_v q_u + \Pi'(q) q_{uv}\}}{-J_{uu}} \tag{52}$$

The first term captures how a change in the coefficient changes the marginal value of raising it. The second captures how history changes the effectiveness of the incentive in moving that coefficient.

Intermediate replay: higher variance can call for stronger incentives. Fix

$$\omega_{\mathrm{c}}(0; u_{\mathrm{myop}}) < \omega < \omega_{\mathrm{c}}(\overline{v}; u_{\mathrm{myop}}) \tag{53}$$

There is a unique interior variance v_{myop} such that $q(v_{\mathrm{myop}}, u_{\mathrm{myop}}) = \beta_*$. At this history, the myopic choice maximizes both current and future payoffs. A small increase in variance lowers the next coefficient below β_*. A stronger incentive can offset this fall, so the optimal incentive increases. Formally,

$$\left.\frac{\partial u_{\mathrm{O}}(v,\omega)}{\partial v}\right|_{v = v_{\mathrm{myop}}} = -\frac{5\delta q_v q_u}{\kappa_0 + 5\delta q_u^2} > 0 \tag{54}$$

At this point $\Pi'(\beta_*) = 0$, so the second effect in (52) vanishes and the first determines the sign. In particular, histories just below v_{myop} favor an incentive below the myopic choice, while histories just above it favor one above.

High replay: higher variance calls for weaker incentives. There exists $\overline{\omega} < 1$ such that, for every $\omega \in (\overline{\omega}, 1)$, the optimum is unique and interior, and $\partial_v u_{\mathrm{O}}(v,\omega) < 0$ for all $v \in (0, \overline{v})$. To get the intuition, consider the case near full replay. The updated coefficient remains close to $\beta_0 > \beta_*$, where raising it worsens future payoff. Higher variance reduces the retained data's weight and makes today's incentive more influential. A stronger incentive then causes a larger unwanted increase in the coefficient, so the firm chooses a weaker one.

To establish this result, expand the optimal incentive as $\omega \uparrow 1$:

$$u_{\mathrm{O}}(v,\omega) = u_{\mathrm{myop}} + (1-\omega)\frac{\delta\Pi'(\beta_0)\{2\beta_0 h(u_{\mathrm{myop}}) - 1\}}{\kappa_0 M(v)} + O\big((1-\omega)^2\big) \tag{55}$$

Since $\Pi'(\beta_0) < 0$, this is a downward departure from the myopic incentive. Higher variance reduces $M(v)$, making that departure larger. More precisely,

$$\frac{\partial u_{\mathrm{O}}}{\partial v} = -(1-\omega)\frac{\delta\Pi'(\beta_0)\{2\beta_0 h(u_{\mathrm{myop}}) - 1\} M_v(v)}{\kappa_0 M(v)^2} + O\big((1-\omega)^2\big) \tag{56}$$

The leading coefficient is strictly negative because $\Pi'(\beta_0) < 0$, $M_v < 0$, and $2\beta_0 h(u_{\mathrm{myop}}) - 1 > 0$. Although higher variance also lowers the coefficient, near full replay its effect on the incentive's effectiveness dominates its effect on the marginal value of the coefficient. Thus the same increase in historical variance that favors stronger incentives around the intermediate-replay reference favors weaker incentives when replay is sufficiently high.

These comparisons explain why observing history matters for incentive design. History determines not just where retraining moves the predictor, but how strongly incentives can affect that movement.

A.5.5 When Does Strategy-Aware Learning Complement Incentives?

The marginal payoff gain from the strategy-aware step is $\Gamma(v,u) = \{\Pi'(q(v,u))\}^2$, with

$$\Gamma_u(v,u) = 2\Pi'(q(v,u))\Pi''(q(v,u))q_u(v,u) \tag{57}$$

Since $\Pi'' < 0$ and $q_u > 0$, incentives increase this training gain when the ordinary fit is above β_*, but reduce it when the fit is below. Lower historical variance or more replay shifts the ordinary fit upward. It can therefore change whether deterrence increases the distortion that strategy-aware learning corrects or already corrects part of it.

Lower-variance history: complements. Suppose

$$\omega_c(v;0) < \omega < \omega_c(\overline{v};0), \qquad v < \overline{v} \tag{58}$$

The retained data have enough influence to keep the ordinary fit above β_* even without incentives. Stronger deterrence raises it further and increases future overprediction. The strategy-aware step then has more to correct, so it makes organizational incentives more valuable. The upper replay bound ensures that this adverse learning effect does not eliminate the positive ordinary incentive. Hence $\Gamma(v, u_O(v)) > \Gamma(v,0)$, giving complementarity.

Higher-variance history: substitutes. Under (48), suppose

$$0 < \omega < \omega_c(v; u_{\text{myop}}) \tag{59}$$

The weaker influence of retained data allows the new reports to pull the fit below β_*. Increasing the incentive then reduces future underprediction by raising the coefficient toward that level. The ordinary optimum lies between the myopic choice and the incentive at which $q = \beta_*$: below the former both payoff components increase, and above the latter both decrease. Incentives therefore leave less for strategy-aware training to improve, giving $\Gamma(v, u_O(v)) < \Gamma(v,0)$ and substitution. Since the cutoff increases with variance, this condition is easier to satisfy for higher-variance histories at the same replay weight.

B The Stylized Organizational Model

B.1 Worker behavior

Let $x > 0$ be productive input, with $\mathbb{E}[x^2] = 1$. The individual outcome is $y = \theta x + \epsilon$, where $\theta > 0$, $\mathbb{E}[\epsilon \mid x, d] = 0$, and $\mathbb{E}[\epsilon^2] < \infty$. Production is exogenous to the predictor and incentive, so the example focuses on strategic reporting rather than effort choice. The indicator d identifies strategic workers and is independent of x, with probability $\alpha \in [0,1]$ of being one. Comparisons between both worker groups require $0 < \alpha < 1$. An honest worker reports $z = x$; a strategic worker can report $z = (1+g)x$, with $g \geq 0$. The algorithm assigns score βz, where $\beta > 0$.

Workers value the allocation at rate $b > 0$, face reporting burden uz, and incur manipulation cost $c_g x g^2/2$, with $c_g > 0$. For example, each reported unit may require a submission step that takes time without verifying whether it is productive. The firm chooses the intensity of this requirement. Relative to reporting honestly, $z = x$, a strategic worker gains expected utility

$$x\left[(b\beta - u)g - \frac{c_g}{2}g^2\right] \tag{60}$$

Here $u \geq 0$ measures the private burden per reported unit, in utility units. The common cost ux cancels when comparing reports, giving

$$g(\beta, u) = \max\left\{\frac{b\beta - u}{c_g}, 0\right\} \tag{61}$$

The reporting requirement changes the private cost of reporting without supplying a verified measure of productive input to training. The allocation benefit is private utility, not an additional wage paid by the firm. Workers bear the reporting burden, while the firm pays a separate implementation cost specified below. Honest workers also bear the burden ux. Fixed production and workforce composition are therefore substantive assumptions: if effort or participation responds to the requirement, the firm must account for those responses as well.

B.2 Firm Payoff and the Static Incentive

With allocation $a = \beta z$, $k > 0$, and organizational implementation cost $ku^2/2$, expected payoff is

$$\begin{aligned} r(\beta, u) &= \mathbb{E}\left[ya - \frac{a^2}{2}\right] - \frac{k}{2}u^2 \\ &= \frac{\theta^2}{2} - \frac{1}{2}\left[(1-\alpha)(\beta-\theta)^2 + \alpha\{\beta(1+g)-\theta\}^2\right] - \frac{k}{2}u^2 \end{aligned} \tag{62}$$

The parameter k measures implementation cost relative to allocation returns. The mean-zero outcome disturbance drops from expected payoff because $\mathbb{E}[x\epsilon \mid d] = 0$. The term $\theta^2/2$ is constant across policies.

In the interior,

$$r_u = \frac{\alpha\beta}{c_g}\{\beta(1+g)-\theta\} - ku, \qquad r_{uu} = -\frac{\alpha\beta^2}{c_g^2} - k < 0 \tag{63}$$

If the available incentive interval contains $[0, b\beta]$, the myopic incentive is

$$u_{\text{myop}}(\beta) = \left[\frac{(\alpha\beta/c_g)\{\beta(1+b\beta/c_g)-\theta\}}{\alpha\beta^2/c_g^2 + k}\right]_{[0,b\beta]} \tag{64}$$

B.3 Training on Current and Retained Data

Define current moments

$$m_{zx}(g) = \mathbb{E}[zx] = 1 + \alpha g, \qquad m_{zz}(g) = \mathbb{E}[z^2] = 1 + 2\alpha g + \alpha g^2 \tag{65}$$

Then $\mathbb{E}[z^2] = m_{zz}(g)$, $\mathbb{E}[zy] = \theta m_{zx}(g)$, and the population least-squares coefficient through the origin is

$$\beta_{\text{OLS}}(g) = \frac{\theta m_{zx}(g)}{m_{zz}(g)} \tag{66}$$

This coefficient fits reported rather than unmanipulated inputs. It is the target of the least-squares procedure used in this model; other estimators need not have the same target.

Retained records have moments $M = \mathbb{E}_{\text{old}}[z^2]$ and $N = \mathbb{E}_{\text{old}}[zy]$. We use m_{zx} only to derive the observed cross moment $\mathbb{E}[zy] = \theta m_{zx}$; training does not observe productive input. Let $\omega \in [0, 1)$ be the replay weight, $\eta \in (0, 1]$ the rate of updating the deployed coefficient, and $\rho \in (0, 1]$ the rate of replacing retained data. The updates are

$$\widehat{\beta} = \frac{(1-\omega)\theta m_{zx}(g) + \omega N}{(1-\omega)m_{zz}(g) + \omega M}, \qquad \beta^+ = (1-\eta)\beta + \eta\widehat{\beta} \tag{67}$$

$$M^+ = (1-\rho)M + \rho m_{zz}(g), \qquad N^+ = (1-\rho)N + \rho\theta m_{zx}(g) \tag{68}$$

These three parameters have different roles. The replay weight determines how much retained data enter the current fit. The replacement rate determines which observations remain available for future training. The coefficient-update rate determines how far the deployed coefficient moves toward the fitted value.

With $M_{\text{train}} = (1-\omega)m_{zz}(g) + \omega M$, differentiation gives

$$\widehat{\beta}_u = \frac{(1-\omega)\alpha}{c_g M_{\text{train}}}\left[2(1+g)\widehat{\beta} - \theta\right], \quad M_u^+ = -\frac{2\rho\alpha(1+g)}{c_g}, \quad N_u^+ = -\frac{\rho\theta\alpha}{c_g} \tag{69}$$

B.4 Dynamic Payoff

Write $L = (\beta, M, N)$ and $F(L, u) = (\beta^+, M^+, N^+)$ for the preceding population transition. With training settings given, the value function satisfies

$$V_t(L) = \max_{u \in \mathcal{U}}\{r(\beta, u) + \delta V_{t+1}(F(L, u))\}, \qquad V_T = 0 \tag{70}$$

Where $0 < u < b\beta$, the transition and continuation value are continuously differentiable, and the optimal incentive is interior, the chain rule gives the necessary first-order condition

$$-r_u = \delta \left[V_{t+1,\beta} \beta_u^+ + V_{t+1,M} M_u^+ + V_{t+1,N} N_u^+ \right] \tag{71}$$

The value derivatives are evaluated at $F(L, u)$, with $\beta_u^+ = \eta \widehat{\beta}_u$, M_u^+, and N_u^+ given above. The first term captures the next coefficient's effect on future payoff; the other terms capture the effect of retained observations holding that coefficient fixed.

In a two-period problem, terminal value is $V_{\text{term}}(\widetilde{\beta}) = \max_{u' \in \mathcal{U}} r(\widetilde{\beta}, u')$. With no subsequent training, the retained-data derivatives vanish and the condition reduces to $r_u + \delta \eta \widehat{\beta}_u V'_{\text{term}}(\beta^+) = 0$. Over longer horizons, incentives also affect later updates through the observations retained today.

B.5 Retraining and Subsequent Responses

For the common-cost special case, write $\beta^+(u) = (1-\eta)\beta + \eta \widehat{\beta}(u)$ for the next coefficient and $M_{\text{train}}(u) = (1-\omega) m_{zz}(g) + \omega M$ for the second moment of reports in the combined training data.

Prediction error before and after workers respond. Let $\mathcal{M}$ be the distribution of retained records. Assuming finite outcome second moments, define prediction error on the training data as

$$R_{\text{fit}}(\widetilde{\beta}; u) = (1-\omega)\mathbb{E}[(\widetilde{\beta} z - y)^2 \mid \beta, u] + \omega \mathbb{E}_{\mathcal{M}}[(\widetilde{\beta} z - y)^2] \tag{72}$$

This criterion evaluates candidate coefficient $\widetilde{\beta}$ on reports generated under the current coefficient and incentive β, u. Workers do not respond anew to each candidate during fitting. The coefficient of $\widetilde{\beta}^2$ is $M_{\text{train}}(u)$, and the linear term is $-2\{(1-\omega)\theta m_{zx}(g) + \omega N\}\widetilde{\beta}$. Thus $\widehat{\beta}(u)$ minimizes the quadratic and

$$R_{\text{fit}}(\widetilde{\beta}; u) - R_{\text{fit}}(\widehat{\beta}(u); u) = M_{\text{train}}(u)\{\widetilde{\beta} - \widehat{\beta}(u)\}^2 \tag{73}$$

Substitute $\widetilde{\beta} = \beta^+(u)$ and $\widetilde{\beta} = \beta$, and use $\beta^+(u) - \widehat{\beta}(u) = (1-\eta)(\beta - \widehat{\beta}(u))$ to obtain

$$R_{\text{fit}}(\beta^+(u); u) - R_{\text{fit}}(\beta; u) = -\eta(2-\eta) M_{\text{train}}(u)\{\beta - \widehat{\beta}(u)\}^2 \leq 0 \tag{74}$$

Retraining reduces prediction error on the weighted combination of new and retained observations. It need not reduce error on each dataset separately.

B.6 Managerial Compensation: Implementation Costs and Optimal Contracts

The manager chooses any $u \in \mathcal{U} = [0, \overline{u}]$ and has the same state information as shareholders. The manager bears the known enforcement-effort cost $c_M(u)$, for example $c_M(u) = \kappa_M u^2/2$ with $\kappa_M > 0$. This cost represents supervision effort and is separate from the firm's expenditure $ku^2/2$. Effective enforcement is not contractible. Shareholders are risk neutral and value $J(u) = r(\beta, u) + \delta C(u)$, minus managerial compensation.

Let Y_0 measure realized current operating payoff and Y_1 measure realized subsequent payoff in present-value terms. These are observable organizational performance measures and therefore contractible. Following the linear-contract approach (Holmström & Milgrom, 1987), restrict compensation to $w_M = w_0 + \boldsymbol{\lambda}^\top \mathbf{Y}$, where $\mathbf{Y} = (Y_0, Y_1)^\top$, $\boldsymbol{\lambda} = (\lambda_0, \lambda_1)^\top \geq 0$, and w_0 is unrestricted. Conditional on u, $\mathbf{Y}$ is normally distributed with mean $\boldsymbol{\mu}(u) = (r(\beta, u), \delta C(u))^\top$ and positive-definite covariance Σ_M, independent of u. These noise assumptions are additional to the worker model and allow common business shocks.

All payments are in present-value units. We abstract from limited liability and renegotiation. The analysis concerns one delegated enforcement decision under a linear contract, holding subsequent policy fixed.

Regularity conditions for Proposition 5. Assume $r(\beta, \cdot), C, c_M$ are twice continuously differentiable on a neighborhood of $\mathcal{U} = [0, \overline{u}]$, both $r(\beta, \cdot)$ and C are concave on this interval, and $c_M(0) = c'_M(0) = 0$ with $c''_M(u) > 0$ throughout. These conditions make the manager's objective strictly concave under nonnegative bonus rates.

Risk assumptions for Proposition 5. The proposition specializes to independent Gaussian performance noise, with $\Sigma_M = \mathrm{diag}(\sigma_0^2, \sigma_1^2)$, where $\sigma_0^2, \sigma_1^2 > 0$. The manager has the constant-absolute-risk-aversion preferences specified below.

Certainty equivalent and shareholder problem. The manager has utility $-\exp[-\gamma_M\{w_M - c_M(u)\}]$, where $\gamma_M > 0$ is constant absolute risk aversion, and reservation certainty equivalent $\overline{U}_M$. A certainty equivalent is the guaranteed net payment equivalent to the risky contract. Under normality it is

$$\mathrm{CE}(u) = w_0 + \boldsymbol{\lambda}^\top \boldsymbol{\mu}(u) - c_M(u) - \frac{\gamma_M}{2}\boldsymbol{\lambda}^\top \Sigma_M \boldsymbol{\lambda} \tag{75}$$

Expected utility equals $-\exp[-\gamma_M \,\mathrm{CE}(u)]$. The shareholders' problem is

$$\begin{aligned} \max_{w_0,\, \boldsymbol{\lambda} \geq 0,\, u \in \mathcal{U}} \quad & J(u) - w_0 - \boldsymbol{\lambda}^\top \boldsymbol{\mu}(u) \\ \text{subject to} \quad & u \in \arg\max_{v \in \mathcal{U}} \{\boldsymbol{\lambda}^\top \boldsymbol{\mu}(v) - c_M(v)\} \\ & w_0 + \boldsymbol{\lambda}^\top \boldsymbol{\mu}(u) - c_M(u) - \frac{\gamma_M}{2}\boldsymbol{\lambda}^\top \Sigma_M \boldsymbol{\lambda} \geq \overline{U}_M \end{aligned} \tag{76}$$

The first constraint is incentive compatibility; the second is participation. Unrestricted fixed pay makes participation bind at an optimum. Hence the expected payment needed to implement u equals

$$\mathbb{E}[w_M \mid u] = \overline{U}_M + c_M(u) + \frac{\gamma_M}{2}\boldsymbol{\lambda}^\top \Sigma_M \boldsymbol{\lambda} \tag{77}$$

Implementation over the full incentive interval. The least risk premium required to implement a candidate u is

$$\begin{aligned} \mathcal{R}_L(u) = \min_{\boldsymbol{\lambda} \geq 0} \quad & \frac{\gamma_M}{2}\boldsymbol{\lambda}^\top \Sigma_M \boldsymbol{\lambda} \\ \text{subject to} \quad & \boldsymbol{\lambda}^\top \{\boldsymbol{\mu}(u) - \boldsymbol{\mu}(v)\} \geq c_M(u) - c_M(v) \quad \text{for every } v \in \mathcal{U} \end{aligned} \tag{78}$$

Set $\mathcal{R}_L(u) = +\infty$ if the constraints are infeasible. When feasible, positive definiteness makes the objective coercive and strictly convex, and the closed constraints give a minimum. Substituting binding participation into (76), shareholders select an implementable requirement to maximize

$$J(u) - c_M(u) - \mathcal{R}_L(u) - \overline{U}_M \tag{79}$$

The fixed salary supporting the minimizing rates is $w_0^* = \overline{U}_M + c_M(u) + \mathcal{R}_L(u) - \boldsymbol{\lambda}^*(u)^\top \boldsymbol{\mu}(u)$. This proves the second-best characterization in Proposition 5, since the reservation value is constant. Under the proposition's concavity assumptions the manager's best response is unique.

Proof of Proposition 5: the implementing contract. Now impose the regularity conditions in Appendix B.6 and independent noise, so $\Sigma_M = \mathrm{diag}(\sigma_0^2, \sigma_1^2)$. For nonnegative rates, the manager's objective $\boldsymbol{\lambda}^\top \boldsymbol{\mu}(v) - c_M(v)$ is strictly concave in v. Consequently an interior requirement u is the unique global best response if and only if

$$\lambda_0 r_u(\beta, u) + \lambda_1 \delta C'(u) = c_M'(u) \tag{80}$$

Write $[a]_+ = \max\{a, 0\}$, let $\mu_j(u)$ be component $j \in \{0, 1\}$ of $\boldsymbol{\mu}(u)$, and define

$$S_L(u) = \sum_{j=0}^{1} \frac{[\mu_j'(u)]_+^2}{\sigma_j^2} \tag{81}$$

For any $u > 0$, strict convexity and $c_M'(0) = 0$ imply $c_M'(u) > 0$. If $S_L(u) > 0$, the least-cost implementing rates and risk premium are

$$\lambda_j^*(u) = \frac{c_M'(u)[\mu_j'(u)]_+}{\sigma_j^2 S_L(u)}, \qquad \mathcal{R}_L(u) = \frac{\gamma_M [c_M'(u)]^2}{2 S_L(u)} \tag{82}$$

To verify optimality, any rates satisfying (80) obey

$$[c_M'(u)]^2 \leq \left(\sum_{j=0}^{1} \lambda_j [\mu_j'(u)]_+ \right)^2 \leq \left(\sum_{j=0}^{1} \sigma_j^2 \lambda_j^2 \right) S_L(u) \tag{83}$$

The first inequality discards nonpositive marginal performance effects; the second is Cauchy–Schwarz. The rates in (82) attain both bounds and satisfy global incentive compatibility by strict concavity. If $r_u(\beta, u) < 0 < \delta C'(u)$, only the deferred measure has a positive marginal effect, giving the bonus rates in Proposition 5 and risk premium

$$\mathcal{R}_L(u) = \frac{\gamma_M \sigma_1^2}{2} \left(\frac{c'_M(u)}{\delta C'(u)} \right)^2 \tag{84}$$

Here γ_M measures aversion to uncertain pay, while σ_1^2 measures the uncertainty in deferred performance. A larger value of either raises the expected compensation needed for the same intended requirement. No current-only contract can implement this interior requirement, because its marginal reward for enforcement is nonpositive while marginal effort cost is positive. □

If enforcement were contractible, shareholders could pay a fixed wage $\overline{U}_M + c_M(u)$ and require the chosen u. Their first-best requirement maximizes $J - c_M$, which is strictly concave under the stated assumptions.

How history affects compensation. Consider implementing the same requirement u under different histories. Suppose stronger enforcement lowers current payoff but raises future payoff, $r_u < 0 < \delta C'$. If retained history makes future payoff more responsive to enforcement, so that $\delta C'(u)$ is larger, a smaller deferred bonus rate is enough to motivate the manager. This reduces the manager's exposure to risk and the compensation needed for bearing it. The comparison concerns the cost of implementing a given requirement; it does not imply that shareholders choose stronger or weaker enforcement across histories. If enforcement instead improves both current and future payoff, the least-cost contract uses bonuses on both measures.

B.7 Proof of Proposition 6: One Permanent Incentive Change

For each constant incentive u, let $L_{t+1}(u) = F(L_t(u), u)$, with the same L_0. Thus changing u affects every period's reporting incentives, not just the current period. By (67),

$$s_t = b\beta_t(u) - u, \qquad B(s_t) = 1 + 2m_1 s_t + m_2 s_t^2, \qquad \mathbb{E}[z_t y_t] = \theta(1 + m_1 s_t) \tag{85}$$

Here $B(s_t) = \mathbb{E}[z_t^2]$, and m_1, m_2 are defined in Section 4.

Allocation and utility along the path. Let $m_x = \mathbb{E}[x] > 0$. An honest worker's utility at date t is $x s_t$. A strategic worker additionally obtains the maximum of $x\{s_t g - c_g g^2/2\}$, which is $x s_t^2/(2c_g)$. Because productive input is independent of costs and strategic status,

$$\begin{aligned} \Delta_{a,t} &= \frac{m_x m_1}{\alpha} \beta_t s_t \\ \mathcal{W}_{H,t} = m_x s_t, \qquad \mathcal{W}_{S,t} &= m_x \left(s_t + \frac{m_1 s_t^2}{2\alpha} \right), \qquad \mathcal{W}_t = m_x \left(s_t + \frac{m_1 s_t^2}{2} \right) \end{aligned} \tag{86}$$

The allocation gap compares groups with the same productive-input distribution. Differentiating s_t gives $s_{t,u} = b\beta_{t,u} - 1$. The product and chain rules then give (12), together with

$$\frac{\partial g_t(c_g)}{\partial u} = \frac{b\beta_{t,u} - 1}{c_g}, \qquad \frac{\partial(\mathcal{W}_{S,t} - \mathcal{W}_{H,t})}{\partial u} = \frac{m_x m_1 s_t}{\alpha}(b\beta_{t,u} - 1) \tag{87}$$

At date 0, the incentive changes but the initial predictor does not, so $\beta_{0,u} = 0$. Stronger enforcement therefore reduces manipulation, narrows allocation and utility gaps, and lowers both groups' utilities. At later dates, retraining changes these effects. The same increase in enforcement widens the allocation gap if $\beta_{t,u} > \beta_t/(2b\beta_t - u)$: the coefficient rises enough to give strategic workers a larger allocation advantage. However, manipulation and both groups' utilities still fall if $\beta_{t,u} < 1/b$, because the higher coefficient does not fully offset the reporting burden. These two conditions can hold simultaneously. Thus, stronger enforcement can eventually widen the allocation gap even while reducing manipulation and lowering both groups' utilities.

The coefficient response generated by retraining. Write $L_{t,u} = \partial L_t(u)/\partial u$. Differentiating the actual state transition gives

$$L_{0,u} = 0, \qquad L_{t+1,u} = F_L(L_t, u) L_{t,u} + F_u(L_t, u) \tag{88}$$

The first term propagates changes to the predictor and retained data; the second is the new effect of the reporting requirement. For the update used here, let $M_{\mathrm{train},t} = (1-\omega)B(s_t) + \omega M_t$. The fitted coefficient has derivative

$$\widehat{\beta}_{t,u} = \frac{(1-\omega)\{\theta m_1 - 2(m_1 + m_2 s_t)\widehat{\beta}_t\}s_{t,u} + \omega(N_{t,u} - \widehat{\beta}_t M_{t,u})}{M_{\mathrm{train},t}} \tag{89}$$

The three components of the next-state derivative are

$$\begin{aligned} \beta_{t+1,u} &= (1-\eta)\beta_{t,u} + \eta\widehat{\beta}_{t,u} \\ M_{t+1,u} &= (1-\rho)M_{t,u} + 2\rho(m_1 + m_2 s_t)s_{t,u} \\ N_{t+1,u} &= (1-\rho)N_{t,u} + \rho\theta m_1 s_{t,u} \end{aligned} \tag{90}$$

Thus the permanent policy change affects current reports, the next fitted coefficient, and the observations retained for later updates.

C Simulation Setup and Results

This appendix specifies the simulation setup and policy-learning results in Section 5. Reported results are rounded to two decimal places; differences are calculated before rounding.

C.1 Worker Choices and Operating Payoff

The simulation allows workers to choose production as well as manipulation. Let $x = (x_1, x_2)$ denote productive inputs, $g = (g_1, g_2)$ proportional inflation, and $h = (h_1, h_2)$ non-manipulable worker characteristics. The firm observes $Z = (z_1, z_2, h_1, h_2)$, with $z_j = (1 + g_j)x_j$. We specify four equally weighted characteristic profiles: $(-0.5, -0.3)$, $(-0.2, 0.5)$, $(0.3, -0.4)$, and $(0.5, 0.4)$. Define $\vartheta_1(h) = 1 + 0.15(h_1 + h_2/2)$ and $\vartheta_2(h) = 1 + 0.15(-h_1/2 + h_2)$. Output is

$$y(x,h) = \frac{\vartheta_1(h)x_1 + \vartheta_2(h)x_2}{2} + 0.08\sin(x_1)\sin(x_2)$$

Worker choices are approximated on a discrete grid. Workers choose each x_j from $\{0, 0.3, 0.6, 0.9, 1.2\}$ and each g_j from $\{0, 0.5, 1, 1.5\}$. Excluding nonzero inflation of a zero input leaves 289 production and reporting plans per profile. A fraction $\alpha = 0.9$ can choose any plan; the remaining workers report truthfully, with $g = 0$.

Private cost multipliers κ_1, κ_2 are drawn once per profile, with $\log \kappa_j \sim N(0, 0.045^2)$. Set the effort-cost coefficient $c_x = 1$, manipulation-cost coefficient $c_g = 0.21$, and cross-term coefficient $\chi = 0.15$. Private cost is

$$\begin{aligned} C_\kappa(x,g) = &\frac{c_x(\kappa_1 x_1^2 + \kappa_2 x_2^2 + 2\chi x_1 x_2)}{4(1+\chi)} \\ &+ \frac{c_g(\kappa_1 g_1^2 x_1 + \kappa_2 g_2^2 x_2 + 2\chi g_1 g_2 \sqrt{x_1 x_2})}{4(1+\chi)} \end{aligned}$$

The incentive raises the private cost of inflated activity. Given prediction $a = f_w(Z)$, where w denotes the predictor's parameters, and incentive u, utility before random shocks is

$$0.44y + 0.45a - C_\kappa(x,g) - u(g_1 x_1 + g_2 x_2)/2$$

Workers also face independent extreme-value utility shocks across their feasible plans.

For output price p, allocation-return scale H, and incentive-cost coefficient k, the firm's payoff is

$$r(w,u) = \mathbb{E}\left[py + H\left(ya - \frac{a^2}{2}\right) - 0.44y - 0.45a\right] - \frac{ku^2}{2} \tag{91}$$

We set $p = 1$, $H = 12$, and $k = 0.65$. The expectation accounts for the distribution of worker types and the idiosyncratic shocks affecting production and reporting choices. Although the simulator uses specified behavioral equations, their private parameters are not supplied to the FQI estimator.

C.2 Prediction and Replay

The predictor is a feedforward neural network with four inputs (z_1, z_2, h_1, h_2), one hidden layer of six tanh units with biases, and a linear output without an output bias. We subtract the network's output at $(0, 0, h_1, h_2)$ from its output at the observed inputs. This gives the raw prediction $\widetilde{f}_w$ and imposes zero prediction when both recorded inputs are zero, consistent with the simulation's production function. The deployed prediction f_w clips this raw output to $[0, 2.5]$.

All 36 parameters are trained by minimizing squared prediction error plus weight decay 10^{-4}. At each review, training takes 12 full-batch gradient steps with step size 0.25, using the entire weighted training distribution. The loss is evaluated on the raw prediction before clipping.

Let $\mathbf{p}_w(u)$ give the population frequencies of observed (z, h, y) records, and let $\mathbf{m}$ give the frequencies retained from earlier observations. The training distribution and the updated retained distribution are

$$\widetilde{\mathbf{m}} = 0.7\mathbf{m} + 0.3\mathbf{p}_w(u), \qquad \mathbf{m}^+ = 0.4\mathbf{m} + 0.6\mathbf{p}_w(u)$$

The next network is trained on records weighted by $\widetilde{\mathbf{m}}$. The training and retention weights are given, not chosen by the incentive policy. The training loss uses observed records, not productive inputs, manipulation, or private costs.

To construct initial and historical models, let s be an overall scale and γ_1, γ_2 weights used for calibration, distinct from the network's output weights. Using plans with unmanipulated reports, 400 gradient steps fit the target

$$s\left[y + \frac{(\gamma_1 - 1)x_1 + (\gamma_2 - 1)x_2}{2}\right]$$

The common initial network uses $s = 0.45, (\gamma_1, \gamma_2) = (0.4, 1.6)$. Historical records are generated without incentives under networks calibrated with $s = 0.85, (\gamma_1, \gamma_2) = (1, 1)$ and $s = 0.7, (\gamma_1, \gamma_2) = (1.8, 0.2)$. Each system then receives the same current network while retaining its own historical records. Unmanipulated reports are used only to construct these controlled initial conditions, not as an additional operating signal for the incentive policy.

C.3 Oracle Benchmarks

There are six reviews, each covering six operating periods. Predictor and incentive stay constant within each interval; outcomes arrive before the next review. With $\delta = 0.95$, $\bar{\delta} = \delta^6$, and $b_\delta = \sum_{j=0}^{5} \delta^j$, total value is

$$\sum_{t=0}^{5} \bar{\delta}^t b_\delta \, r(w_t, u_t)$$

For gains averaged across histories, the denominator is the mean of the two complete no-intervention values, 62.21. Each policy generates its own predictors, retained records, reports, and outputs. History 2 is the benchmark history and history 1 is the alternative history. Incentive points in the path figure are choices at review dates. The connecting lines are visual guides; incentives remain fixed between reviews. Cumulative payoff includes every operating period.

Let $Q_i(u)$ be the optimal total value in history i conditional on first incentive u. The history-aware oracle maximizes each Q_i separately. The current-information oracle maximizes $\{Q_1(u) + Q_2(u)\}/2$ over a common first incentive; subsequent choices may differ once current predictors distinguish the histories. These objectives define the oracle benchmarks, which are approximated numerically without a guarantee of global optimality.

We optimize over $u \in [0, 0.6]$ using L-BFGS-B, with multiple starting points and numerical derivatives. The myopic search evaluates a 301-point grid, refines each local peak, and includes both endpoints. Dynamic searches start from myopic, zero-incentive, constant-incentive, and previously found feasible paths. Each dynamic search can also use improved continuation paths found by the other, so the current-information benchmark has access to continuations found in the history-aware search. Final values are recomputed from complete trajectories.

C.4 When Is Dynamic Incentive Design Worthwhile?

The sensitivity analysis uses incentive-cost coefficients $k \in \{0.1625, 0.65, 2.6\}$. Figure 3 shows replay weights $\omega \in \{0.2, 0.4, 0.6\}$, and Table 2 also includes $\omega \in \{0, 0.35, 0.7\}$. Two further cases use strategic-worker fractions $\alpha = 0.3$ and 0.6, with $k = 0.65$ and $\omega = 0.7$, and generate historical records under the corresponding population composition. Every case is optimized separately. Each grid and optimization budget are fixed before evaluation, and all cases are reported.

Within each case, the two histories share the same initial predictor and current payoff function. Each policy then generates its own behavior. Let V^{M}, V^{C}, and V^{H} be the equally weighted average values for myopic, current-information dynamic, and history-aware control. The current-information policy uses a common initial incentive chosen on this same average objective. Figure 3 reports $V^{\mathrm{H}} - V^{\mathrm{M}}$ and $V^{\mathrm{H}} - V^{\mathrm{C}}$, each divided by no-intervention value 62.21.

Interpreting the two gains. The left panel measures the gain from accounting for future learning and history relative to choosing myopically at every review. The right asks how much observing history adds when the firm already accounts for future learning. Their difference is $V^{\mathrm{C}} - V^{\mathrm{M}}$, so the right-panel gain is part of the left-panel gain, not an additional amount to be added. A large benefit from dynamic incentives need not imply a large additional benefit from history. A common initial incentive can also be close to one history's optimum while sacrificing payoff in the other; averaging includes both consequences. Later predictors can distinguish the histories, so the information restriction need not persist throughout the horizon.

Why replay affects the gains differently. At a fixed current network, training uses $\omega\mathbf{m} + (1 - \omega)\mathbf{p}_w(u)$. A larger ω puts less weight on new observations affected by today's incentive, which can reduce its influence on the next predictor. At the same time, greater weight on retained observations can make incentive choice depend more on history. These effects help explain why the left-panel gain falls and the right-panel gain rises in the displayed cases, but they do not establish general monotonicity. Today's records also enter $\mathbf{m}^+ = 0.4\mathbf{m}+0.6\mathbf{p}_w(u)$, so greater replay can increase their importance in later updates. The overall payoff effect depends on subsequent responses, incentives, and the remaining horizon.

Why incentive cost matters. History is valuable when using it changes incentives in a way that improves net payoff, not simply because the retained data differ. At $\omega = 0.6$, raising k from 0.65 to 2.6 reduces the gain over myopic control from 2.98 to 1.23 percentage points and the additional gain from history from 1.70 to 0.15 percentage points. This does not mean that retained data no longer affect training or that the policies are identical. It means that observing history adds less net payoff in this case.

Changing the fraction of strategic workers changes both current behavior and later training. The overall gain from incentives can therefore change differently from the additional gain from observing history. Table 2 reports both comparisons rather than interpreting a larger overall gain as evidence of greater history value.

C.5 Policy Learning with FQI

Policy learning uses 3,200 training trajectories in 1,600 pairs, each containing one organization from each history. At the first review, each pair receives the same incentive, drawn uniformly from $\{0, 0.02, \ldots, 0.6\}$. Later incentives are drawn independently across organizations: a uniform draw, repeating the previous incentive, and moving one grid point down or up have probabilities 0.2, 0.4, 0.2, and 0.2, respectively. Choices stay within the endpoints, and assignment probabilities are recorded. A separate 400 diagnostic trajectories come from 200 disjoint pairs.

The current-information policy uses 51 features: 36 network parameters, predictions at 12 fixed probe inputs, and the previous incentive, observed payoff, and output. The history-aware policy adds 49 features computed from retained records: the 36 components of the regularized prediction-loss gradient at the current parameters, and 13 weighted averages covering each observed feature, its square, the individual outcome, and each feature–outcome product.

Table 2: Payoff differences across incentive costs, replay weights, and strategic-worker fractions, averaged equally over both histories. H, C, and M denote history-aware, current-information, and myopic benchmarks; zero denotes no intervention

k	ω	α	H minus zero	H minus M	H minus C
0.1625	0	0.9	16.57	14.57	0.00
0.1625	0.2	0.9	8.81	6.69	0.90
0.1625	0.35	0.9	7.38	5.02	1.11
0.1625	0.4	0.9	7.05	4.95	1.28
0.1625	0.6	0.9	5.41	3.81	1.52
0.1625	0.7	0.9	4.99	3.20	1.83
0.65	0	0.9	14.98	13.06	0.00
0.65	0.2	0.9	7.36	5.60	1.02
0.65	0.35	0.9	6.31	4.30	1.13
0.65	0.4	0.9	5.99	4.59	1.57
0.65	0.6	0.9	4.36	2.98	1.70
0.65	0.7	0.9	4.17	2.61	1.76
2.6	0	0.9	8.96	7.33	0.00
2.6	0.2	0.9	4.77	3.80	0.01
2.6	0.35	0.9	3.42	2.58	0.04
2.6	0.4	0.9	2.93	2.14	0.09
2.6	0.6	0.9	1.85	1.23	0.15
2.6	0.7	0.9	1.63	0.78	0.15
0.65	0.7	0.3	11.30	2.50	0.38
0.65	0.7	0.6	9.67	3.44	0.79

Both policies use information available before choosing the incentive. The history-aware policy requires access to the network, its training loss, and retained records and weights. Initially, both histories have identical current-information features, with previous-period quantities set to zero, but different history summaries. Later predictors can distinguish the histories. Fitting doesn't use the simulator's private parameters.

Finite-horizon FQI works backward from the last review, fitting an ExtraTrees regressor at each review with 256 trees and minimum leaf size one. The terminal target contains only current payoff. Learned myopic control instead pools observations across reviews to predict current payoff, using the same current-information features and tree settings. All three policies use the same 19,200 training transitions. Regression settings were fixed after preliminary comparisons and before collecting the balanced training sample; diagnostic trajectories are not used for policy selection.

Averaged across both histories, history-aware FQI gains 4.14%, current-information FQI gains 2.20%, and learned myopic control gains 1.60%. The corresponding dynamic oracle gains are 4.17% and 2.42%. Rounding the history-aware oracle path to the feasible incentive menu gives 4.15%, while exact myopic optimization on the grid gives 1.57%. Both learned dynamic policies have lower average payoff than their corresponding oracle. Their difference includes approximation and estimation error, not only the value of retained information. These comparisons concern the specified population and do not establish transfer to new setting.

C.6 FQI Training and Evaluation

FQI estimates the current and future payoff from each incentive. Let $t = 0, \ldots, 5$ index reviews, o_t denote the policy inputs described above, $R_t = b_\delta r_t$ be the discounted payoff within a six-period review interval, and $\bar{\delta} = \delta^6$ be the discount factor between reviews. Write $\widehat{Q}_t(o, u)$ for estimated remaining payoff from choosing u at review t given information o. Starting with $\widehat{Q}_6 = 0$, we work backward through the reviews and construct the regression target

$$q_t^{\text{target}} = R_t + \bar{\delta} \max_{u' \in \{0, 0.02, \ldots, 0.6\}} \widehat{Q}_{t+1}(o_{t+1}, u') \tag{92}$$

At each review, we regress this target on (o_t, u_t) using only the training trajectories. The target combines the recorded payoff with the estimated value of the best next incentive; at the last review, it

contains only current payoff. The fitted policy chooses the incentive with the highest estimated value, $\widehat{\pi}_t(o) \in \arg\max_{u \in \{0, 0.02, \ldots, 0.6\}} \widehat{Q}_t(o, u)$, breaking ties in favor of the smallest incentive. Myopic policy learning omits the future-value term.

During evaluation, the prediction network is retrained after each review using new and retained observations. The incentive policy is not retrained: it applies the previously fitted decision rules to the available information at each review.

C.7 Training Samples

We repeat randomized data collection and policy fitting five times. Each replication supplies a fresh set of 3,200 training trajectories to all three policies under the same paired design. The training-data seeds are 952501–952505, paired with fitting seeds 8501–8505. These runs do not reuse the original training dataset and are not used to tune the estimators. Every fitted policy is evaluated by its complete expected-payoff trajectory in the benchmark history, using the same no-intervention denominator as Table 1.

The sample standard deviation measures variation across the five complete data-collection and fitting runs, and the standard error of their mean is that standard deviation divided by $\sqrt{5}$. The history-aware minus current-information difference is paired within each run because both policies use the same training data.

Table 3: Replications of policy learning in the benchmark history. The three policy columns report percentage gains over no intervention; H minus C is their paired percentage-point difference. SD is the sample standard deviation across runs; SE is the standard error of the five-run mean. The original run in Table 1 is not included in these summaries

Data seed	Fit seed	Myopic	Current info.	History aware	H minus C
952501	8501	1.38	3.21	6.47	3.26
952502	8502	1.49	3.16	6.45	3.29
952503	8503	1.21	3.12	6.38	3.26
952504	8504	1.33	3.18	6.45	3.27
952505	8505	1.16	3.13	6.40	3.27
Mean		1.31	3.16	6.43	3.27
SD		0.13	0.04	0.04	0.01
SE of mean		0.06	0.02	0.02	0.01