AI, Headquarters and Guijie: A Principal-Agent Analysis

Zehao Zhang

August 28, 2025

1 Model

Consider a principal P (headquarters) and an agent A (local manager, "guijie"). HQ deploys an AI system that both detects manipulation and provides guidance; the dashboard can be strategically gamed, and AI has dual-use effects.

Technology and KPIs with AI The campaign outcome is $y \in \{0, 1\}$ (success or failure). The agent allocates effort across two activities

$$e = (e_c, e_g) \in \mathbb{R}^2_+,$$

where e_c raises true commercial impact ("core") and e_g is gaming/manipulation that inflates the dashboard without creating real value.

AI has two design knobs chosen by HQ: guidance $g \geq 0$ and detection intensity $\ell \geq 0$. Guidance directly raises the productivity of core and (leakily) of gaming; detection directly raises the sensitivity of manipulation detection. Formally:

- Success arrives with probability $f(e_c; g)$ with $\partial f/\partial e_c > 0$, $\partial^2 f/\partial e_c^2 < 0$, and $\partial^2 f/(\partial e_c\, \partial g) > 0$ (AI-complementarity).
- The measured KPI is $m(e_c, e_g; g) = \delta(g)\, e_c + \psi(g)\, e_g$, where $\delta'(g) > 0$ and $\psi'(g) = \sigma\, \delta'(g)$ with leakage parameter $\sigma \in [0, 1]$ (dual-use guidance).
- Manipulation is detected with probability $d(e_g, \ell, g)$, increasing in e_g and ℓ, convex in e_g, and with detection synergy $\partial^2 d/(\partial e_g\, \partial g) > 0$.

The agent bears a separable convex cost $G(e) = G_c(e_c) + G_g(e_g)$ with $G'_i > 0$, $G''_i > 0$.

Dashboard design and timing HQ commits to a linear dashboard and AI design before effort choices. Let $b \geq 0$ be the incentive slope (bonus sensitivity), $\lambda \in [0, 1]$ the weight on true success versus the measured KPI, $g \geq 0$ the guidance intensity, and $\ell \geq 0$ the detection intensity, with convex costs $K(g)$ and $C(\ell)$.

Payment rule: the realized wage is

$$w = w_0 + b\big(\lambda\, y + (1-\lambda)\, m(e)\big) - \mathbb{1}\{\text{detected}\}\, P,$$

where $P > 0$ is a clawback/penalty applied if gaming is detected. In expectations,

$$\mathbb{E}[w] = w_0 + b\Big(\lambda\, f(e_c; g) + (1-\lambda)\, (\delta(g)e_c + \psi(g)e_g)\Big) - P\, d(e_g, \ell, g).$$

Timing: (i) HQ chooses (b, λ, ℓ, g); (ii) the agent chooses e; (iii) y is realized, m is measured, detection occurs with probability $d(e_g, \ell, g)$, and payments are made.

Payoffs HQ's expected payoff is

$$v_P = r\, f(e_c; g) - \mathbb{E}[w] - C(\ell) - K(g),$$

where $r > 0$ is the revenue from a success. The agent's expected payoff is

$$v_A = \mathbb{E}[w] - G(e).$$

Strategic interaction For an interior agent optimum $e^*(b, \lambda, \ell, g)$, first-order conditions satisfy

$$\begin{aligned} G_c'(e_c) &= b\big(\lambda\, \partial f/\partial e_c(e_c; g) + (1-\lambda)\, \delta(g)\big), \\ G_g'(e_g) &= b(1-\lambda)\, \psi(g) - P\, \partial_{e_g} d(e_g, \ell, g). \end{aligned}$$

These equations highlight: (i) a direct AI effect via ℓ that raises detection and lowers e_g; (ii) an indirect AI effect via g that raises core productivity and dashboard sensitivity to core (through $\delta(g)$) but—with leakage σ—also raises returns to gaming, while detection synergy in g steepens $\partial_{e_g} d$.

2 Equilibrium and Comparative Statics with Dual-Use AI

Given (b, λ, ℓ, g), the agent chooses $e^*(b, \lambda, \ell, g)$ that solves the conditions above. Anticipating this, HQ chooses (b, λ, ℓ, g) to maximize v_P subject to incentive compatibility.

Proposition 1 (Dual-use AI: direct vs indirect effects)**.** *Suppose f is increasing and concave with positive AI complementarity, d is increasing in e_g and ℓ and convex in e_g with detection synergy in g, K and C are convex, and G is separable and convex. Then any interior equilibrium exhibits:*

1. *(Direct effect) e_g is decreasing in ℓ and P, with a corner $e_g^* = 0$ whenever $P\,\partial_{e_g} d \geq b(1-\lambda)\,\psi(g)$.*

2. *(Indirect guidance) e_c is increasing in g (AI complements core). With leakage $\sigma > 0$, e_g is locally increasing in g via $\psi'(g)$ but decreasing in g via detection synergy if $\partial^2 d/(\partial e_g\,\partial g)$ is large; thus e_g can be non-monotone in g.*

3. *(Design) If leakage is moderate (small σ) or detection synergy is strong, the optimal g^* crowds back from gaming: $\partial e_g^*/\partial g < 0$ at g^* while $\partial e_c^*/\partial g > 0$.*

4. *Optimal detection ℓ^* is (weakly) increasing in the gaming productivity slope $\psi'(g)$ and decreasing in synergy strength.*

Sketch. *Follows from the agent FOCs and the envelope theorem: g shifts both marginal benefits and the detection slope; signs depend on leakage σ and synergy $\partial^2 d/(\partial e_g\,\partial g)$.*

Example (linear-quadratic). Let $G_c(e_c) = \frac{1}{2}c_c e_c^2$, $G_g(e_g) = \frac{1}{2}c_g e_g^2$, $f(e_c; g) = (1+\eta g)e_c$, $\delta(g) = \delta_0 + \alpha g$, $\psi(g) = \psi_0 + \sigma\alpha g$, and $d(e_g, \ell, g) = \ell(1+\kappa g)e_g$. Then

$$e_c^* = \frac{b\big(\lambda(1+\eta g) + (1-\lambda)\,\delta(g)\big)}{c_c}, \quad e_g^* = \max\left\{0,\ \frac{b(1-\lambda)\,\psi(g) - P\,\ell(1+\kappa g)}{c_g}\right\}.$$

Hence $\partial e_g^*/\partial g < 0$ when $P\,\ell\,\kappa > b(1-\lambda)\,\sigma\alpha$ (guidance-induced detection dominates dual-use leakage), generating a U-shaped gaming response in g.