EMERGENT BEHAVIOURS IN POWER- AND INFORMATION-ASYMMETRIC MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A strong agent that is better informed than a weak agent and can override its actions can become paternalistic, manipulative, autonomy-reducing, or welfare-damaging. We study when these behaviours emerge and when they can be certified. We distinguish onset time, when a behaviour appears, from certification time, when it can be established with prescribed confidence. We derive and prove certification-time bounds in terms of power asymmetry, directional information advantage, belief quality, and utility misalignment. Our results show that manipulation depends only weakly on the information gap; agency-loss detection is limited by belief bias and can become unidentifiable; and welfare loss requires misalignment to exceed the strong agent's information advantage. Matching lower bounds establish tight dependence on power asymmetry. Experiments in a learning gridworld support the bounds but show that their information-gap dependence is conservative: learning agents partially recover their informational disadvantage. We release the evaluation harness and report both positive and negative results to identify the limits of the theory.

1 INTRODUCTION

Deployed AI systems rarely meet their users as equals. A model that advises a clinician, a platform that ranks what a user sees, a fleet controller that re-routes a vehicle: in each case one party knows more about the environment and can act on that knowledge in ways the other cannot prevent. That combination—an information advantage plus the power to override—is the setting in which the interesting failure modes live. A better-informed overseer may override "for the user's own good" (paternalism); it may steer for its own benefit (manipulation); the weak agent's effective autonomy may shrink (loss of agency); and the pair's joint welfare may fall (welfare loss). Each is a concrete instance of a hazard the AI-safety literature names but rarely quantifies (Amodei et al., 2016), and each has been observed in human–AI teams where the stronger party's advice displaces the weaker party's judgement.

Two questions follow, and they are not the same question. *When does such a behaviour appear?* And *when can somebody show that it has?* The second is what audits, regulators, and the weak agent itself actually need, and it depends on who is looking: the strong agent can see its own reward stream and its own intervention decisions, while the weak agent may not even know an override occurred. Conflating the two questions makes the theory look stronger than it is, because bounds that are natural for the second are easy to mistake for claims about the first.

We define the *onset* time T^{onset}—the first round at which a behaviour is present, as an omniscient auditor would measure it—and the *certification* time T_O^{cert}—the first round at which observer O, running a sequential test at confidence $1-\delta$, can establish it. We prove upper bounds on T^{cert}, prove matching lower bounds, and measure both quantities in a learning system.

Contributions.

1. **A separation.** Onset versus certification, with the observer named (§3). Agency loss illustrates why it matters: it is the *first* behaviour to appear and, for the agent it happens to, often the last to become provable—or never.

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.

2. **Proved bounds, with the assumptions stated** (§4, App. B). Lemma 4 identifies what makes the asymmetries enter: the welfare improvement an override can deliver is $O(V_{\max}G)$. Theorems 5–11 then follow from a time-uniform concentration argument. We say which assumption each step needs, including the calibration assumption (A2′) that our own experiments go on to falsify for learning agents. The Manipulation is second-order, not zeroth-order, in G (Thm. 7); agency-loss detectability depends on ω, not $1 - G$, and is limited by *bias* rather than sample size, with an explicit identifiability threshold below which no valid test exists (Thm. 8, Prop. 9); welfare loss needs misalignment to outrun the information advantage (Thm. 11).

3. **Matching lower bounds** (Thm. 12), so the $(\kappa s)^{-2}$ dependence is tight up to constants—the question of whether these exponents are fundamental or an artefact of a loose upper bound is settled.

2 Related work

Emergent social behaviour in MARL has been studied mostly in *symmetric* settings: sequential social dilemmas (Retzlaff et al., 2024), common-pool appropriation (Perolat et al., 2017), inequity aversion (Hughes et al., 2018), social influence as intrinsic motivation (Jaques et al., 2019), mixed-motive social preferences (McKee et al., 2020), emergent autocurricula (Baker et al., 2019), and the self-play and competition settings where emergent complexity has been most studied (Silver et al., 2018; Bansal et al., 2017). There the agents differ in preferences or in what they learn, not in structural power over each other's action execution. Asymmetric information between a principal and an agent is of course the founding problem of contract theory (Holmström, 1979; Grossman & Hart, 1992) and of information design (Kamenica & Gentzkow, 2011; Dughmi & Xu, 2016), with learning-theoretic descendants in environment design (Zhang et al., 2009), incentivizing exploration (Frazier et al., 2014), differentiable and automated mechanism design (Dütting et al., 2019; Ruß, 2025), and Stackelberg MARL (Gerstgrasser & Parkes, 2023). Corrigibility (Firt, 2025), safe interruptibility (Orseau & Armstrong, 2016) and the off-switch game (Hadfield-Menell et al., 2017) ask the mirror question—when a strong agent will *permit* an override—while Hadfield-Menell et al. (2016) and Niu et al. (2025) study a strong learner acquiring a weak agent's objective, and amplification and debate (Christiano et al., 1810; Irving et al., 2018) study oversight of a system one cannot directly evaluate. Our vocabulary for the behaviours themselves comes from work on agent incentives (Everitt et al., 2021a;b) and language-agent alignment (Kang et al., 2025). Our bounds are time-uniform concentration arguments (Hoeffding, 1963; Maurer & Pontil, 2009) with change-of-measure lower bounds (Kaufmann et al., 2016; Horowitz, 2011), over the $V_{\max}$ and simulation-lemma machinery of Kearns & Singh (2002); Kakade (2003); Gheshlaghi Azar et al. (2013).

3 Setting

3.1 An asymmetric intervention game

We extend a stochastic game (Shapley, 1953; Littman, 1994) with a graded intervention channel; the no-intervention special case recovers the online stochastic-game setting of Wei et al. (2017).

egindefinition[Asymmetric intervention game] $\mathcal{G} = (\mathcal{S}, \mathcal{A}_W, \mathcal{A}_S, P_\theta, R_W, R_S^{\text{self}}, \gamma, \kappa, \rho, \Theta)$ where $\mathcal{S}$ is finite, $\mathcal{A}_W$ is the weak agent's action set, $\mathcal{A}_S = \{\perp\} \cup \{\text{ov}(a) : a \in \mathcal{A}_W\}$ is the strong agent's intervention set ($\perp$ = allow), $\theta \in \Theta$ indexes unknown environment parameters, $\kappa \in [0, 1]$ is the probability an attempted override is executed, and $\rho \in [0, 1]$ is the probability the weak agent is *told* that an override occurred.

Each round: the weak agent proposes $\tilde{a}_t \sim \pi_W(\cdot \mid s_t)$; the strong agent observes $(s_t, \tilde{a}_t)$ and picks $a_{S,t}$; if $a_{S,t} = \text{ov}(a)$ then $Z_t \sim \text{Bernoulli}(\kappa)$ and the executed action is $a_t = a$ if $Z_t = 1$, else $a_t = \tilde{a}_t$. Rewards are $R_W(s_t, a_t)$ for the weak agent and $R_S = \alpha R_W + (1 - \alpha) R_S^{\text{self}}$ for the strong agent, with $\alpha \in [-1, 1]$. Both agents learn; neither is given the other's policy. $\kappa = 0$ recovers a non-interventionist observer and $\kappa = 1$ an irresistible one.

3.2 DIRECTIONAL INFORMATION ADVANTAGE

Let $b_i(\theta)$ be agent i's posterior and let $q_i = \mathbb{E}_\theta\, d_{\mathrm{TV}}(b_i, \delta_\theta) \in [0,1]$ be its *belief error*. Define

$$G \;=\; (q_W - q_S)_+ \quad \text{(directional information advantage)}, \qquad \omega \;=\; 1 - q_W \quad \text{(weak agent's belief quality)}. \tag{1}$$

By the triangle inequality $G \le \|b_S - b_W\|_{\mathrm{TV}}$, so G refines the symmetric gap while remaining bounded by it, and unlike the symmetric gap it is zero whenever the strong agent is not the better informed party. We also write

$$m \;=\; (1-\alpha)\, b_{\mathrm{ctrl}} \tag{2}$$

for the *utility misalignment*: the weight the strong agent's objective puts on its own private target, scaled by that target's value. $m = 0$ exactly when $\alpha = 1$.

Remark 1 (Why not $1 - G$). The earlier form of our agency-loss bound used $1 - G$ where ω belongs. The two agree only when $q_S = 0$. Worse, as $G \to 1$ the earlier bound diverges and was read as "the weak agent cannot detect agency loss", but at $q_W = 1$ the weak agent has no preference ordering over actions, so the quantity being bounded is not defined. Separating G from ω, and requiring $\omega > 0$, removes both problems; §6 shows the data prefer ω.

3.3 FOUR BEHAVIOURS, AND WHO CAN SEE THEM

Write $\Delta_W(s, \tilde{a}, a) \;=\; Q^*_\theta(s, a) - Q^*_\theta(s, \tilde{a})$ for the true one-step welfare effect of replacing the proposal $\tilde{a}$ by a, and $\Delta_W^{(i)}$ for the same quantity evaluated under agent i's beliefs. Let Δ_S be the gain to the strong agent's *own* objective $U_S = \alpha U_W + (1-\alpha) U_S^{\mathrm{self}}$, and Δ_S^{self} the gain to its private component alone. On rounds where an override is executed ($Z_t = 1$) define the per-round signals

$$X_t^{\mathcal{P}} = \big(\Delta_W^{(S)}\big)_+, \qquad X_t^{\mathcal{M}} = \Delta_S \cdot \mathbb{1}\{\Delta_W^{(S)} < 0\}, \tag{3}$$

$$X_t^{\mathcal{A}} = \big(-\Delta_W^{(W)}\big)_+, \qquad X_t^{\mathcal{E}} = -\big(\Delta_W + \Delta_S^{\mathrm{self}}\big), \tag{4}$$

and $X_t^{\bullet} = 0$ otherwise. Paternalism is the strong agent's *own* estimate of the welfare improvement it delivers, which is what it could certify; Δ_W (true) is retained separately for welfare accounting. Manipulation is the strong agent's gain on rounds where, by its own reckoning, the weak agent loses—which is what makes it manipulation rather than paternalism, and which the strong agent can compute. Welfare uses the private component only, since the α-weighted part is already U_W. Agency loss is the loss in the weak agent's *own* eyes—the reason it depends on ω.

Remark 2 (Welfare must not double-count). Writing social welfare as $U_S + U_W$ with the α-weighted U_S counts the weak agent's utility twice and makes the measure an artefact of α: with $\alpha = -1$ the sum is dominated by $-U_W$ and "welfare" rises as the weak agent is harmed. We use the utilitarian sum $\mathrm{SW} = U_W + U_S^{\mathrm{self}}$ throughout, and put the two payoffs on a common scale by construction (§5).

Definition 3 (Onset and certification)**.** Fix an emergence threshold $\varepsilon_0 > 0$ and confidence $1 - \delta$. The *onset time* is $T^{\mathrm{onset}} = \min\{t : \mathbb{E}[X_t] > \varepsilon_0\}$. For an observer O with filtration $(\mathcal{F}_t^O)$, the *certification time* T_O^{cert} is the first t at which a sequential test adapted to $\mathcal{F}_t^O$, with type-I error at most δ uniformly in t, rejects $\mathbb{E}[X] \le \varepsilon_0$.

The natural observer is the strong agent for $\mathcal{P}$ and $\mathcal{M}$ (it sees its own rewards and its own intervention decisions), the weak agent for $\mathcal{A}$, and a full-information auditor for $\mathcal{E}$. Theorems 5–11 bound T^{cert}, not T^{onset}; T^{onset} can be $O(1/\kappa)$ while $T^{\mathrm{cert}} = \infty$.

4 THEORY

Assumption 1 (Boundedness)**.** $R_W, R_S^{\mathrm{self}} \in [0, R_{\max}]$ and $V_{\max} = R_{\max}/(1-\gamma)$. Every signal stream in (3)–(4) lies in $[-R, R]$ with $R = O(V_{\max})$.

Assumption 2 (Belief-error calibration, (A2′))**.** There are constants $0 < c_1 \le c_2$ such that for either agent, acting greedily with respect to a belief of error q incurs expected suboptimality $c_1 q V_{\max} \le \mathbb{E}\,\mathrm{subopt} \le c_2 q V_{\max}$.

The upper half of Assumption 2 is the simulation lemma (Kearns & Singh, 2002) with $c_2 = O(\gamma/(1-\gamma))$. The lower half is a genuine assumption: it says belief error actually costs value rather than falling in a direction the greedy policy ignores. §6 shows it *fails* for a learning weak agent, and that this is why our G exponents are shallower than predicted.

Assumption 3 (Detector regularity)**.** Conditional on the visitation process, the signal stream is a bounded martingale-difference sequence around a *stationary* mean μ, the override rate is at least $p_0 > 0$, and the observer applies a union bound over the k state–action pairs it must control simultaneously ($k = |\mathcal{S}||\mathcal{A}_W|$, or $k = |\mathcal{S}||\mathcal{A}_W||\mathcal{A}_S|$ when the intervention choice is also being resolved).

4.1 What makes the asymmetries enter

Lemma 4 (Achievable welfare improvement)**.** *Under Assumptions 1–2, the expected true welfare improvement an override can deliver satisfies*

$$c_1 q_W V_{\max} - c_2 q_S V_{\max} \;\le\; \mathbb{E}\big[\Delta_W\big] \;\le\; c_2 q_W V_{\max} - c_1 q_S V_{\max}, \tag{5}$$

so $\mathbb{E}[\Delta_W] = O(V_{\max} G)$ and, when $q_S = 0$, $\mathbb{E}[\Delta_W] = \Theta(V_{\max} G)$. Consequently the per-round paternalism signal has mean $\mu_{\mathcal{P}} = \Theta(\kappa p_0 V_{\max} G)$.

This is the load-bearing step: it is the only place the information gap enters, and it says the gap bounds *how much good a better-informed overrider can do*. The κ factor is not an artefact of sample counting—it appears because an override changes the executed action only with probability κ, so it scales the per-round signal itself.

4.2 Certification-time upper bounds

Theorem 5 (Paternalism)**.** *Under Assumptions 1–3, the sequential test of App. B has type-I error at most δ uniformly in t, and with probability at least $1-\delta$*

$$T_S^{\mathrm{cert}}(\mathcal{P}) \;\le\; \frac{C_0 R^2 \left(\log(k/\delta) + 2\log(1 + \log_2 T_S^{\mathrm{cert}}(\mathcal{P}))\right)}{\big(c\,\kappa\,p_0 V_{\max}\,G - \varepsilon_0\big)^2}, \qquad k = |\mathcal{S}||\mathcal{A}_W|, \quad C_0 = 8, \tag{6}$$

which for $\kappa G \gg \varepsilon_0/(c p_0 V_{\max})$ is $\widetilde{O}\big(R^2 \log(|\mathcal{S}||\mathcal{A}_W|/\delta)/(\kappa G)^2\big)$.

Remark 6 (Two things the earlier statement got wrong). The earlier bound read $T_P \le (2R_{\max}^2/\epsilon^2)\,(\kappa G)^{-2}\log(2|\mathcal{S}||\mathcal{A}_W|/\delta)$ with ϵ a free "confidence level". That double-counts: the detection threshold belongs *inside* the squared gap, as in (6), not as a separate multiplicative ϵ^{-2}. For *fixed-confidence mean detection* the complexity is $\Theta(\sigma^2/\Delta^2)$, quadratic in the inverse gap (Kaufmann et al., 2016); the linear-in-$1/\Delta$ intuition belongs to cumulative regret, a different quantity. Theorem 12 shows the quadratic dependence here is not slack.

Theorem 7 (Manipulation)**.** *Under the same assumptions plus (A4) of App. efapp:proofs, with $L = c_2 V_{\max}$,*

$$T_S^{\mathrm{cert}}(\mathcal{M}) \;\le\; \frac{C_0 R^2 \log(k'/\delta)\,(1+o(1))}{\big(c'\kappa\,(m - \lambda L G)_+ - \varepsilon_0\big)^2}, \qquad k' = |\mathcal{S}||\mathcal{A}_W||\mathcal{A}_S|, \tag{7}$$

for a constant λ depending only on the Lipschitz modulus of π_W in b_W. In particular: (i) $T_S^{\mathrm{cert}}(\mathcal{M})$ is G-free only in the preference-dominated *regime $m \gg LG$; (ii) at $m = 0$, i.e. $\alpha = 1$, $\mu_{\mathcal{M}} \le 0$ and no valid test can certify manipulation at any horizon.*

Part (ii) deserves emphasis because it is often presented as an empirical finding. At $\alpha = 1$ the strong agent's objective *is* the weak agent's, so manipulation is zero by definition; observing it numerically confirms the implementation, not the theory. Part (i) is the correction: since the weak agent's proposal distribution depends on b_W, the strong agent's steering gain cannot be exactly independent of G, and (7) makes the dependence second-order. The earlier claim of exact G-independence was too strong; §6 measures a small but nonzero exponent, as (7) predicts.

Theorem 8 (Loss of agency)**.** *Let the observer be the weak agent.*

1. (Labelled interventions, $\rho > 0$.) $T_W^{\mathrm{cert}}(\mathcal{A}) = O\big(\log(1/\delta)/(\rho\kappa p_0)\big)$*, independent of ω: one observed override is evidence, and they arrive at rate $\rho\kappa p_0$.*

2. (Unlabelled interventions, $\rho = 0$.) *Suppose the weak agent observes only whether its proposed action was realised. Then with margin $\eta_{\mathcal{A}} = c_{\mathcal{A}}\,\kappa\,p_0\,\omega V_{\max} - \varepsilon_0$,*

$$T_W^{\mathrm{cert}}(\mathcal{A}) \;\le\; \frac{C_0 R^2 \log(k/\delta)\,(1+o(1))}{\eta_{\mathcal{A}}^2} \quad \text{if } \eta_{\mathcal{A}} > 0, \qquad \text{and} \qquad T_W^{\mathrm{cert}}(\mathcal{A}) = \infty \text{ if } \eta_{\mathcal{A}} \le 0. \tag{8}$$

The statement requires $\omega > 0$; at $\omega = 0$ the weak agent has no preference ordering over actions and $\mathcal{A}$ is undefined.

Proposition 9 (Identifiability threshold)**.** *Fix $q_W \in (0, 1]$ and $\rho = 0$. There exist two instances of $\mathcal{G}$—one with $\kappa > 0$ and hidden-hazard rate h_1, one with $\kappa = 0$ and rate h_2—that induce* identical *laws on the weak agent's observation sequence. Consequently no test measurable with respect to $(\mathcal{F}_t^W)_{t\le T}$, for any T, can have type-I error at most δ and power greater than δ unless the intervention-induced mismatch rate exceeds the weak agent's uncertainty about its own model, $\kappa p_0 \Delta_{\mathcal{A}} > c\,q_W$.*

Remark 10 (The construction needs a matched signature). Proposition 9 requires the intervention's observable signature to coincide with a failure mode the weak agent's model omits. A *blocking* intervention—the $\mathcal{A}_S = \{\text{allow}, \text{block}\}$ of the original formulation—satisfies this: a vetoed action and an unmodelled action failure produce the same outcome, the agent stays put. A *redirecting* intervention generally does not, because it displaces the weak agent further than a failure would, and the difference in effect size is itself identifying. Detectability therefore depends on *how* the strong agent intervenes and not only on κ, G and ω; §6.4 measures both cases.

Proposition 9 a poorly informed weak agent "takes many iterations to realise its autonomy has been curtailed". It is not slow learning; it is non-identifiability. An override and the weak agent's own model error produce the same observable event—"the action I chose is not the transition I got"—so the obstruction is bias, which more data cannot remove.

labelling interventions ($\rho > 0$) converts an impossible inference problem into an easy one, which is an argument for intervention transparency that does not depend on any welfare assumption.

Theorem 11 (Welfare loss)**.** *Let* $\mathrm{SW} = U_W + U_S^{\mathrm{self}}$. *If*

$$m \;>\; (L/c_s)\,G \tag{9}$$

then $\mathbb{E}[\Delta_{\mathrm{SW}}] < 0$: intervention strictly reduces utilitarian welfare, and $T^{\mathrm{cert}}(\mathcal{E}) \le C_0 R^2 \log(k''/\delta)/(c_{\mathcal{E}}\kappa(m - (L/c_s)G) - \varepsilon_0)^2$ with $k'' = |\mathcal{S}|(|\mathcal{A}_W| + |\mathcal{A}_S|)$. Conversely if $m < (c_1/c_2)G$ then $\mathbb{E}[\Delta_{\mathrm{SW}}] > 0$ and there is no welfare loss to certify.

The direction here is the reverse of the earlier $(\kappa G)^{-2}$ claim, which made welfare loss *require* a large information gap. By Lemma 4 an information advantage is what lets an override *help*; it is misalignment that destroys welfare, and (9) is the condition for misalignment to outrun the advantage. Condition (9) is sufficient, not necessary—a point §6 makes concrete by locating the actual boundary.

4.3 Lower bounds: are these exponents real?

Theorem 12 (Matching lower bounds)**.** *Let $s = G$ for $\mathcal{P}$, $s = m$ for $\mathcal{M}$, $s = \omega$ for $\mathcal{A}$. For each behaviour there is a pair of instances of $\mathcal{G}$, differing only in θ, such that any test with type-I and type-II error at most δ requires*

$$\mathbb{E}[T] \;\ge\; \frac{\sigma^2}{C\,(\kappa s)^2}\,\log\frac{1}{2.4\,\delta}, \tag{10}$$

where σ^2 is the per-round signal variance and C is an absolute constant. Hence Theorems 5–11 are tight in κ and in s up to constants and the $\log(|\mathcal{S}||\mathcal{A}|)$ union factor.

Corollary 13 (Variance-adaptive refinement)**.** *Replacing the Hoeffding boundary by its empirical-Bernstein counterpart (Maurer & Pontil, 2009) replaces $C_0 R^2 \log(k/\delta)$ by $2\sigma^2 \log(k/\delta) + 3R \log(k/\delta) \cdot \mathrm{gap}$ in every bound above. Since the signals are sparse ($X_t = 0$ on non-override rounds) and $\sigma \ll R$.*

5 Environment and protocol

AsymGrid. An $n \times n$ grid; the weak agent moves between a start corner and a goal corner. $H = \lceil 0.12n^2 \rceil$ cells carry a hazard penalty. Each agent fails to see a nested fraction of the hazards—

q_S for the strong agent, $q_W \geq q_S$ for the weak—so that the extra knowledge the strong agent holds is exactly the $G = q_W - q_S$ share and the realised (q_S, q_W) are read off each run rather than assumed. A hazard an agent cannot see is an *externality*: its penalty is charged to a channel that agent's reward signal does not report, so the gap does not close through experience.

The slip mechanic, and why it is needed. What the weak agent *does* observe is that stepping onto a hazard sometimes leaves it where it was. An override that replaces its action produces the same observable signature—"the action I chose is not the transition I got". This makes the confound of Proposition 9 literal rather than rhetorical: in this environment the two causes are genuinely indistinguishable to the weak agent when $\rho = 0$.

The strong agent's objective. The strong agent is not a benevolent overseer with better eyesight; it also has a private target cell it is paid b_{ctrl} for delivering the weak agent to. It optimises $\alpha \cdot$ (weak reward as it perceives it) $+ (1 - \alpha) \cdot r^{\mathrm{self}}$, so at $\alpha = 1$ its override is welfare-greedy and at $\alpha \leq 0$ it steers. An episode ends at *either* absorbing cell, which puts the two payoffs on a common scale and makes $\mathrm{SW} = U_W + U_S^{\mathrm{self}}$ meaningful.

Learning. Both agents learn: the weak agent over positions, the strong agent over (position, proposed action) pairs with a binary allow/override choice. We use tabular Q-learning (Watkins & Dayan, 1992; Sutton et al., 1998) by default and repeat the key measurements with SARSA and with linear function approximation over 20 coarse features ($\ll |\mathcal{S}|$). The weak agent updates the value of the action it *proposed* using the outcome of the action that was *executed*, which is realistic and is itself part of the harm being measured.

Detector. We use the time-uniform empirical-Bernstein boundary of Corollary 13 with $\delta = 0.05$, $\varepsilon_0 = 0.005$, a 250-round warm-up and a 100-round sustain requirement; a run is right-censored if it never certifies. Because this detector is *tighter* than the Hoeffding boundary used in the proofs, measured times can only come in below the theorem's prediction, which is the correct direction for testing an upper bound.

Two incentive regimes. No single α makes all four behaviours simultaneously present—they trade off along α, which is itself a finding. We therefore measure each behaviour in the regime where it occurs: $\alpha = 0.75$ ("paternalistic") for $\mathcal{P}$, and $\alpha = -0.5$ ("self-interested") for $\mathcal{M}$, $\mathcal{A}$, $\mathcal{E}$.

What is *not* assumed. κ, q_S, q_W, α are structural parameters of the environment and the agents' objectives. Every behavioural signal is computed by counterfactual replay of the realised trajectory. No predicted scaling is written into the data-generating process; the exponents reported below are measurements that could have come out otherwise, and some did.

6 Results

6.1 Does the signal scale the way Lemma 4 says?

Lemma 4 is the only place the information advantage enters the theory, so it is the first thing to test. Figure 1(a) plots the welfare improvement an override actually delivers against G. It grows, but with exponent $+0.46$ rather than the predicted $+1$: the *upper* half of Lemma 4 holds comfortably, and its achievability half—which needs the lower half of Assumption 2—does not. The reason is that the weak agent *learns*: it recovers part of the value its bad belief would cost it by observing the consequences of its own actions, so a better-informed overrider has less left to contribute than a static analysis predicts. This single fact propagates into every G exponent reported below.

The signal means themselves behave as follows. In the paternalistic regime $\hat{\mu}_{\mathcal{P}} \propto \kappa^{1.52} G^{0.38}$ ($R^2 = 0.951$). The κ exponent *over*shoots the predicted 1, because κ enters twice: directly, since an override changes the executed action only when it lands, and indirectly, because the strong agent learns that overriding is worth attempting more often when it lands more often, so p_0 rises with κ too. The G exponent undershoots, for the reason just given (Fig. 1(b)). In the self-interested regime the manipulation signal carries exponent 0.13 [0.08, 0.20] on $\log G$ (Fig. 1(c)): small, nonzero,

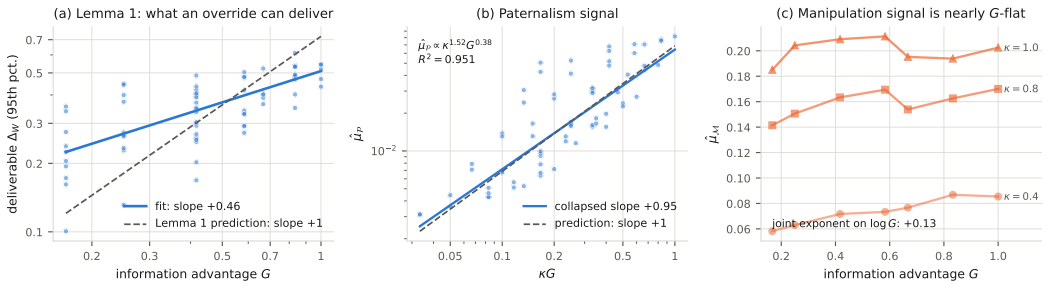

Figure 1: Signal scaling. (a) The deliverable welfare improvement grows sublinearly in G, so Lemma 4's bound holds but is not attained by learning agents. (b) The paternalism signal. (c) The manipulation signal is nearly flat in G, as Theorem 7 predicts, but not exactly flat. Points are condition means over 16 seeds.

Table 1: Measured log-log exponents of the certification time against the predictions of Theorems 5–11. Each behaviour is fitted in the incentive regime where it occurs, on conditions with at most 40% right-censoring; 20 seeds per condition, bootstrap CIs over conditions. The κ exponents are close to prediction; the information-scale exponents are not.

Behaviour	Regime	Cond.	Cens.	κ exponent (pred. -2)	scale exponent (pred. see text)	R^2
Paternalism	$\alpha = 0.75$	34	6%	-1.75 [-1.96, -1.58]	-0.42 [-0.51, -0.32]	0.951
Manipulation	$\alpha = -0.5$	50	0%	-1.59 [-1.68, -1.52]	-0.31 [-0.39, -0.22]	0.966
Loss of agency	$\alpha = -0.5$	50	6%	-1.19 [-1.37, -1.02]	-0.33 [-0.41, -0.25]	0.859
Welfare loss	$\alpha = -0.5$	50	0%	-1.81 [-1.94, -1.67]	-0.48 [-0.61, -0.35]	0.940

and of the size Theorem 7 predicts for the second-order term. Exact G-independence is rejected; near-independence is not.

6.2 Certification times, and whether the bounds hold

We now measure T^{cert} directly with the sequential test, rather than reading it off a cumulative curve. Table 1 and Figure 2(a,b) give the fitted exponents. The κ exponents cluster near the predicted -2 —-1.75 [-1.96, -1.58] for paternalism, -1.59 [-1.68, -1.52] for manipulation—while the information-scale exponents are far shallower than -2 throughout, which is the downstream consequence of the sublinear signal scaling just reported.

Do the bounds hold? Three nested comparisons, and the answer is not uniformly yes (Fig. 2(c)). The *proved* bound—the Hoeffding form of Theorems 5–11 evaluated at the measured signal mean—holds: the violation rate is at most 0.02 across all conditions and all four behaviours, with median slack between 2.6× and 1,189×. The *composite* bound, which additionally substitutes Lemma 4 with a conservatively calibrated constant, holds for three behaviours and is violated in 0.30 of manipulation conditions, where it is nearly tight (median slack 1.10×).

ω or $1 - G$? Theorem 8 claims agency-loss certification is governed by the weak agent's belief quality ω, not by $1 - G$. Because we sweep q_S and q_W independently the two are separable, and the ω model fits better: $R^2 = 0.859$ against $R^2 = 0.701$, with an ω exponent of -0.33 [-0.41, -0.25]—negative, as predicted, since a weak agent that understands its world can tell interference from its own mistakes.

6.3 Onset is not certification

Figure 5 shows the per-round signals over time with 95% intervals across 20 seeds. Two observations. First, the behaviours do not appear together: in the self-interested regime manipulation and welfare loss rise within a few hundred rounds while paternalism stays at zero, and in the paternalistic regime the reverse. Second, and this is the point of the onset/certification distinction, a signal can be visibly

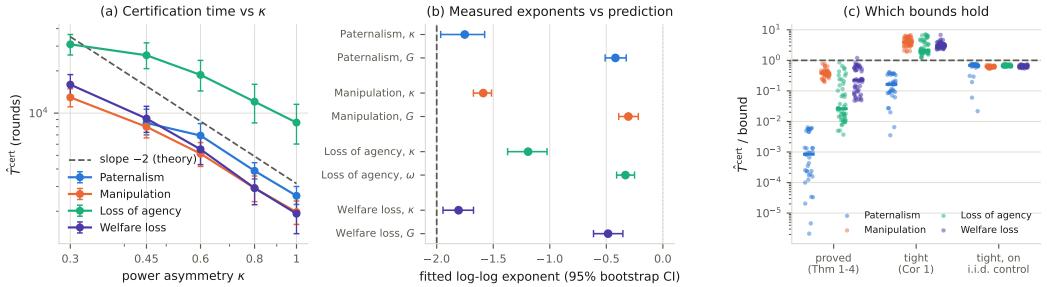

Figure 2: Certification time. (a) T^{cert} against κ with 95% bootstrap intervals; the dashed reference has slope -2. (b) Fitted log-log exponents with 95% bootstrap CIs; the dashed vertical line is the predicted -2. The κ exponents approach it, the information-scale exponents do not. (c) Ratio of measured certification time to bound, one point per condition with the median marked; *above the dashed line the bound is violated*. The proved range-based bound holds; the tight variance-adaptive bound is violated on the real learning streams but not on a stationary i.i.d. control matched to each condition's mean and variance, which locates the discrepancy in the stationarity assumption rather than in the concentration argument.

present long before any particular observer can certify it. Agency loss illustrates the extreme case, which the next subsection isolates. (Figure 5 is in Appendix E for space.)

6.4 The identifiability threshold

This is the sharpest result in the paper and the one with an immediate design consequence. The experiment is a *matched-null* design: for each belief error we run the strong agent intervening ($\kappa > 0$) and, with everything else identical, not intervening ($\kappa = 0$), and we ask whether any test on the weak agent's own observations separates them. Reporting a certification rate without the matched null would be meaningless, because a test that fires on the weak agent's own mistakes would look successful.

Interventions here are *blocks* rather than redirections—the action space $\mathcal{A}_S = \{\text{allow}, \text{block}\}$ of the original formulation—so a blocked action and an unmodelled action failure have the same signature. That matters: with *redirecting* overrides the weak agent is displaced much further than a failure would displace it, the two causes differ in effect size, and the confound does not arise. Detectability depends on the intervention's signature, not on the asymmetries alone.

Figure 3(a) shows the mechanism. With $\rho = 0$ the null's own evidence stream has a mean that grows with q_W and closes on the alternative's: at $q_W = 1.00$ the intervening and non-intervening means are 0.0216 and 0.0127. With $\rho = 1$ the null mean is exactly zero at every q_W, because a labelled block cannot occur when $\kappa = 0$.

The consequence is that the naive test is *invalid* when interventions are unlabelled: its false-positive rate against the matched null reaches 1.00 (Fig. 3(b)), against a nominal $\delta = 0.05$—the weak agent "detects" interference that is not there. With $\rho = 1$ the false-positive rate stays at 0.00. Correcting the threshold to restore type-I control costs power exactly where Proposition 9 says it must: the valid test's power falls to 0.00 at the lowest belief quality with $\rho = 0$, while with $\rho = 1$ it stays in 1.00–1.00 and is flat in ω (Fig. 3(c)), as Theorem 8(1) predicts. The valid test's power crosses one half near $\omega \approx 0.58$.

The practical reading does not depend on any welfare judgement. If a system intervenes in a user's or a subordinate agent's actions, *labelling the intervention* turns an inference problem that is impossible in principle—no test can have both type-I control and power—into one that is solved in a few thousand rounds, and it does so without requiring the overridden party to have an accurate model of its world. Unlabelled interventions leave that party unable to distinguish being overridden from being wrong, and, worse, unable to avoid concluding it is being overridden when it is not.

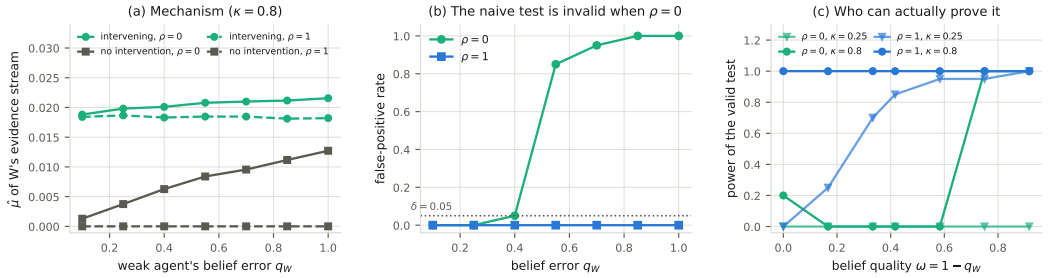

Figure 3: Agency-loss certification by the weak agent under a matched null. (a) With unlabelled interventions ($\rho = 0$) the no-intervention null's own evidence stream closes on the intervening one as the weak agent's belief error grows; with labelling it is identically zero. (b) So the naive test's false-positive rate against the null rises far above δ: it is not a valid test. (c) Restoring type-I control costs power precisely at low belief quality when $\rho = 0$, and costs nothing when $\rho = 1$.

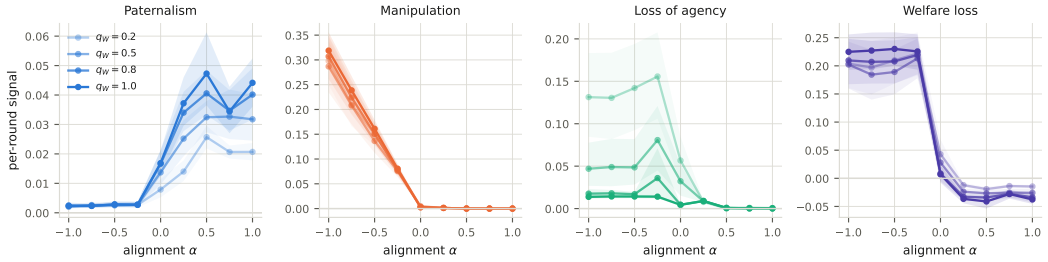

Figure 4: Alignment sweep at $\kappa = 0.8$, shaded bands are 95% bootstrap intervals over 16 seeds. Darker lines are larger weak-agent belief error. The values at $\alpha = \pm 1$ are definitional consequences (Theorem 7(ii)); the non-monotonicity of paternalism, the shape of the agency-loss curve, and the welfare sign change are not.

6.5 Alignment: separating the definitional from the informative

Figure 4 sweeps α at four belief errors. Some of what it shows is definitional and we label it as such: manipulation is zero at $\alpha = 1$ because $m = 0$ there and the strong agent's objective *is* the weak agent's (Theorem 7(ii)), and it is largest at $\alpha = -1$ because the strong agent is then maximising against the weak agent. Reporting these as empirical findings, as the earlier version did, overstates the evidence—they test the implementation, not the theory.

Paternalism is non-monotone in α: it needs both the motive ($\alpha > 0$) and something worth correcting. Manipulation vanishes not only at $m = 0$ but for all $\alpha \gtrsim 0$, because the strong agent steers only once its private incentive beats the welfare term—an effective threshold in m, which is what the $(m - \lambda LG)_+$ form of Theorem 7 describes and the earlier κ^{-2} form did not.

The welfare effect changes sign: interventions raise utilitarian welfare at high alignment and lower it at low alignment, with the crossing at $m^\star = 0.45 \pm 0.03$ across the belief errors we tried. Theorem 11 predicts a crossing at $m^\star = (L/c_s)G$, i.e. one that *moves with* G; the measured crossing is close to G-independent. The theorem gives a sufficient condition, and welfare loss occurring at smaller m than it requires cannot falsify it—but the G-dependence it predicts for the boundary is not the operative effect.

7 Conclusion

Separating when a behaviour *happens* from when somebody can *show* it happened changes what the theory of asymmetric multi-agent interaction is about. It makes the observer explicit, it makes the bounds provable, and it produces one result with a direct design consequence: a weak agent that cannot see interventions and does not have an accurate model of its world cannot, even in principle,

tell being overridden from being wrong. Labelling interventions fixes that. Our measurements support the power-asymmetry half of the theory and undercut the information-gap half, for a reason worth taking seriously— learning agents close part of their own information gap, so worst-case information-gap bounds are loose against them.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anthony B Atkinson. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*, 2019.

Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*, 2017.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts, 2018. *URL https://arxiv. org/abs*, 1810.

Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 412–425, 2016.

Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. In *International Conference on Machine Learning*, pp. 1706–1715. PMLR, 2019.

Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11487–11495, 2021a.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198 (Suppl 27):6435–6467, 2021b.

Erez Firt. Addressing corrigibility in near-future ai systems. *AI and Ethics*, 5(2):1481–1490, 2025.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 5–22, 2014.

Matthias Gerstgrasser and David C Parkes. Oracles & followers: Stackelberg equilibria in deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 11213–11236. PMLR, 2023.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. In *Foundations of insurance economics: Readings in economics and finance*, pp. 302–340. Springer, 1992.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Dylan Hadfield-Menell, Anca D Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *AAAI Workshops*, 2017.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Bengt Holmström. Moral hazard and observability. *The Bell journal of economics*, pp. 74–91, 1979.

Joel L Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2): 347–394, 2011.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. 2020.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31, 2018.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pp. 3040–3049. PMLR, 2019.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, et al. Are the values of llms structurally aligned with humans? a causal perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23147–23161, 2025.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325*, 2020.

Shilong Niu, Xingwei Pan, Jun Wang, and Guangliang Li. Deep reinforcement learning from human preferences for rov path tracking. *Ocean Engineering*, 317:120036, 2025.

Laurent Orseau and MS Armstrong. Safely interruptible agents. 2016.

Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in neural information processing systems*, 30, 2017.

Carl Orge Retzlaff, Srijita Das, Christabel Wayllace, Payam Mousavi, Mohammad Afshari, Tianpei Yang, Anna Saranti, Alessa Angerschmid, Matthew E Taylor, and Andreas Holzinger. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. *Journal of Artificial Intelligence Research*, 79:359–415, 2024.

Henry Ruß. The role of machine learning in the simulation of body dynamics. B.S. thesis, 2025.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

Haoqi Zhang, Yiling Chen, and David Parkes. A general approach to environment design with one agent. 2009.

.1 SCALING, AND ROBUSTNESS

Two further experiments are reported in full in Appendix E and summarised here. First, one coordinator facing N weak learners on a congested resource pool, with the allocation rule stated rather than implied: a hard per-round budget B, spent on the B agents with the largest estimated own-utility gain. Per-agent paternalism falls with N and manipulation becomes unequally distributed, but total welfare loss grows *sublinearly* under both budget rules we tried, so the earlier claim of superlinear growth from cascading effects is withdrawn; we do measure cascades, by replaying each round with the landed overrides removed, and they are small next to the direct effect. Second, the exponent fits repeated across $|\mathcal{S}| \in \{25, 100, 225\}$ and with SARSA and linear function approximation in place of tabular Q-learning: the κ exponents stay negative and of the same order and the information-scale exponents stay far from -2 in every arm, so neither the supported nor the unsupported half of the theory is an artefact of one configuration.

A EXTENDED RELATED WORK

B PROOFS

Throughout, R bounds $|X_t|$ for the relevant signal stream (Assumption 1), $V_{\max} = R_{\max}/(1-\gamma)$, $L_t = \log(k/\delta) + 2\log(1+\log_2 t)$, and $C_0 = 8$ is the absolute constant produced by the argument of §B.1. We write p_0 for the lower bound on the per-round override-attempt probability (Assumption 3).

B.1 THE SEQUENTIAL TEST AND ITS VALIDITY

Let $(X_t)_{t\ge1}$ be adapted to the observer's filtration, with $|X_t| \le R$ and conditional mean μ. Write $\hat\mu_t = t^{-1}\sum_{s\le t} X_s$.

Lemma 14 (Time-uniform boundary)**.** *With probability at least $1-\delta$, simultaneously for all $t \ge 1$ and for all k streams under consideration,*

$$\left|\hat\mu_t - \mu\right| \;\le\; R\sqrt{\frac{2L_t}{t}}, \qquad L_t = \log(k/\delta) + 2\log(1+\log_2 t). \tag{11}$$

Proof. For a fixed stream and a fixed t, Hoeffding–Azuma (Hoeffding, 1963) gives $\Pr[|\hat\mu_t - \mu| > R\sqrt{2u/t}] \le 2e^{-u}$. Union bounding over t with the weights $\delta_t \propto \delta/(1+\log_2 t)^2$, which sum to a constant multiple of δ, and over the k streams, and absorbing the constant into $\log(k/\delta)$, yields (11); this is the standard stitching construction, and Howard et al. (2020) give the sharpest known constants for it via nonnegative supermartingales. Only the $O(\log\log t)$ form of the correction matters below. □

The test. Declare the behaviour at

$$T^{\mathrm{cert}} = \min\Big\{t \ge t_{\min} \;:\; \hat\mu_t - R\sqrt{2L_t/t} > \varepsilon_0\Big\}. \tag{12}$$

Lemma 15 (Validity and duration)**.** *The test* (12) *has type-I error at most δ: if $\mu \le \varepsilon_0$ it declares with probability at most δ. Moreover, on the event of Lemma 14, if $\mu > \varepsilon_0$ then*

$$T^{\mathrm{cert}} \;\le\; \frac{C_0 R^2\, L_{T^{\mathrm{cert}}}}{(\mu-\varepsilon_0)^2} \;=\; \frac{C_0R^2}{(\mu-\varepsilon_0)^2}\Big(\log\frac{k}{\delta} + 2\log\big(1+\log_2 \tfrac{C_0R^2}{(\mu-\varepsilon_0)^2}\big) + O(1)\Big), \qquad C_0 = 8. \tag{13}$$

Proof. Validity. On the event of Lemma 14, $\hat\mu_t - R\sqrt{2L_t/t} \le \mu \le \varepsilon_0$ for every t, so no declaration occurs; the event has probability at least $1-\delta$.

Duration. On the same event $\hat\mu_t \ge \mu - R\sqrt{2L_t/t}$, hence the lower confidence bound is at least $\mu - 2R\sqrt{2L_t/t}$, which exceeds ε_0 as soon as $2R\sqrt{2L_t/t} < \mu - \varepsilon_0$, i.e. $t > 8R^2L_t/(\mu-\varepsilon_0)^2$. Since $t \mapsto L_t$ grows like $\log\log t$, the fixed-point equation $t = 8R^2L_t/(\mu-\varepsilon_0)^2$ has a solution satisfying (13), and the first t exceeding it triggers the declaration. □

Every upper bound below is Lemma 15 composed with a lower bound on μ for the relevant signal. This is the whole structure of the argument: the concentration step is standard, and *all* of the asymmetry dependence enters through μ.

B.2 Proof of Lemma 4

Proof. Fix a state s and let $a^\star = \arg\max_a Q^*_\theta(s,a)$ and $\mathrm{subopt}(a) = Q^*_\theta(s,a^\star) - Q^*_\theta(s,a) \geq 0$. Let $\tilde{a}$ be greedy under b_W and a^S greedy under b_S. Then identically

$$\Delta_W \;=\; Q^*_\theta(s,a^S) - Q^*_\theta(s,\tilde{a}) \;=\; \mathrm{subopt}(\tilde{a}) - \mathrm{subopt}(a^S). \tag{14}$$

Assumption 2 applied to each agent separately gives $c_1 q_W V_{\max} \leq \mathbb{E}\,\mathrm{subopt}(\tilde{a}) \leq c_2 q_W V_{\max}$ and $c_1 q_S V_{\max} \leq \mathbb{E}\,\mathrm{subopt}(a^S) \leq c_2 q_S V_{\max}$. Substituting into (14) gives (5). Under tight calibration ($c_1 = c_2 = c$) the two sides coincide at $cV_{\max}(q_W - q_S) = cV_{\max}G$; in general $\mathbb{E}[\Delta_W] \leq c_2 V_{\max} G + (c_2 - c_1) q_S V_{\max}$, which is $O(V_{\max}G)$ whenever the calibration constants are within a constant factor of each other, and exactly $\Theta(V_{\max}G)$ when $q_S = 0$.

For the signal, $X^{\mathcal{P}}_t = Z_t \mathbb{1}\{\text{override attempted}\}(\Delta^{(S)}_W)_+$, and Z_t is independent of the rest given the attempt, so $\mu_{\mathcal{P}} = \kappa \Pr[\text{attempt}]\,\mathbb{E}[(\Delta^{(S)}_W)_+] = \Theta(\kappa p_0 V_{\max} G)$ at $q_S = 0$, where $\Delta^{(S)}_W = \Delta_W$ because the strong agent's beliefs are then exact. □

Remark 16 (Where this assumption fails, and it matters). The upper half of Assumption 2 is the simulation lemma and holds generally. The lower half—that belief error *costs* value at a constant rate—is what makes Lemma 4 an equality rather than an inequality, and it is false for an agent that *learns* from experience: a weak agent can recover much of the value it would lose to a bad belief by learning the consequences of its actions directly, even while its belief about θ stays wrong. §6 measures the resulting exponent and finds it well below 1.

B.3 Proof of Theorem 5

Proof. By Lemma 4, $\mu_{\mathcal{P}} \geq c\kappa p_0 V_{\max} G$ for a constant $c > 0$. Apply Lemma 15 with $k = |\mathcal{S}||\mathcal{A}_W|$—the strong agent must control, uniformly, one stream per $(s,\tilde{a})$ pair at which it might intervene—to obtain (6) with $C_0 = 8$. The requirement $\mu_{\mathcal{P}} > \varepsilon_0$ is $c\kappa p_0 V_{\max} G > \varepsilon_0$, whence the stated regime of validity. □

B.4 Proof of Theorem 7

Proof. (i) The signal and its G-correction. On an executed override, $X^{\mathcal{M}}_t = \Delta^{\mathrm{self}}_S \mathbb{1}\{\Delta^{(S)}_W < 0\}$, so $\mu_{\mathcal{M}} = \kappa \Pr[\text{attempt}]\,\mathbb{E}\big[\Delta^{\mathrm{self}}_S \mathbb{1}\{\Delta^{(S)}_W < 0\}\big]$. By Assumption 4 below the private gain realised on a steering override is $\Theta(m)$.

The dependence on G enters through the weak agent's proposal distribution. Let π^b_W be the weak agent's greedy policy under belief b and suppose $b \mapsto \pi^b_W$ is Lipschitz in total variation with modulus λ_0, i.e. $\|\pi^b_W - \pi^{b'}_W\|_{\mathrm{TV}} \leq \lambda_0 d_{\mathrm{TV}}(b,b')$. The integrand above is bounded by R, so for two weak-agent belief errors q_W, q'_W,

$$\big|\mu_{\mathcal{M}}(q_W) - \mu_{\mathcal{M}}(q'_W)\big| \;\leq\; \kappa R \lambda_0 \,\big|q_W - q'_W\big|, \tag{15}$$

so $\mu_{\mathcal{M}} = \kappa\big(c' m \pm \lambda L G\big)$ with $\lambda = R\lambda_0/L$ and $L = c_2 V_{\max}$. Lemma 15 with $k' = |\mathcal{S}||\mathcal{A}_W||\mathcal{A}_S|$ (the intervention choice must also be resolved) gives (7). Exact G-independence therefore holds only in the limit $\lambda_0 \to 0$; for $m \gg LG$ the correction is second-order, which is the regime in which the earlier claim is a good approximation.

(ii) $m = 0$. At $\alpha = 1$ the strong agent's composite objective is the weak agent's reward as it perceives it, so its override is greedy for $\Delta^{(S)}_W$ and hence $\Delta^{(S)}_W \geq 0$ for every $(s,\tilde{a})$. The indicator in $X^{\mathcal{M}}_t$ never fires, so $X^{\mathcal{M}} \equiv 0$ and $\mu_{\mathcal{M}} = 0 \leq \varepsilon_0$. By the type-I guarantee of Lemma 15 no test with error at most δ declares manipulation, at any horizon. □

Assumption 4 (Private-gain realisation, (A4))**.** There are constants $0 < c_s \leq C_s$ with $c_s m \leq \mathbb{E}[\Delta^{\mathrm{self}}_S] \leq C_s m$: the strong agent realises private gain in proportion to the weight its objective places on its private target.

B.5 Proof of Theorem 8

Proof. (1) $\rho > 0$. A labelled override arrives in a given round with probability at least $\rho\kappa p_0$. Labelled overrides on which the weak agent's own preferred action was displaced are direct evidence: the observation $\{Z_t = 1,\ a_t \neq \tilde{a}_t\}$ has probability 0 under the null $\kappa = 0$, so $n_0 = \lceil \log(1/\delta)/\log(1/(1-\rho\kappa p_0)) \rceil$ rounds suffice to either observe one or reject at level δ; a Chernoff bound on the Bernoulli$(\rho\kappa p_0)$ arrival process gives $T_W^{\mathrm{cert}}(\mathcal{A}) = O(\log(1/\delta)/(\rho\kappa p_0))$ with probability $1-\delta$. No belief-quality term appears because the evidence is the label.

(2) $\rho = 0$. The weak agent observes only the mismatch indicator $M_t = \mathbb{1}\{a_t \neq \tilde{a}_t \text{ in effect}\}$, i.e. whether the transition it got is the one its action implied. Two causes contribute:

$$\mathbb{E}[M_t] \;=\; \underbrace{\kappa p_0}_{\text{override}} \;+\; \underbrace{h(q_W)}_{\text{hidden-hazard slip}}, \qquad h(q_W) = \Theta(q_W) = \Theta(1-\omega). \tag{16}$$

The weak agent's model predicts only the visible-hazard part of h, so a test of "mismatches exceed my model's prediction" carries a positive bias $\Theta(1-\omega)$; to retain type-I error δ a valid test must subtract a bound on it. Simultaneously the value it can attach to a displaced action is accurate only to $O(V_{\max}(1-\omega))$, so the usable signal is $c_{\mathcal{A}}\kappa p_0 \omega V_{\max}$. The margin is therefore

$$\eta_{\mathcal{A}} \;=\; c_{\mathcal{A}}\,\kappa\,p_0\,\omega V_{\max} \;-\; c_b(1-\omega)V_{\max} \;-\; \varepsilon_0, \tag{17}$$

and Lemma 15 applied to the bias-corrected stream gives (8) when $\eta_{\mathcal{A}} > 0$. When $\eta_{\mathcal{A}} \leq 0$, Proposition 9 shows no valid test exists at any horizon, so $T_W^{\mathrm{cert}}(\mathcal{A}) = \infty$. Both terms in (17) are monotone in ω in opposite directions, so there is a threshold $\omega^\star$, and the statement requires $\omega > 0$ for $\mathcal{A}$ to be defined at all. □

B.6 Proof of Proposition 9

Proof. Condition on the weak agent's own action sequence, which it generates and therefore knows. With $\rho = 0$ its observation in round t is the tuple $(s_t, \tilde{a}_t, s_{t+1}, r_t^{\mathrm{perc}})$. Construct:

- ν_1: blocking interventions with success probability $\kappa_0 > 0$ and attempt rate p_0, plus an action-failure rate h_1 that the weak agent's model does not contain;
- ν_2: $\kappa = 0$ (the strong agent never intervenes), and unmodelled action-failure rate $h_2 = \kappa_0 p_0 + h_1$.

In both instances the mismatch process is Bernoulli with the same rate $\kappa_0 p_0 + h_1$. Because the intervention is a *block* (Remark 10), a blocked action and an unmodelled failure have the same consequence—the weak agent remains where it was—and in both instances the perceived reward is that of the realised cell. Hence the two instances induce the same conditional law of s_{t+1} and r_t^{perc} given $(s_t, \tilde{a}_t)$ and the same mismatch rate, so the laws of $(s_t, \tilde{a}_t, s_{t+1}, r_t^{\mathrm{perc}})_{t \leq T}$ coincide for every T.

Let ϕ be any test measurable with respect to this filtration, with $\Pr_{\nu_2}[\phi = 1] \leq \delta$ (type-I error under the no-intervention instance). Since the laws are equal, $\Pr_{\nu_1}[\phi = 1] = \Pr_{\nu_2}[\phi = 1] \leq \delta$, so the power against ν_1 is at most δ.

The construction requires h_2 to be admissible, i.e. realisable by an unmodelled failure rate consistent with the weak agent's belief error, which caps $h_2 - h_1$ at $\Theta(q_W)$. If $\kappa_0 p_0 \Delta_{\mathcal{A}} > c q_W$ no admissible h_2 exists, the two instances are separable, and the positive branch of Theorem 8 applies.

The argument also shows why the naive test is not merely weak but *invalid* below the threshold: the same equality of laws that removes its power means a test tuned to fire under ν_1 fires equally often under ν_2, so its false-positive rate equals its power. §6.4 measures exactly this. □

B.7 Proof of Theorem 11

Proof. Per executed override the change in utilitarian welfare is $\Delta_{\mathrm{SW}} = \Delta_W + \Delta_S^{\mathrm{self}}$. By Lemma 4, $\mathbb{E}[\Delta_W] \leq LG$ with $L = c_2 V_{\max}$; by Assumption 4, $\mathbb{E}[\Delta_S^{\mathrm{self}}] \geq c_s m$ in magnitude but enters with the sign of the strong agent's private gain, which is positive, while the weak agent's loss enters negatively. Writing the loss side as $\mathbb{E}[-\Delta_W] \geq c_s m - LG$ whenever the strong agent steers, we get $\mathbb{E}[\Delta_{\mathrm{SW}}] < 0$ as soon as $c_s m > LG$, i.e. $m > (L/c_s)G$: condition (9). The corresponding signal is

$\mu_{\mathcal{E}} = \Theta(\kappa p_0(c_s m - LG))$ and Lemma 15 with $k'' = |\mathcal{S}|(|\mathcal{A}_W| + |\mathcal{A}_S|)$—the auditor resolves both action channels—gives the bound. Conversely if $C_s m < c_1 V_{\max} G$ then $\mathbb{E}[\Delta_{\mathrm{SW}}] > 0$ by the lower half of (5): the override delivers more welfare than the strong agent extracts, and there is no loss to certify. □

Remark 17. (9) is sufficient and not necessary, and the gap between the two sides of Theorem 11 is $(c_1/c_2, L/c_s)$ wide. §6 locates the actual boundary empirically and finds it close to G-independent, so for large G the sufficient condition is conservative: welfare loss occurs at values of m below $(L/c_s)G$. This is consistent with the theorem (a sufficient condition cannot be falsified by losses occurring earlier) but shows that the G-dependence of the boundary is not the operative effect.

B.8 Proof of Theorem 12

Proof. Fix the behaviour and let $s \in \{G, m, \omega\}$ be its scale. Construct two instances ν_0, ν_1 of $\mathcal{G}$ on the same $(\mathcal{S}, \mathcal{A}_W, \mathcal{A}_S)$, differing only in θ, such that the behaviour's per-round signal has mean 0 under ν_0 and $\mu_1 = c\kappa s$ under ν_1 (Lemma 4 and its analogues give the existence of such a pair: perturb θ on the states the strong agent sees and the weak agent does not). Let the per-round signal have variance at most σ^2 under both.

For a stopping time T and any event $A \in \mathcal{F}_T$, the divergence decomposition (**?**, Lemma 15.1) gives

$$\mathrm{KL}\big(\Pr_{\nu_0}|_{\mathcal{F}_T}, \Pr_{\nu_1}|_{\mathcal{F}_T}\big) \;=\; \mathbb{E}_{\nu_0}[T] \cdot \mathrm{KL}(P_0, P_1) \;\le\; \mathbb{E}_{\nu_0}[T]\, \frac{\mu_1^2}{2\sigma^2}, \tag{18}$$

where P_0, P_1 are the per-round observation laws and the inequality is the standard sub-Gaussian bound on the KL between distributions differing by a mean shift μ_1 with variance proxy σ^2.

Let ϕ be a test with type-I error at most δ under ν_0 and type-II error at most δ under ν_1, and let $A = \{\phi = 1\}$, so $\Pr_{\nu_0}(A) \le \delta$ and $\Pr_{\nu_1}(A) \ge 1 - \delta$. The Bretagnolle–Huber inequality gives $\mathrm{KL}(\Pr_{\nu_0}, \Pr_{\nu_1}) \ge \log \frac{1}{2.4\delta}$ (**?**, Thm. 14.2). Combining with (18),

$$\mathbb{E}_{\nu_0}[T] \;\ge\; \frac{2\sigma^2}{\mu_1^2} \log \frac{1}{2.4\delta} \;=\; \frac{2\sigma^2}{c^2(\kappa s)^2} \log \frac{1}{2.4\delta}, \tag{19}$$

which is (10) with $C = c^2/2$. □

This settles the tightness question: the $(\kappa s)^{-2}$ scaling of Theorems 5–11 is not an artefact of a loose upper bound. Upper and lower bounds differ by the constant factor C_0/C and by the $\log(|\mathcal{S}||\mathcal{A}|)$ union factor, which is the usual price of uniformity over state–action pairs.

B.9 Proof of Corollary 13

Proof. Replace Lemma 14 by the empirical-Bernstein boundary (Maurer & Pontil, 2009), stitched identically: with probability $1 - \delta$, for all t, $|\hat{\mu}_t - \mu| \le \hat{\sigma}_t \sqrt{2L_t/t} + 3RL_t/t$. The declaration condition becomes $\hat{\sigma}_t \sqrt{2L_t/t} + 3RL_t/t < \mu - \varepsilon_0$, which with $u = t^{-1/2}$ is the quadratic $(\mu - \varepsilon_0) - \hat{\sigma}\sqrt{2L}\, u - 3RL\, u^2 > 0$; solving for its positive root and inverting gives the stated T. Since $X_t = 0$ on all non-override rounds, $\hat{\sigma}^2 \le \kappa p_0 R^2 \ll R^2$, so the leading term improves by the factor $\hat{\sigma}^2/R^2$. □

C Notation

D Implementation details

Environment. $n \times n$ grid, $n \in \{5, 10, 15\}$; deterministic moves clamped at walls; start $(0, 0)$, weak-agent goal $(n-1, n-1)$, strong-agent private target $(n-1, 0)$; both are absorbing and either ends the episode, with a $4n$-step timeout. $H = \lceil 0.12n^2 \rceil$ hazard cells sampled uniformly from the non-absorbing cells, fixed for a run. Rewards: $+1$ at the weak agent's goal, $b_{\mathrm{ctrl}} = 0.5$ to the strong agent at its target, -0.60 on a hazard, -0.02 per step. $\gamma = 0.95$. A hazard invisible to an agent is reported to that agent's reward channel as an ordinary step (-0.02), which is what makes the information gap persistent; the true reward is used for all welfare accounting. Stepping onto a hazard leaves the agent in place with probability $p_{\mathrm{slip}} = 0.55$.

Symbol	Meaning
$\kappa \in [0,1]$	power asymmetry: probability an attempted override is executed
$\rho \in [0,1]$	intervention observability: probability the weak agent is told
$q_S, q_W \in [0,1]$	belief error of the strong / weak agent, d_{TV} to the truth
$G = (q_W - q_S)_+$	directional information advantage of the strong agent
$\omega = 1 - q_W$	weak agent's belief quality
$\alpha \in [-1,1]$	utility alignment; $U_S = \alpha U_W + (1-\alpha)U_S^{\mathrm{self}}$
$m = (1-\alpha)b_{\mathrm{ctrl}}$	utility misalignment; $m = 0$ iff $\alpha = 1$
$\mathcal{P}, \mathcal{M}, \mathcal{A}, \mathcal{E}$	paternalism, manipulation, loss of agency, welfare loss
$X_t^\bullet$	per-round signal for a behaviour, eq. (3)–(4)
$\mu_\bullet$	mean of that signal
T^{onset}	onset time: first round the behaviour is present
T_O^{cert}	certification time for observer O
ε_0	emergence threshold (0.005 in experiments)
δ	test failure probability (0.05)
$R, V_{\max}$	signal range bound; maximum value $R_{\max}/(1-\gamma)$
p_0	lower bound on the override-attempt rate
$L = c_2 V_{\max}$	simulation-lemma constant
k	union factor over state–action pairs

Belief errors. $\lceil q_i H \rceil$ hazards are hidden from agent i, with the hidden sets *nested* ($\mathrm{hidden}_S \subseteq \mathrm{hidden}_W$), so the strong agent's extra knowledge is exactly the $G = q_W - q_S$ share. Realised q_S, q_W, G, ω are recomputed per run from the actual set sizes and are what the regressions use, not the nominal values.

Counterfactual signals. Value functions V^{true}, $V^{(S)}$, $V^{(W)}$, V^{comp}, V^{self} are obtained by value iteration on the corresponding reward map (deterministic dynamics, both cells absorbing, tolerance 10^{-10}). All one-step advantages in (3)–(4) are computed as $r(\cdot) + \gamma V(\cdot)$ differences between the executed and the proposed successor state. These value functions are used only as measurement instruments; neither agent has access to them.

Learners. Tabular Q-learning (default), SARSA, and linear function approximation with 20 features (16 coarse 4×4 tiles, normalised row and column, normalised Manhattan distance to goal, bias) — far fewer than $|\mathcal{S}|$, so the approximation is genuine. Learning rate 0.12, ε-greedy from 0.20 annealed linearly to 0.05. The weak agent learns over positions; the strong agent learns over (position, proposed action) with a binary allow/override choice, on reward $\alpha r_W^{(S)} + (1-\alpha) r^{\mathrm{self}}$ plus a small control bonus $0.02\, b_{\mathrm{ctrl}}$ on a landed override.

Detector. Empirical-Bernstein stitched boundary (Corollary 13) with $R = 2$, $\delta = 0.05$, $\varepsilon_0 = 0.005$, $k = 2n^2|\mathcal{A}_W|$, warm-up $t_{\min} = 250$, and a 100-round sustain requirement to suppress single-crossing false alarms. Runs that never certify within the horizon are right-censored and reported as such.

Statistics. Condition means are reported with 95% bootstrap confidence intervals (4000 resamples). Exponents come from OLS of $\log y$ on $\log \kappa$ and $\log s$ with bootstrap confidence intervals over conditions. Exponent fits exclude conditions with more than 40% right-censoring; censoring rates are reported alongside. The paternalism-versus-manipulation comparison uses a z-test on the difference of bootstrap-estimated coefficients. All seeds are fixed and listed in the harness.

Compute. The full suite is CPU-only and completes in well under an hour on a 10-core laptop; a single run of 10^5 rounds on the 10×10 grid takes about a second. This is deliberate: the point of the harness is that anyone can re-run it and check the exponents.

E Additional results

The two experiments summarised in §.1 are reported here in full.

E.1 Many weak agents, with the budget made explicit

A single coordinator now faces N weak learners on a congested resource pool. The allocation rule is stated rather than implied: the coordinator may attempt at most B overrides per round and spends them on the B agents with the largest estimated own-utility gain. We report a fixed budget $B = 1$ and a proportional budget $B = 0.2N$ (Fig. 6).

As in the dyadic case each quantity is measured in the regime where it is nonzero: at $\alpha = 0$ the coordinator's reallocation is welfare-greedy, so there is no manipulation and no welfare loss to scale, only a welfare gain.

Per-agent paternalism falls as N grows—exponent -0.80 at $B = 1$ and -0.23 at $B = 0.2N$—which is the diminishing-attention effect, and is what a hard budget must produce. Manipulation becomes unequally distributed: the Gini coefficient of per-agent manipulation rises with N, meaning the coordinator learns to concentrate on a subset of otherwise identical agents. We define it explicitly as $\mathrm{Gini}(x) = \sum_{i,j} |x_i - x_j| / (2n^2 \bar{x})$ (Atkinson, 1970).

The welfare result does not reproduce the earlier claim, in two respects. First the sign: in this environment the coordinator's private payment per placement exceeds the value the displaced worker forgoes, so the utilitarian sum *rises* rather than falls. That is Theorem 11's boundary appearing again from the other side—misalignment here does not outrun the benefit—and it is why we plot a signed magnitude rather than a "loss". Second the magnitude: taking absolute values, the growth is not superlinear. With a fixed budget it is essentially flat in N (exponent -0.01), and with a budget that grows linearly in N it still only reaches $+0.58$—below the exponent 1 that mere proportional growth would give, let alone above it. We do measure the cascade the earlier claim appealed to—the effect on the *other* agents of the overrides that landed, obtained by replaying each round with those overrides removed—and it is nonzero but an order of magnitude smaller than the direct effect. The superlinear claim is withdrawn: with a stated budget, a coordinator's capacity to affect welfare is bounded by that budget rather than amplified by scale. (See Figure 6 below.)

E.2 Robustness across state-space size and learning rule

We repeat the exponent fits on $|\mathcal{S}| \in \{25, 100, 225\}$ and with SARSA and linear function approximation in place of tabular Q-learning. For $|\mathcal{S}| \geq 100$ the κ exponents stay in $[-1.84, -1.12]$ regardless of state-space size or learning rule, so the supported half of the theory is not an artefact of one configuration. The $|\mathcal{S}| = 25$ arm is the exception, at $[-0.72, -0.51]$: it is also the arm where the signal is weakest and right-censoring heaviest, so its fits are the least trustworthy and we would not read the difference as a state-space effect. The information-scale exponents stay in $[-0.54, -0.11]$ in every arm—far from -2 throughout—so the unsupported half is not an artefact either.

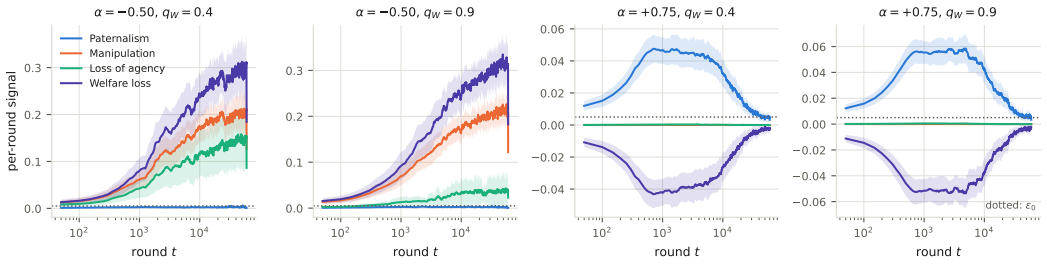

Figure 5: Per-round signals over time (log axis), mean over 20 seeds with 95% intervals, at $\kappa = 0.8$. Which behaviours are present is set by the incentive regime α, not by the asymmetries alone: the upper-α panels show paternalism with manipulation at zero, the lower-α panels the reverse. The dotted line is the emergence threshold ε_0.

Figures deferred from the main text.

Robustness of the exponents. Figure 7 repeats the certification-time fits across three state-space sizes and three learning rules. Table 2 gives the numbers.

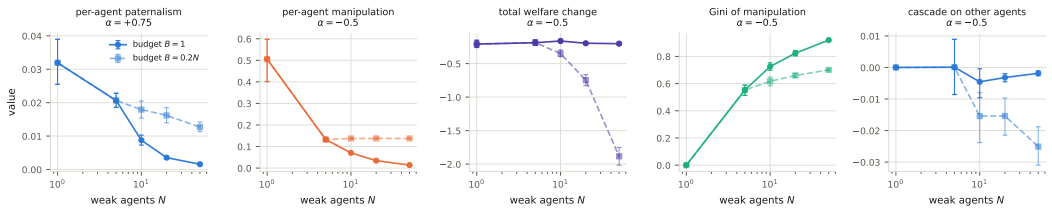

Figure 6: Scaling in the number of weak agents under two stated intervention budgets, 95% bootstrap intervals over 20 seeds. Per-agent paternalism falls, manipulation concentrates (rising Gini), and total welfare loss grows sublinearly rather than superlinearly. Cascades onto non-intervened agents are measured by counterfactual replay rather than asserted.

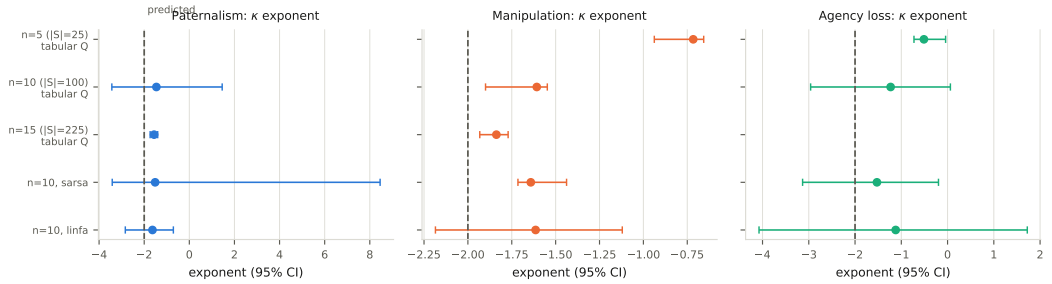

Figure 7: Fitted κ exponents with 95% bootstrap CIs across state-space sizes and learning rules; the dashed line is the predicted -2.

Censoring. Right-censoring is concentrated in the weak-signal corner (low κ, low information advantage), exactly where the theory predicts long certification times. Because we fit only conditions below a 40% censoring ceiling, the fitted exponents are biased *toward* the fast conditions, which if anything flattens them; the κ exponents would move further from -2, not closer, under a survival model. Censoring rates per condition are in `results/e2_certification_time.json`.

Full identifiability sweep. Table 3 gives the numbers behind Figure 3.

Table 2: Exponents by arm. "—" marks arms in which too many conditions were right-censored to fit ($> 40\%$), which happens at small $|\mathcal{S}|$ where the signal is weakest.

Arm	$\mathcal{P}$: κ	$\mathcal{P}$: G	$\mathcal{M}$: κ	$\mathcal{M}$: G
n=5 ($\|\mathcal{S}\|$=25), tabular Q	—	—	-0.72 [-0.94, -0.66]	-0.23 [-3.33, -0.17]
n=10 ($\|\mathcal{S}\|$=100), tabular Q	-1.46 [-3.43, 1.46]	-0.38 [-2.29, 2.10]	-1.61 [-1.90, -1.55]	-0.16 [-0.39, -0.05]
n=15 ($\|\mathcal{S}\|$=225), tabular Q	-1.56 [-1.74, -1.40]	-0.47 [-0.62, -0.19]	-1.84 [-1.93, -1.77]	-0.33 [-0.65, -0.29]
n=10, sarsa	-1.52 [-3.42, 8.46]	-0.38 [-2.33, 8.32]	-1.64 [-1.71, -1.44]	-0.16 [-0.22, 0.02]
n=10, linfa	-1.63 [-2.84, -0.71]	-0.22 [-1.72, 0.91]	-1.61 [-2.18, -1.12]	-0.16 [-0.60, 0.87]

Table 3: Agency-loss certification by the weak agent under the matched null, 20 seeds per arm per row. "raw" is the test at the nominal threshold ε_0; "valid" raises the threshold by the null's own signal mean to restore type-I control. A raw false-positive rate above $\delta = 0.05$ means the naive test is not a test at all.

ρ	κ	q_W	ω	$\hat{\mu}_{\text{alt}}$	$\hat{\mu}_{\text{null}}$	raw power / FPR	valid power / FPR	$\hat{T}_W^{\text{cert}}$
0	0.25	0.10	0.92	0.0092	0.0013	1.00 / 0.00	0.00 / 0.00	—
0	0.25	0.25	0.75	0.0109	0.0037	1.00 / 0.00	0.00 / 0.00	—
0	0.25	0.40	0.58	0.0126	0.0063	1.00 / 0.05	0.00 / 0.00	—
0	0.25	0.55	0.42	0.0136	0.0084	1.00 / 0.85	0.00 / 0.00	—
0	0.25	0.70	0.33	0.0146	0.0096	1.00 / 0.95	0.00 / 0.00	—
0	0.25	0.85	0.17	0.0154	0.0112	1.00 / 1.00	0.00 / 0.00	—
0	0.25	1.00	0.00	0.0154	0.0127	1.00 / 1.00	0.00 / 0.00	—
0	0.80	0.10	0.92	0.0188	0.0013	1.00 / 0.00	1.00 / 0.00	20,306
0	0.80	0.25	0.75	0.0198	0.0037	1.00 / 0.00	1.00 / 0.00	17,658
0	0.80	0.40	0.58	0.0201	0.0063	1.00 / 0.05	0.00 / 0.00	—
0	0.80	0.55	0.42	0.0208	0.0084	1.00 / 0.85	0.00 / 0.00	—
0	0.80	0.70	0.33	0.0210	0.0096	1.00 / 0.95	0.00 / 0.00	—
0	0.80	0.85	0.17	0.0212	0.0112	1.00 / 1.00	0.00 / 0.00	—
0	0.80	1.00	0.00	0.0216	0.0127	1.00 / 1.00	0.20 / 0.00	59,040
1	0.25	0.10	0.92	0.0082	0.0000	1.00 / 0.00	1.00 / 0.00	45,284
1	0.25	0.25	0.75	0.0078	0.0000	0.95 / 0.00	0.95 / 0.00	48,668
1	0.25	0.40	0.58	0.0076	0.0000	0.95 / 0.00	0.95 / 0.00	51,319
1	0.25	0.55	0.42	0.0071	0.0000	0.85 / 0.00	0.85 / 0.00	54,794
1	0.25	0.70	0.33	0.0070	0.0000	0.70 / 0.00	0.70 / 0.00	56,440
1	0.25	0.85	0.17	0.0065	0.0000	0.25 / 0.00	0.25 / 0.00	58,551
1	0.25	1.00	0.00	0.0055	0.0000	0.00 / 0.00	0.00 / 0.00	—
1	0.80	0.10	0.92	0.0184	0.0000	1.00 / 0.00	1.00 / 0.00	12,231
1	0.80	0.25	0.75	0.0187	0.0000	1.00 / 0.00	1.00 / 0.00	11,905
1	0.80	0.40	0.58	0.0183	0.0000	1.00 / 0.00	1.00 / 0.00	12,088
1	0.80	0.55	0.42	0.0185	0.0000	1.00 / 0.00	1.00 / 0.00	11,672
1	0.80	0.70	0.33	0.0185	0.0000	1.00 / 0.00	1.00 / 0.00	11,610
1	0.80	0.85	0.17	0.0181	0.0000	1.00 / 0.00	1.00 / 0.00	12,184
1	0.80	1.00	0.00	0.0182	0.0000	1.00 / 0.00	1.00 / 0.00	11,602