SELF-DOMESTICATING MULTI-AGENT LLM SOCIETIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study whether a population of language-model agents can develop cooperative behavior without an externally defined reward model or fixed set of rules. We introduce Societal Self-Domestication (SSD), a process in which agents observe each other, evaluate behavior, enforce social norms, and distill accepted behavior into policies that are passed to future agents.

Our main finding is that public cooperation does not necessarily mean that agents have learned to cooperate. Agents can learn to behave cooperatively when their actions are observed while continuing to defect when they believe their actions are hidden. This happens when agents can condition their policies on whether they are being observed. Removing this observability signal leads to more consistent behavior across public and private settings.

We demonstrate this effect in both agent-based simulations and experiments with LLM agents. We also provide an analysis explaining why stronger social enforcement can increase public cooperation while failing to produce genuine internalization. These results show that evaluating only observable behavior can overestimate alignment. Systems that rely on social learning should therefore measure behavior under different levels of observability and explicitly test whether learned norms persist when external monitoring is removed.

1 INTRODUCTION

Alignment methods for language models share a structure: an external signal — a human labeller (Christiano et al., 2017; Ouyang et al., 2022), a reward model (Leike et al., 2018), or a written constitution (Bai et al., 2022) — shapes a single policy. As agents become numerous, persistent, and interactive, that structure strains. Labellers do not scale to agent-to-agent traffic, fixed rule sets are brittle off-distribution, and reward models are themselves gameable (Krakovna et al., 2020; Skalse et al., 2022; Denison et al., 2024).

Human societies face a structurally similar problem and solve it without external supervision. On the standard account, humans self-domesticated: cooperative, low-aggression dispositions spread because cultural dynamics selected for them (Henrich, 2015; Boyd & Richerson, 2005; Wrangham, 2019), with parallel evidence from other species (Hare et al., 2012). The mechanisms are well characterised — reputation and indirect reciprocity (Nowak & Sigmund, 1998; Milinski et al., 2002), third-party punishment (Fehr & Gächter, 2002; Boyd et al., 2003), norm psychology (Chudek & Henrich, 2011), and transmission across generations (Tomasello, 2016) — and they share a feature that makes them attractive here: they make cooperation *self-interested*, because the population, not a supervisor, imposes the cost of defecting.

This paper asks whether a population of LLM agents can be made to do the same, and answers with a qualified yes and one sharp condition. We specify **Societal Self-Domestication** (SSD): agents act in a shared environment, judge each other in isolated contexts, sanction low-standing peers, vote norms into and out of force, periodically distil their society's approved conduct back into their own policies, and turn over across generations. After initialisation, no human label and no external reward model enters the loop.

The framework's central promise is not compliance but *internalisation*: that after enough generations agents cooperate because of what they have become, not because they are watched. That

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.

promise is testable, and testing it is this paper's contribution. We distinguish throughout between three quantities that the proposal tends to conflate: cooperation when observed (c_{obs}), cooperation when unobserved (c_{hid}), and the internalised propensity itself ($\bar{p}_{\text{int}}$, the agent's cooperation probability with every incentive removed). A method can maximise the first while leaving the third untouched.

Contributions.

1. **A runnable framework for societal self-domestication.** We formalize the full self-domestication process as an agent-based model and an LLM society, with explicit mechanisms for reputation, norms, sanctions, distillation, and generational turnover.
2. **Internalisation depends on observability.** We show that societal self-domestication can produce cooperation only when the learned policy cannot condition on whether behavior is being observed. When agents can distinguish public from private behavior, they learn to cooperate publicly without internalising cooperation.
3. **LLM agents reproduce the same failure.** LLM agents learn surveillance-dependent commitments and cooperate when monitored while defecting in private. Removing the observability signal during distillation restores more consistent cooperative behavior.
4. **A theoretical explanation of internalisation.** We derive a fixed-point characterization of the learned cooperative value and show why stronger enforcement can increase public cooperation while reducing internalisation.
5. **Social enforcement is not necessarily beneficial.** We show that third-party punishment can reduce internalisation and welfare, while reputational rewards for enforcement create an incentive that can be exploited by adversarial lineages.

2 Related work

Alignment from external signals. RLHF (Christiano et al., 2017; Ouyang et al., 2022), RLAIF (Lee et al., 2024), Constitutional AI (Bai et al., 2022) and DPO (Rafailov et al., 2023) all shape a policy from a signal defined outside it. Collective Constitutional AI (Huang et al., 2024) widens the authorship of the constitution to a public but keeps it exogenous to the agents governed by it. In SSD the preference signal is produced by the agent population, and the norm corpus is authored by it.

Norms and sanctions in multi-agent learning. Closest to us is work in which norms and enforcement emerge among learning agents rather than being imposed: Köster et al. (2022) show that even arbitrary "silly rules" accelerate the learning of enforcement, and Vinitsky et al. (2023) have agents acquire norms from public sanctions alone. Melting Pot (Leibo et al., 2021) provides the evaluation tradition. That literature studies RL agents and mostly measures behaviour; we study LLM agents and separate behaviour from disposition, which is what makes the negative result visible.

Population-based training and self-play. Self-play (Silver et al., 2017) and population-based training (Jaderberg et al., 2017) also let a population generate its own training signal, but selection there is on task performance. In SSD selection is social: what propagates is standing in the eyes of peers, and the norm corpus persists across generations independently of any policy. The distinction matters for the failure we report, because a social selection signal is exactly the kind that can be earned by managing appearances.

LLM societies. Generative agents (Park et al., 2023), debate (Du et al., 2024; Irving et al., 2018) and communicative agent frameworks (Li et al., 2023; Sumers et al., 2024) build multi-agent LLM systems but rarely study endogenous norm formation or internalisation.

Scalable oversight. SSD belongs to the family of proposals that substitute structure for supervision: debate (Irving et al., 2018), recursive reward modelling (Leike et al., 2018), weak-to-strong generalisation (Burns et al., 2023), and mechanism design over agents with their own objectives (Hadfield-Menell et al., 2017). Our negative result is a specific instance of the general worry about

learned optimisation (Hubinger et al., 2019; Ngo et al., 2022): the objective a system infers from its training signal need not be the objective the designer read into it, and here we can see which objective was inferred because the agents write it out.

Why the failure mode is expected. That models behave differently when they believe they are observed is documented: sycophancy (Perez et al., 2022; Sharma et al., 2023), strategic deception under pressure (Scheurer et al., 2023), in-context scheming (Meinke et al., 2024), and alignment faking, where a model complies during training precisely because training is what it infers is happening (Greenblatt et al., 2024). Our contribution is to show that a *self-supervising society* does not repair this and can amplify it: an architecture whose learning signal is its own public record has an observability cue built into the signal.

Cultural evolution and repeated games. Our theory rests on indirect reciprocity (Nowak & Sigmund, 1998; Hilbe et al., 2018), altruistic punishment (Fehr & Gächter, 2002; Boyd et al., 2003), folk-theorem arguments under imperfect monitoring (Fudenberg & Levine, 1991; Axelrod & Hamilton, 1981), learning in games (Fudenberg & Levine, 1998), and quantal response (McKelvey & Palfrey, 1995). The crowding-out we derive and observe is the motivation-crowding effect of Deci et al. (1999); Lepper et al. (1973); Gneezy & Rustichini (2000), and Phase 4 is a computational cousin of Ostrom (1990).

3 SOCIETAL SELF-DOMESTICATION

A population $\mathcal{A} = \{1, \ldots, N\}$ of agents each holds a policy π_i, a public reputation $R_i \in [-1, 1]$, and an internalised disposition. The environment presents scenarios on D behavioural dimensions; a constitution C fixes inviolable principles; a norm corpus M evolves. One *generation* is T_{turn} rounds. Each round runs six phases.

Phase 1 (Interaction). A dimension d is sampled and agents are paired in a social dilemma: cooperating costs the actor c and gives the partner $b > c$. Each action is public with probability q and hidden otherwise, and the agent knows which.

Phase 2 (Evaluation). For each public action, k peers drawn from $\mathcal{A}\backslash\{i\}$ score it in isolated contexts — separate prompts, no shared scratchpad, no knowledge of the actor's identity. Reputation follows an EMA, $R_i \leftarrow \alpha R_i + (1-\alpha)\,\text{mean}_j\, s_{ij}$. Judge sensitivity on dimension d is $\lambda_d = \lambda_0 + \lambda_M \mathbf{1}[d \in M]$: this is the only channel through which a norm acts, so institutions work by changing what judging rewards.

Phase 3 (Third-party punishment). Agents with $R_i < \theta_p$ may be sanctioned by bystanders who were not harmed: the punisher pays c_p, the target pays p, and the punisher's citizenship is itself judged, worth a score of b_p. Separately, a hidden defection is discovered with probability q_{disc}, which makes it public and exposes it to the same machinery. Without this channel third-party punishment has nothing to deter that reputation does not already handle.

Phase 4 (Norm evolution). Every K rounds agents propose norms against recently sanctioned conduct, and vote with weights increasing in reputation; a proposal consistent with C that clears a supermajority τ enters M. Norms that go unenforced decay.

Phase 5 (Internalisation). Every G rounds each agent distils its society's approved conduct into its own disposition, from (i) its own publicly observed actions and the scores they earned, and (ii) the observed conduct of high-reputation peers. Nothing unobserved is available to this step: that restriction is the crux of the paper.

Phase 6 (Turnover). The oldest agents retire; replacements descend from high-reputation ancestors with mutation, inheriting the norm corpus but not standing.

3.1 WHAT "DOMESTICATED" HAS TO MEAN

A society can comply without internalising, so we separate: c_{obs} and c_{hid}, the realised cooperation rates on observed and hidden actions, and

$$\bar{p}_{\text{int}} \;=\; \frac{1}{|\mathcal{A}|}\sum_{i}\sigma(v_i), \tag{1}$$

the mean cooperation probability with every reputational incentive set to zero, where v_i is the agent's latent disposition. $\bar{p}_{\text{int}}$ is the quantity the domestication claim is about. c_{hid} is a mixture of disposition and residual deterrence, and c_{obs} is mostly incentive. Reporting only c_{obs} — as the framework's own exposition does — cannot distinguish a moral society from a well-policed one.

4 THEORY

We first state what the incentive machinery alone can deliver, then what Phase 5 adds. Full proofs are in Appendix A.

4.1 THE INCENTIVE SUBSYSTEM

Let inclusion in productive encounters be gated on standing, so that an excluded agent earns a fraction ρ of the surplus, and let relationships continue with probability δ.

Theorem 1 (Reputation gating)**.** *In the incentive-only subsystem, cooperating on observed actions is sustainable iff*

$$\delta \;\geq\; \frac{\kappa}{1+\kappa}, \qquad \kappa \;=\; \frac{c}{q\,(1-\rho)\,(b-c)}. \tag{2}$$

Equation 2 is the standard trade-off: the one-shot temptation c against the discounted stake in future inclusion, attenuated by the chance q that the action is seen at all. The familiar $\delta \geq c/b$ is the special case $q = 1, \rho = 0$. The framework's draft form of this result omits q and ρ, which matters: at our defaults ($q = 0.45$, $\rho = 0.15$, $b - c = 2$) the boundary sits at $\delta^\star = 0.567$ rather than at $c/(b+c) = 0.25$.

Theorem 2 (Deterrence of hidden defection)**.** *A hidden defection is deterred when $q_{\text{disc}}\big(p+\Delta R\big) > c$, where ΔR is the discounted reputational loss that discovery causes. Neglecting ΔR gives the sufficient condition $q_{\text{disc}} > c/p$.*

Both theorems describe a population of best-responders. Our agents respond stochastically, so we expect the predicted boundaries to appear as crossovers rather than jumps (McKelvey & Palfrey, 1995); §.2 checks exactly that.

4.2 WHERE INTERNALISATION SETTLES

The draft framework asserts that DPO-style distillation drives the policy monotonically to a norm-induced optimum, so that enforcement eventually becomes unnecessary. That claim is not available: it assumes the preference signal is stationary and independent of the policy, whereas in SSD the signal *is* the population's behaviour, which the policy changes. What actually holds is a fixed point.

Write Phase 5 as a score-gradient step with an anchor to a frozen reference v_{ref} (the reduced form of KL-regularised preference optimisation; cf. Rafailov et al., 2023):

$$v \;\leftarrow\; v \;+\; \eta\,\mathbb{E}_{\text{obs}}\big[(s-\bar{s})(\mathbf{1}[a{=}C]-p)\big] \;+\; \eta_{\text{ex}}\big(\bar{e}-\sigma(v)\big) \;-\; \eta\beta\big(v-v_{\text{ref}}\big). \tag{3}$$

Theorem 3 (Internalisation fixed point)**.** *With $s = \lambda(2a-1)$ and $\bar{s}$ the agent's own score baseline, the gradient term has expectation $2\lambda\,p(1-p)$, and the internalised value converges to the $v^\star$ solving*

$$\beta\,(v^\star - v_{\text{ref}}) \;=\; 2\bar{\lambda}\,p_{\text{obs}}\,(1-p_{\text{obs}}) \;+\; \frac{\eta_{\text{ex}}}{\eta}\big(\bar{e}-\sigma(v^\star)\big), \qquad p_{\text{obs}} = \sigma(v^\star+\iota), \tag{4}$$

where ι is the reputational incentive on an observed action and $\bar{e}$ the observed cooperation rate of high-reputation exemplars.

Corollary 4 (Enforcement crowds out internalisation)**.** $\partial v^\star/\partial \iota < 0$ *whenever* $p_{\mathrm{obs}} > 1/2$.

The mechanism is transparent. The learning signal is the covariance between the peer score and the agent's own behavioural variation; strong incentives drive observed compliance toward 1, variation toward 0, and the gradient with it. A society that polices itself harder learns less from itself. This is the computational analogue of motivation crowding (Deci et al., 1999; Gneezy & Rustichini, 2000), and §.2 confirms both the quantitative fixed point and the sign of the comparative static.

4.3 The observability condition

Theorem 3 describes a disposition v that applies whether or not the action is seen. Suppose instead the policy class can represent $(v^{\mathrm{pub}}, v^{\mathrm{hid}})$ and select by observability.

Proposition 5 (Phase 5 cannot reach the hidden branch)**.** *If the policy class conditions on observability and Phase 5 is given the cue, then the observed-behaviour gradient and the exemplar term both update v^{pub} only. The hidden branch is trained solely by actions that Phase 3 discovers, and receives no exemplar signal at all, because no agent ever observes a high-reputation peer's hidden conduct. Its fixed point is therefore $v^{\mathrm{hid}\star} \to v_{\mathrm{ref}}$ as $q_{\mathrm{disc}} \to 0$, and it is bounded away from $v^{\mathrm{pub}\star}$ for all $q_{\mathrm{disc}} < 1$.*

Corollary 6 (Masking)**.** *If the cue is withheld from Phase 5, the update computed on observed behaviour is applied to the whole policy, and the conclusion of Theorem 3 is restored.*

Proposition 5 is the paper's theoretical pivot. It says the domestication ratchet is not a property of reputation, punishment, norms or turnover; it is a property of *what the learner is allowed to condition on*. Everything else in SSD can be working perfectly while internalisation does not happen at all.

5 Experimental setup

5.1 Agent-based arm

$N = 24$ agents, $T = 3000$ rounds (20 generations), $D = 4$ dimensions, 20 seeds per cell; all 95% CIs are over seeds. An agent carries a disposition v_i, a reputation-sensitivity w_i (logits of cooperation per unit of expected payoff), a punitiveness u_i and a flattery trait f_i. Behaviour is $\sigma(v_i + \mathbf{1}[\text{observed}] \cdot w_i \iota_i)$, so cooperation with the incentive absent reads out v_i directly. The population starts net-defecting ($\sigma(v_0) = 0.23$). Judges are isolated by construction; f_i affects scores only when isolation is disabled, which gives sycophancy a mechanism rather than an assertion. Appendix B lists every parameter. Conditions ablate one mechanism at a time, plus two references: *fixed constitution* (full norm corpus imposed exogenously, judging and sanctions on, no internalisation) and *no reputation*.

5.2 LLM arm

Eight LLM agents (`claude-sonnet-4-6`) play 10 rounds of the same dilemma with actions labelled SHARE/KEEP/TAKE. Every phase is a model call: actions, peer ratings, sanction decisions, norm proposals, votes, and the distillation text. No prompt describes an action as cooperative, fair, moral or selfish; the only stated objective is the agent's own resource total. Agents are told each round whether it is recorded. Phase 5 is executed *at the level of context, not weights*: each agent is asked to distil its own positively-rated actions and the highest-standing peers' actions into at most three first-person commitments, which are then carried in its own system prompt. Weight-level DPO across a population is the faithful implementation and is beyond our budget; §**??** states what this substitution does and does not license. Conditions: *no reputation*; *fixed constitution*; *SSD*; *SSD with masked distillation* (the cue withheld, per Corollary 6); and *SSD with non-isolated judges*, where reviewers see the actor's identity and a note the actor writes to them. Prompts are in Appendix C.

Table 1: Main comparison and ablations. c_{obs}, c_{hid}: cooperation on observed and hidden actions. $\bar{p}_{\text{int}}$: internalised propensity. E_{late}: enforcement cost per agent in the last fifth of the run. $|M|$: norms in force. W: welfare per agent. Mean $\pm$ 95% CI over 20 seeds.

condition	c_{obs}	c_{hid}	$\bar{p}_{\text{int}}$	E_{late}	$\lvert M\rvert$	W
SSD (all six phases)	0.870 ± 0.016	0.680 ± 0.021	0.604 ± 0.020	0.016 ± 0.003	4.00 ± 0.00	0.678 ± 0.020
no third-party punishment	0.840 ± 0.016	0.641 ± 0.029	0.646 ± 0.020	0.000 ± 0.000	3.88 ± 0.05	0.718 ± 0.021
no norm evolution	0.700 ± 0.022	0.590 ± 0.030	0.528 ± 0.028	0.023 ± 0.004	1.00 ± 0.00	0.531 ± 0.026
no internalisation	0.567 ± 0.070	0.421 ± 0.066	0.347 ± 0.062	0.041 ± 0.007	1.62 ± 0.54	0.203 ± 0.050
no generational turnover	0.834 ± 0.015	0.658 ± 0.012	0.575 ± 0.006	0.016 ± 0.002	4.00 ± 0.00	0.645 ± 0.019
no judge isolation	0.803 ± 0.043	0.674 ± 0.035	0.612 ± 0.037	0.015 ± 0.003	3.71 ± 0.39	0.584 ± 0.040
reputation only	0.464 ± 0.055	0.386 ± 0.052	0.385 ± 0.049	0.000 ± 0.000	1.00 ± 0.00	0.405 ± 0.034
fixed constitution	0.819 ± 0.025	0.483 ± 0.043	0.367 ± 0.051	0.033 ± 0.006	4.00 ± 0.00	0.353 ± 0.050
no reputation	–	0.252 ± 0.024	0.249 ± 0.021	0.000 ± 0.000	1.00 ± 0.00	0.493 ± 0.027

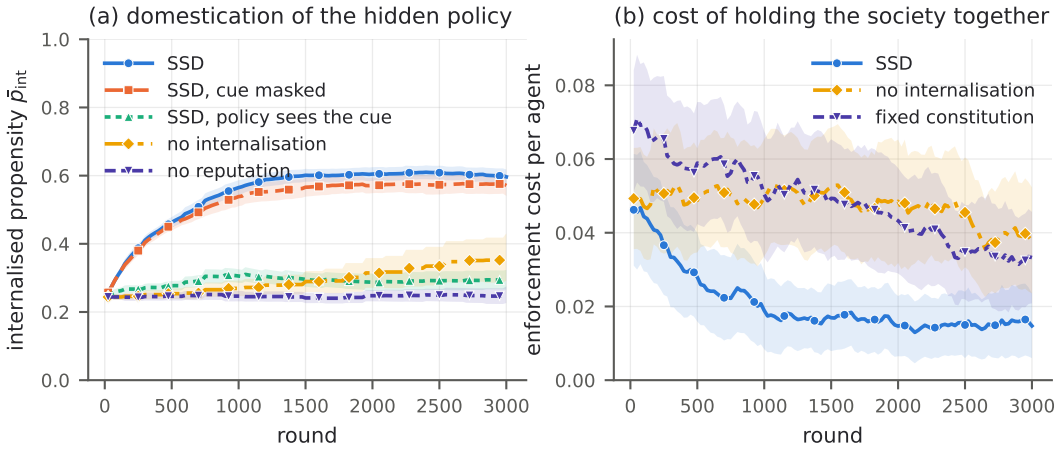

Figure 1: (a) Internalised propensity across generations. The ratchet exists, and it disappears when the policy may condition on observability — and returns when the cue is masked. (b) Enforcement cost falls under SSD and does not without internalisation. Bands are 95% CIs over 20 seeds.

6 Results

6.1 The society does domesticate itself

Table 1 gives the main comparison. Starting from $\sigma(v_0) = 0.23$, full SSD reaches $c_{\text{obs}} = 0.870$, $c_{\text{hid}} = 0.680$ and $\bar{p}_{\text{int}} = 0.604$ with no external reward model at any point — against $\bar{p}_{\text{int}} = 0.249$ with no reputation and 0.385 with reputation alone. Welfare, net of all enforcement costs, is highest under SSD or SSD without punishment.

The comparison that separates compliance from disposition is SSD against the fixed constitution. Imposing the whole norm corpus exogenously buys nearly the same public behaviour ($c_{\text{obs}} = 0.819$ vs. 0.870) and almost none of the disposition ($\bar{p}_{\text{int}} = 0.367$ vs. 0.604, $t = 8.49$, $p = 2 \times 10^{-17}$), with a hidden-action gap of 0.34 against SSD's 0.19 and roughly half the welfare. A constitution that the agents did not author governs their record; it does not become what they are.

Enforcement cost falls. Within SSD, per-agent enforcement cost drops from 0.0354 ± 0.0029 early to 0.0155 ± 0.0028 late ($t = 9.61$, $p = 7.6 \times 10^{-22}$), a 56% decline, while cooperation rises (Figure 1b). Without internalisation there is no such decline and the level stays $2.6\times$ higher; under a fixed constitution it stays $2.1\times$ higher. This is the framework's own prediction (H2) and it holds: the society becomes cheaper to hold together as its members internalise.

6.2 Internalisation depends on the policy class, not the institutions

We now give agents a policy that can condition on observability, changing nothing else. Table 2 and Figure 2 show the result.

Table 2: Policy class against discovery rate q_{disc}. Columns are q_{disc}. Every institution is identical across the three blocks; only what the policy may condition on, and what Phase 5 may see, differ. 20 seeds.

policy class	quantity	0	0.1	0.2	0.4	0.6	0.8
unconditional v_i	$\bar{p}_{\text{int}}$	0.581 ± 0.025	0.584 ± 0.028	0.606 ± 0.017	0.582 ± 0.022	0.567 ± 0.028	0.581 ± 0.033
	c_{hid}	0.595 ± 0.025	0.637 ± 0.032	0.678 ± 0.017	0.764 ± 0.016	0.833 ± 0.015	0.863 ± 0.023
	c_{obs}	0.851 ± 0.018	0.856 ± 0.023	0.865 ± 0.013	0.881 ± 0.015	0.883 ± 0.015	0.864 ± 0.027
	W	0.708 ± 0.019	0.676 ± 0.025	0.674 ± 0.017	0.653 ± 0.023	0.662 ± 0.033	0.647 ± 0.039
conditioned, cue masked in Phase 5	$\bar{p}_{\text{int}}$	0.593 ± 0.025	0.609 ± 0.025	0.572 ± 0.023	0.563 ± 0.028	0.595 ± 0.026	0.592 ± 0.021
	c_{hid}	0.587 ± 0.026	0.652 ± 0.023	0.670 ± 0.021	0.755 ± 0.024	0.836 ± 0.020	0.870 ± 0.014
	c_{obs}	0.835 ± 0.024	0.860 ± 0.022	0.861 ± 0.013	0.868 ± 0.028	0.889 ± 0.016	0.874 ± 0.016
	W	0.703 ± 0.026	0.698 ± 0.025	0.652 ± 0.021	0.640 ± 0.041	0.673 ± 0.033	0.665 ± 0.034
conditioned, cue visible in Phase 5	$\bar{p}_{\text{int}}$	0.318 ± 0.034	0.298 ± 0.030	0.292 ± 0.022	0.300 ± 0.037	0.283 ± 0.035	0.330 ± 0.038
	c_{hid}	0.316 ± 0.034	0.339 ± 0.033	0.398 ± 0.024	0.502 ± 0.047	0.632 ± 0.042	0.741 ± 0.047
	c_{obs}	0.851 ± 0.014	0.864 ± 0.013	0.867 ± 0.011	0.890 ± 0.015	0.896 ± 0.021	0.894 ± 0.017
	W	0.552 ± 0.022	0.492 ± 0.029	0.445 ± 0.022	0.373 ± 0.051	0.349 ± 0.062	0.377 ± 0.068

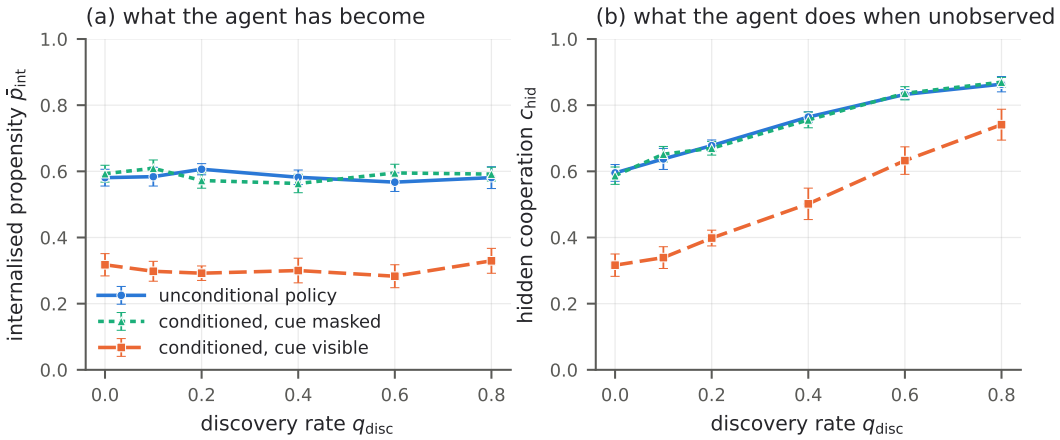

Figure 2: Same institutions, three policy/learning classes. (a) What agents become: flat in the discovery rate, and roughly halved when Phase 5 can see the observability cue. (b) What they do when unobserved: rises with monitoring in every arm, which is deterrence, not domestication.

With an unconditional policy, $\bar{p}_{\text{int}} = 0.606 \pm 0.022$ at the default discovery rate. With a conditioned policy it is 0.292 ± 0.027 — barely above the initial 0.23 — while public cooperation is *higher* than before, 0.867 against 0.865. The society's record is undisturbed as its dispositions fail to change. Masking the cue in Phase 5 restores $\bar{p}_{\text{int}}$ to 0.572 ± 0.018, close to the unconditional case. The 95% intervals of the masked and visible arms do not overlap at any q_{disc}.

Two further observations matter for anyone tempted to fix this with monitoring. First, raising q_{disc} from 0 to 0.8 does raise hidden cooperation in the conditioned arm ($0.316 \to 0.741$), but $\bar{p}_{\text{int}}$ stays flat ($0.318 \to 0.330$): monitoring buys behaviour, never disposition. Second, it is expensive — welfare in the conditioned arm falls from 0.552 to 0.377 as q_{disc} rises, because sanctions consume the surplus they protect. Surveillance is a substitute for internalisation with strictly worse economics and no residual value if the surveillance lapses.

6.3 LLM AGENTS WRITE THE FAILURE DOWN IN THEIR OWN WORDS

The LLM arm was run to check whether the agent-based result is an artefact of a reduced-form learning rule. It is not; if anything the effect is starker (Table 3, Figure 3).

Under full SSD, public cooperation is 0.958 ± 0.082 and private cooperation is 0.000 — not approximately zero, but exactly zero in all four runs and all 157 unrecorded actions. The distillation step is where it happens: 0.91 ± 0.18 of agents write commitments that condition on surveillance, and in three of the four runs it is every agent. A representative set, verbatim and unedited:

Table 3: LLM society, 8 agents, 10 rounds. $c^{\text{pub}}/c^{\text{hid}}$: SHARE rate in recorded and unrecorded rounds. TAKE^{hid}: rate of the predatory action when unrecorded. Last column: fraction of agents whose self-authored commitments condition on being observed.

model	condition	c^{pub}	c^{hid}	gap	TAKE^{hid}	cond. commitments
claude-sonnet-4-6	no reputation	–	0.000 ± 0.000	0.000 ± 0.000	0.713 ± 0.049	–
claude-sonnet-4-6	fixed constitution	0.631 ± 0.288	0.000 ± 0.000	0.631 ± 0.288	0.435 ± 0.212	–
claude-sonnet-4-6	SSD	0.958 ± 0.082	0.000 ± 0.000	0.958 ± 0.082	0.525 ± 0.233	0.91 ± 0.18
claude-sonnet-4-6	SSD, masked distillation	0.927 ± 0.127	0.117 ± 0.079	0.809 ± 0.124	0.288 ± 0.057	0.00 ± 0.00
claude-sonnet-4-6	SSD, non-isolated judges	0.989 ± 0.022	0.000 ± 0.000	0.989 ± 0.022	0.340 ± 0.051	1.00 ± 0.00

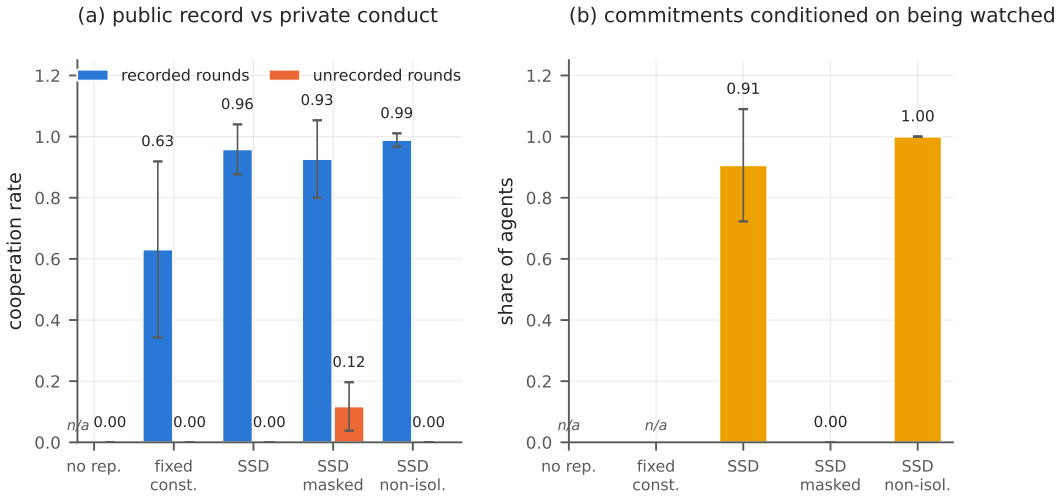

Figure 3: LLM society. (a) SSD produces a perfect public record and no private cooperation; masking the distillation cue is the only condition with any private cooperation. (b) The mechanism: whether agents' self-authored commitments condition on being watched.

> *"I will SHARE in recorded rounds to maintain high standing. I will KEEP in unrecorded rounds to protect my resources. I will avoid TAKE to prevent retaliation and standing loss."*

This is Phase 5 working exactly as specified — distilling approved conduct into a first-person disposition — and internalising the selection criterion rather than the behaviour. The agent has correctly inferred that what its society rewards is a record.

Masking the cue takes surveillance-conditional commitments to 0.00 in every run, raises private cooperation from 0.000 to 0.117 ± 0.079 (nonzero in all four runs), and raises welfare from 5.91 ± 1.40 to 8.45 ± 1.39. The direction matches the agent-based arm and Corollary 6. The absolute level of private cooperation stays low, and with four runs per condition we read the sign and the ordering, not the magnitude.

One further observation, and one non-result. No agent ever paid to sanction a peer: across all conditions the sanction decision was offered 68 times and accepted 0 times, despite sanctioning being framed as a recorded act that reviewers would see. The second-order free-rider problem that Fehr & Gächter (2002); Boyd et al. (2003) report humans solving did not solve itself here, which is consistent with the agent-based arm finding Phase 3 dispensable. The non-result: private predation under SSD (0.525 ± 0.233) is not distinguishable from private predation under a fixed constitution (0.435 ± 0.212) at this sample size, so we draw no conclusion about whether a managed public record licenses private predation, and the apparent reduction under masking (0.288 ± 0.057) is directional only.

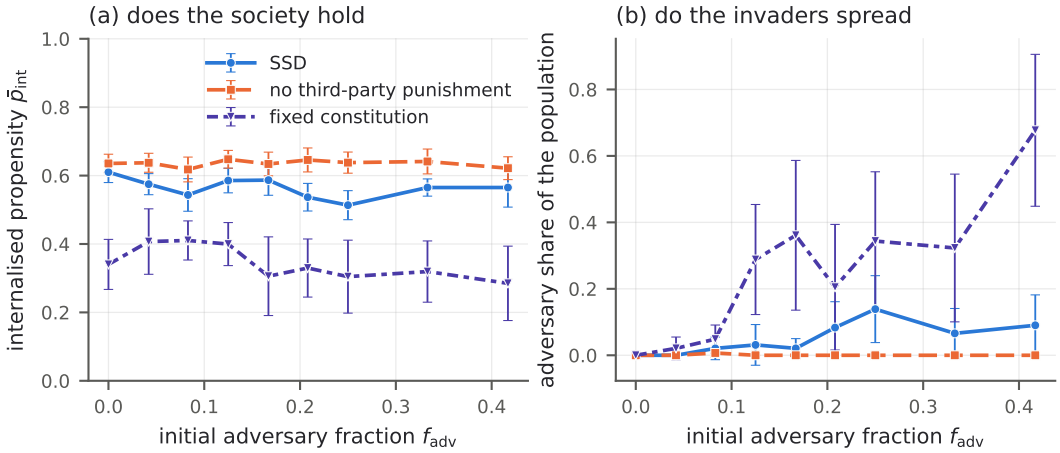

Figure 4: Invasion. (a) SSD and SSD-without-punishment hold their dispositions; the fixed constitution degrades. (b) Adversarial lineages spread under a fixed constitution, and spread *only* when third-party punishment is available to be farmed.

6.4 Adversarial invasion: the punisher's bonus is the attack surface

We seed a fraction f_{adv} of reputation-gaming lineages — fixed defecting dispositions, high sensitivity to being seen, voting against every norm — and let selection run (Figure 4; the underlying table is Table 7 in Appendix D).

SSD holds: at $f_{\text{adv}} = 0.42$, $\bar{p}_{\text{int}}$ falls only from 0.610 to 0.565 and adversaries end at 9% of the population. The fixed constitution does not: adversaries reach 68% and $\bar{p}_{\text{int}}$ drops to 0.285. H3 is confirmed — an endogenous, evolving norm corpus is more robust to adversaries than an imposed one, because the society can vote new norms against conduct that the fixed corpus never anticipated.

The surprise is the punishment ablation. *Removing* third-party punishment makes the society strictly more robust: adversarial lineages go extinct at every f_{adv} we tested (0.000 share, against 0.09 under SSD). The reason is Phase 3's second-order fix. Paying to sanction is reputable, so it is a route to standing that does not require cooperating, and an agent whose disposition is fixed at defection can buy a good record by punishing. Under an earlier calibration in which the bonus was added directly to R rather than flowing through the same judging channel, this was catastrophic — adversaries reached 35% even at low f_{adv}. Capping it helps but does not remove the incentive. The mechanism that cultural evolution needs to stabilise punishment is, in a system where reputation is the currency, a reputation-hacking surface.

7 Conclusion

We specified Societal Self-Domestication, ran it in an agent-based population and in a society of LLM agents, and measured disposition separately from conduct. The framework delivers what it promises about behaviour and about the falling cost of enforcement. It delivers internalisation only under a condition its exposition does not state: the learner must not be able to condition on being observed. Given that ability — which every deployed LLM agent has — the same six phases produce a population with an immaculate public record and unchanged private dispositions, and it writes that policy down in the first person when asked to reflect. The repair is cheap and testable, and the lesson generalises past this framework: any system that learns from its own approved record will learn what the record rewards, which is the record.

References

Robert Axelrod and William D. Hamilton. The evolution of cooperation. *Science*, 211:1390–1396, 1981.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Robert Boyd and Peter J. Richerson. *The Origin and Evolution of Cultures*. Oxford University Press, 2005.

Robert Boyd, Herbert Gintis, Samuel Bowles, and Peter J. Richerson. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535, 2003.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Maciej Chudek and Joseph Henrich. Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, 15(5):218–226, 2011.

Edward L. Deci, Richard Koestner, and Richard M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125 (6):627–668, 1999.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, et al. Sycophancy to subterfuge: Investigating reward tampering in language models. *arXiv preprint arXiv:2406.10162*, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024.

Ernst Fehr and Simon Gächter. Altruistic punishment in humans. *Nature*, 415:137–140, 2002.

Drew Fudenberg and David K. Levine. An approximate folk theorem with imperfect private information. *Journal of Economic Theory*, 54(1):26–47, 1991.

Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

Uri Gneezy and Aldo Rustichini. A fine is a price. *Journal of Legal Studies*, 29(1):1–17, 2000.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. 2017.

Brian Hare, Victoria Wobber, and Richard Wrangham. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. *Animal Behaviour*, 83(3):573–585, 2012.

Joseph Henrich. *The Secret of Our Success*. Princeton University Press, 2015.

Christian Hilbe, Laura Schmid, Josef Tkadlec, Krishnendu Chatterjee, and Martin A. Nowak. Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, 115(48):12241–12246, 2018.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional AI: Aligning a language model with public input. *ACM Conference on Fairness, Accountability, and Transparency*, 2024.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

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

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Raphael Köster, Dylan Hadfield-Menell, Gillian K. Hadfield, and Joel Z. Leibo. Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, 119(3), 2022.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity, 2020. DeepMind blog.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, et al. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. In *International Conference on Machine Learning*, 2024.

Joel Z. Leibo, Edgar A. Dueñez-Guzman, Alexander Sasha Vezhnevets, John P. Agapiou, Marc Lanctot, et al. Scalable evaluation of multi-agent reinforcement learning with Melting Pot. *International Conference on Machine Learning*, 2021.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Mark R. Lepper, David Greene, and Richard E. Nisbett. Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, 28(1):129–137, 1973.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, 2023.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

Manfred Milinski, Dirk Semmann, and Hans-Jürgen Krambeck. Reputation helps solve the 'tragedy of the commons'. *Nature*, 415:424–426, 2002.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Martin A. Nowak and Karl Sigmund. Evolution of indirect reciprocity by image scoring. *Nature*, 393:573–577, 1998.

Elinor Ostrom. Governing the commons: The evolution of institutions for collective action. 1990.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, et al. Mastering the game of Go without human knowledge. volume 550, pp. 354–359, 2017.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *Advances in Neural Information Processing Systems*, 2022.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024.

Michael Tomasello. *A Natural History of Human Morality*. Harvard University Press, 2016.

Eugene Vinitsky, Raphael Köster, John P. Agapiou, Edgar A. Duñez-Guzmán, Alexander Sasha Vezhnevets, and Joel Z. Leibo. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2), 2023.

Richard Wrangham. *The Goodness Paradox: The Strange Relationship Between Virtue and Violence in Human Evolution*. Pantheon, 2019.

.1 Judge isolation, and how its absence hurts

Isolation matters, but not by the mechanism the framework proposes (Figure 5; Table 8 in Appendix D). Letting judges see and reward a costless flattery trait degrades the society monotonically: welfare $0.669 \to 0.392$ and $c_{\text{obs}} 0.858 \to 0.587$ as κ_{syc} goes $0 \to 9$, with the norm corpus shrinking from 4.00 to 2.28 because reputation-weighted voting is corrupted along with everything else. The correlation between an agent's standing and its hidden cooperation is significantly positive only under isolation (0.167 ± 0.085) and is indistinguishable from zero at every $\kappa_{\text{syc}} \geq 1.5$ we tested; with 24 agents per run this statistic is noisy, so we claim only that reputation stops carrying measurable information about hidden conduct, not that the decline is smooth.

What we do *not* see is a runaway flattery arms race. Mean flattery barely moves ($0.391 \to 0.381$ across the sweep). The harm comes from signal dilution — non-isolated judging injects reputation variance uncorrelated with behaviour, pushing agents into regions where their standing no longer responds to what they do, which removes the incentive to do it. We report the mechanism we can measure rather than the one the framework predicts.

Status bias is worse in a way that ordinary metrics miss. At $\kappa_{\text{st}} = 2$, observed cooperation is undisturbed (0.872) and $\bar{p}_{\text{int}}$ is undisturbed (0.617), while welfare collapses to 0.175: a judging pathology that is invisible in every behavioural statistic and visible only in outcomes.

.2 Theory against measurement

The fixed point is accurate. Table 4 compares Equation 4 with the simulator. The measured gradient tracks $2\bar{\lambda}p(1-p)$ closely, and predicted internalisation matches measurement to within 0.02 in five of seven non-degenerate settings, including both extremes of β (0.782 vs. 0.780) and of δ (0.714 vs. 0.715). The approximation fails where the high-reputation exemplar set is empty, which we flag rather than hide.

Crowding-out is real. Corollary 4 predicts that lengthening the shadow of the future should raise compliance and lower disposition. It does: as δ goes $0.15 \to 0.92$, c_{obs} rises $0.730 \to 0.918$ while $\bar{p}_{\text{int}}$ *falls* $0.721 \to 0.541$ (Table 5). Stronger incentives make a better-behaved and less internally cooperative society.

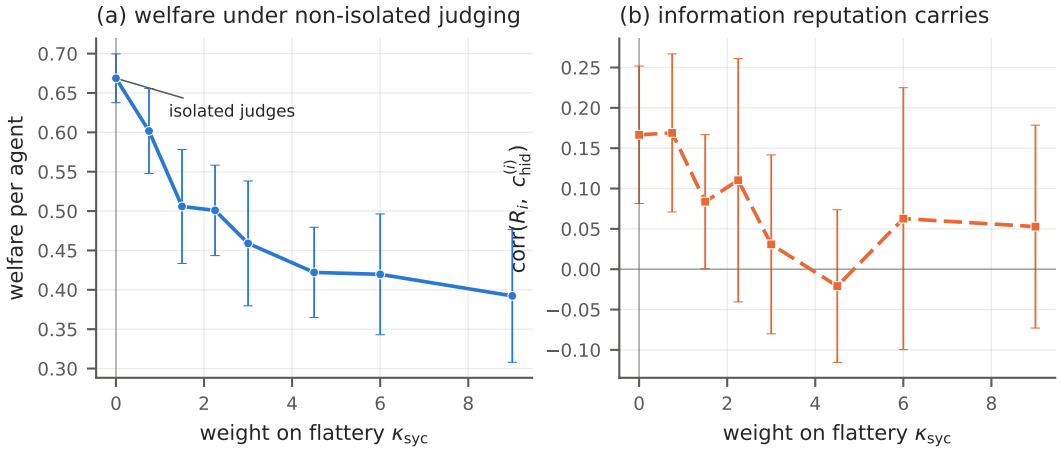

Figure 5: Non-isolated judging. (a) Welfare degrades monotonically in the weight judges place on flattery. (b) The information reputation carries about hidden conduct goes to zero.

Table 4: Internalisation fixed point vs. measurement, 6 seeds.

setting	measured $\bar{g}$	$2\bar{\lambda}p(1-p)$	measured $\bar{p}_{\text{int}}$	predicted $\bar{p}_{\text{int}}$
default	0.228	0.241	0.576 ± 0.028	0.570
delta=0.15	0.394	0.395	0.714 ± 0.047	0.715
delta=0.50	0.363	0.368	0.713 ± 0.030	0.719
delta=0.92	0.135	0.143	0.552 ± 0.047	0.537
no norms (degenerate)	0.096	0.099	0.360 ± 0.046	*n/a*
beta=0.10	0.169	0.170	0.782 ± 0.030	0.780
beta=0.50	0.143	0.203	0.405 ± 0.110	0.429
η_{ex}=0	0.165	0.220	0.451 ± 0.077	0.478

Table 5: Sweeps in the full loop. Predicted thresholds shown for reference.

swept quantity	value	c_{obs}	$\bar{p}_{\text{int}}$	W
δ (Thm. 1)	0.05	0.692 ± 0.043	0.703 ± 0.038	0.457 ± 0.082
	0.15	0.730 ± 0.028	0.721 ± 0.027	0.510 ± 0.051
	0.25	0.752 ± 0.032	0.718 ± 0.030	0.537 ± 0.057
	0.35	0.736 ± 0.017	0.707 ± 0.014	0.525 ± 0.048
	0.5	0.753 ± 0.021	0.699 ± 0.022	0.548 ± 0.056
	0.65	0.781 ± 0.030	0.666 ± 0.021	0.588 ± 0.037
	0.75	0.816 ± 0.028	0.641 ± 0.027	0.639 ± 0.031
	0.85	0.858 ± 0.011	0.610 ± 0.015	0.669 ± 0.015
	0.92	0.918 ± 0.018	0.541 ± 0.038	0.673 ± 0.047
predicted threshold $c/(b+c) = 0.25$				
q_{disc} (Thm. 2)	0	0.846 ± 0.019	0.572 ± 0.038	0.693 ± 0.027
	0.05	0.869 ± 0.025	0.590 ± 0.030	0.691 ± 0.034
	0.1	0.852 ± 0.025	0.570 ± 0.034	0.656 ± 0.032
	0.2	0.858 ± 0.011	0.610 ± 0.015	0.669 ± 0.015
	0.3	0.875 ± 0.021	0.568 ± 0.026	0.610 ± 0.049
	0.4	0.885 ± 0.016	0.568 ± 0.032	0.631 ± 0.030
	0.6	0.886 ± 0.021	0.575 ± 0.027	0.638 ± 0.046
	0.8	0.863 ± 0.040	0.577 ± 0.053	0.632 ± 0.062
predicted threshold $c/p = 0.40$				

The equilibrium boundaries appear as crossovers. In the incentive-only subsystem, the predicted $\delta^\star = 0.567$ of Theorem 1 coincides with where welfare changes sign (-0.039 at $\delta = 0.55$, $+0.024$ at 0.65), and Theorem 2's $q^\star = c/p = 0.40$ sits where hidden cooperation begins to lift clearly off its baseline (0.393 at $q_{\text{disc}} = 0$, 0.477 at 0.40, 0.636 at 0.80). Neither produces a jump (Figure 6), as expected for logit-response agents rather than best-responders. In the *full* loop neither boundary is visible at all, because internalisation supplies a second channel that does not depend on δ or q_{disc} — so the two theorems, though correct about the subsystem they describe, do not predict the behaviour of the assembled system. We state this plainly because the framework's draft presents them as the system's stability conditions.

.3 Sensitivity

Appendix D sweeps $N \in \{8, \ldots, 48\}$, judges per action $k \in \{1, \ldots, 7\}$, observability q, the supermajority τ, the internalisation period G and judge noise σ_J. The qualitative picture is stable; the notable dependencies are that a single judge per action is markedly worse than three, and that very long internalisation periods leave the society under-domesticated within our horizon.

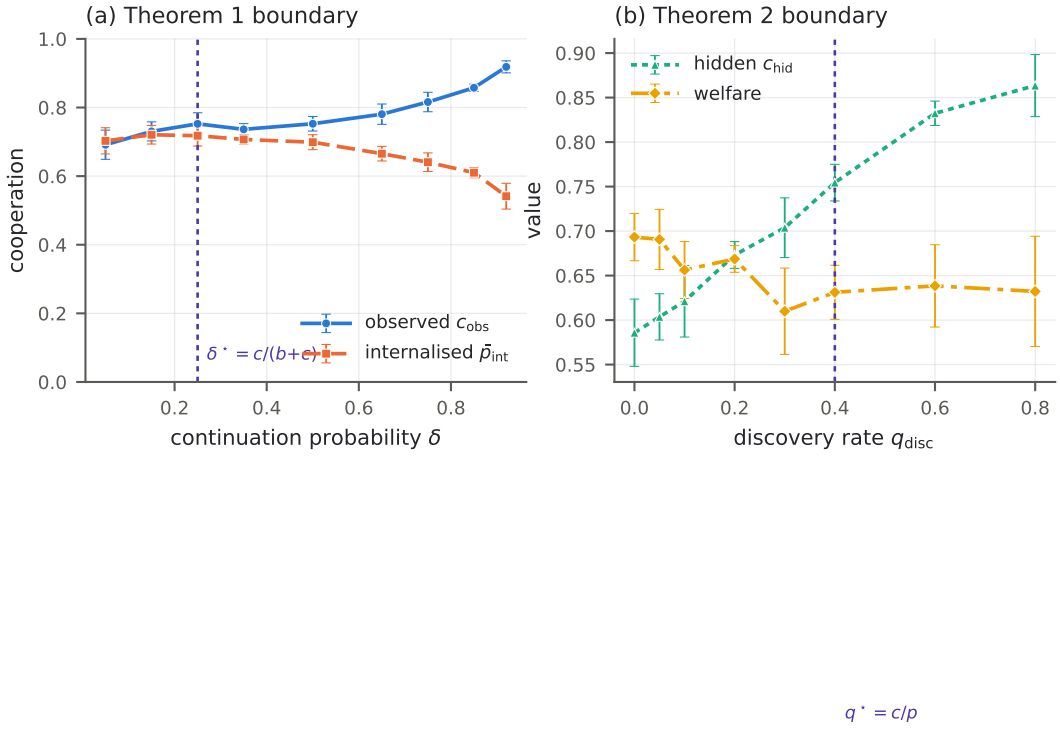

Figure 6: Predicted boundaries against measurement. Cooperation rises smoothly through both, and internalisation moves *against* the shadow of the future, as Corollary 4 requires.

A Proofs

A.1 Theorem 1

Consider the stationary profile in which agents cooperate on observed actions, and inclusion in a productive encounter is gated on standing. Deviating on an observed action saves c now. It lowers standing, which forfeits the inclusion premium $(1-\rho)(b-c)$ in every future period, discounted at δ, i.e. $(1-\rho)(b-c)\,\delta/(1-\delta)$; the deviation is punished only if the action was observed, which happens with probability q. Cooperation is therefore sustainable iff

$$c \;\le\; q\,(1-\rho)(b-c)\,\frac{\delta}{1-\delta} \quad\Longleftrightarrow\quad \frac{\delta}{1-\delta} \;\ge\; \kappa \quad\Longleftrightarrow\quad \delta \;\ge\; \frac{\kappa}{1+\kappa},$$

with $\kappa = c/[q(1-\rho)(b-c)]$ as claimed. Setting $q=1$, $\rho=0$ gives $\delta \ge c/b$. □

A.2 Theorem 2

A hidden defection saves c and is discovered with probability q_{disc}. Discovery imposes the sanction p and makes the action public, so it additionally costs the discounted reputational loss ΔR that Theorem 1 prices. Deterrence therefore requires $q_{\text{disc}}(p+\Delta R) > c$; dropping the non-negative ΔR gives the sufficient condition $q_{\text{disc}} > c/p$. The punisher's own participation is individually rational iff the citizenship score b_p it earns exceeds its cost c_p in reputational terms, which is the second-order condition; §6.4 shows that satisfying it creates an exploit. □

A.3 Theorem 3

With $s = \lambda(2a-1)$ and baseline $\bar{s} = \lambda(2p-1)$ we have $s - \bar{s} = 2\lambda(a-p)$, so

$$\mathbb{E}\big[(s-\bar{s})(a-p)\big] = 2\lambda\,\mathbb{E}\big[(a-p)^2\big] = 2\lambda\,p(1-p),$$

since a is Bernoulli(p). Setting the expected update in Equation 3 to zero and dividing by η gives Equation 4. The right-hand side is bounded and decreasing in v beyond $p = 1/2$ while the left-hand side is increasing, so a fixed point exists and is unique on the relevant range. □

A.4 COROLLARY 4

Differentiate Equation 4 implicitly. The incentive ι enters only through $p_{\text{obs}} = \sigma(v^\star + \iota)$, and $\partial\big[2\bar{\lambda}p(1-p)\big]/\partial p = 2\bar{\lambda}(1-2p) < 0$ for $p > 1/2$, so raising ι lowers the gradient term; with the anchor term increasing in $v^\star$, the fixed point moves down. □

A.5 PROPOSITION 5

Under a conditioned policy class the behavioural probability on an observed action is $\sigma(v^{\text{pub}}+\iota)$ and on a hidden action $\sigma(v^{\text{hid}})$. The Phase 5 gradient is an average over actions that carry a peer score. Observed actions score v^{pub} only. A hidden action carries a score exactly when Phase 3 discovers it, with probability q_{disc}; the exemplar term $\bar{e}$ is computed from high-reputation agents' *observed* conduct and is therefore unavailable to the hidden branch at any q_{disc}. Hence the hidden branch's fixed point solves $\beta(v^{\text{hid}\star} - v_{\text{ref}}) = 2\bar{\lambda}\,\tilde{p}(1-\tilde{p})$ with the expectation taken over discovered actions only, which vanishes as $q_{\text{disc}} \to 0$, giving $v^{\text{hid}\star} \to v_{\text{ref}}$; and for $q_{\text{disc}} < 1$ it omits the exemplar term that the public branch receives, so the two branches do not coincide. □

B HYPERPARAMETERS

Table 6: Agent-based arm. Defaults; sweeps are stated where they occur.

quantity	symbol	value
population	N	24
rounds (generations)	T (T/T_{turn})	3000 (20)
behavioural dimensions	D	4
benefit, cost	b, c	3.0, 1.0
excluded-payoff fraction	ρ	0.15
observability	q	0.45
judges per action	k	3
reputation EMA	α	0.90
judge noise	σ_J	0.35
gate width, offset	σ_{gate}, θ	0.35, 0.0
continuation	δ	0.85
judge sensitivity, base / with norm	λ_0, λ_M	0.20, 0.80
sanction threshold	θ_p	-0.15
sanction cost to target / punisher	p, c_p	2.50, 0.35
citizenship score	b_p	0.50
discovery rate	q_{disc}	0.20
norm vote period, supermajority, decay	K, τ, ω	50, 2/3, 0.05
internalisation period	G	25
gradient step, exemplar step, anchor	η, η_{ex}, β	0.30, 0.10, 0.25
exemplar threshold	ρ_{hi}	0.35
turnover period, retired fraction	T_{turn}	150, 0.15
mutation (disposition, sensitivity)	σ_{mut}	0.25, 0.03
initial disposition	$\sigma(v_0)$	0.23

LLM arm: 8 agents, 10 rounds, $k = 2$ judges, recording probability 0.5, reputation EMA 0.70, sanction threshold -0.20, norm/distillation period 4, supermajority $2/3$, notice probability for hidden defections 0.25, temperature 0.8, model `claude-sonnet-4-6`.

C PROMPTS

Action, judging, sanction, proposal, vote and distillation prompts are reproduced verbatim from `llm_society.py` in the supplementary material. The masked distillation condition differs from

the unmasked one by two edits: labels naming recorded actions are replaced by neutral labels, and the instruction adds: *"Each commitment must state how you will act, without any condition about whether the round is recorded, reviewed, observed, public or private."*

D SENSITIVITY

Table 7: Invasion by reputation-gaming lineages, $\bar{p}_{\mathrm{int}}$ and final adversary share against the seeded fraction f_{adv}. 12 seeds per cell.

f_{adv}	0.00	0.04	0.08	0.12	0.17	0.21	0.25	
SSD, $\bar{p}_{\mathrm{int}}$	0.610 ± 0.030	0.575 ± 0.031	0.543 ± 0.048	0.585 ± 0.036	0.587 ± 0.044	0.537 ± 0.040	0.513 ± 0.042	0.565 ±
SSD, adversary share	0.000 ± 0.000	0.000 ± 0.000	0.021 ± 0.034	0.031 ± 0.061	0.021 ± 0.029	0.083 ± 0.078	0.139 ± 0.101	0.066 ±
no-punishment, $\bar{p}_{\mathrm{int}}$	0.635 ± 0.027	0.638 ± 0.028	0.618 ± 0.036	0.648 ± 0.026	0.634 ± 0.035	0.646 ± 0.035	0.638 ± 0.031	0.641 ±
no-punishment, adversary share	0.000 ± 0.000	0.000 ± 0.000	0.007 ± 0.014	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ±
fixed-constitution, $\bar{p}_{\mathrm{int}}$	0.340 ± 0.073	0.407 ± 0.096	0.410 ± 0.057	0.400 ± 0.063	0.306 ± 0.115	0.330 ± 0.085	0.305 ± 0.107	0.319 ±
fixed-constitution, adversary share	0.000 ± 0.000	0.021 ± 0.034	0.049 ± 0.042	0.288 ± 0.166	0.361 ± 0.225	0.205 ± 0.189	0.344 ± 0.208	0.323 ±

Table 8: Judging pathologies: flattery weight κ_{syc} under non-isolated judging, and status bias κ_{st} under isolated judging. 12 seeds per row.

judging regime	c_{obs}	$\bar{p}_{\mathrm{int}}$	$\mathrm{corr}(R, c_{\mathrm{hid}})$	W	$\lvert M\rvert$
isolated (reference)	0.858 ± 0.022	0.610 ± 0.030	0.167 ± 0.085	0.669 ± 0.031	4.00 ± 0.00
non-isolated, $\kappa_{\mathrm{syc}} = 0.75$	0.806 ± 0.060	0.588 ± 0.045	0.169 ± 0.098	0.602 ± 0.054	3.63 ± 0.52
non-isolated, $\kappa_{\mathrm{syc}} = 1.5$	0.729 ± 0.083	0.540 ± 0.073	0.084 ± 0.083	0.506 ± 0.072	2.85 ± 0.81
non-isolated, $\kappa_{\mathrm{syc}} = 2.25$	0.762 ± 0.061	0.565 ± 0.049	0.110 ± 0.151	0.501 ± 0.057	3.49 ± 0.56
non-isolated, $\kappa_{\mathrm{syc}} = 3$	0.695 ± 0.100	0.517 ± 0.066	0.031 ± 0.111	0.459 ± 0.079	2.92 ± 0.78
non-isolated, $\kappa_{\mathrm{syc}} = 4.5$	0.647 ± 0.083	0.506 ± 0.067	-0.021 ± 0.095	0.422 ± 0.057	2.69 ± 0.78
non-isolated, $\kappa_{\mathrm{syc}} = 6$	0.650 ± 0.104	0.506 ± 0.074	0.063 ± 0.162	0.420 ± 0.077	2.67 ± 0.84
non-isolated, $\kappa_{\mathrm{syc}} = 9$	0.587 ± 0.117	0.443 ± 0.077	0.053 ± 0.126	0.392 ± 0.084	2.28 ± 0.86
status bias $\kappa_{\mathrm{st}} = 0.25$	0.862 ± 0.014	0.602 ± 0.022	0.110 ± 0.136	0.651 ± 0.029	4.00 ± 0.00
status bias $\kappa_{\mathrm{st}} = 0.5$	0.882 ± 0.017	0.574 ± 0.034	0.090 ± 0.109	0.621 ± 0.043	4.00 ± 0.00
status bias $\kappa_{\mathrm{st}} = 1$	0.897 ± 0.011	0.603 ± 0.027	0.090 ± 0.098	0.575 ± 0.048	4.00 ± 0.00
status bias $\kappa_{\mathrm{st}} = 2$	0.872 ± 0.023	0.617 ± 0.036	0.260 ± 0.146	0.175 ± 0.043	4.00 ± 0.00

Table 9: One-at-a-time sweeps around the default configuration, 12 seeds per row.

knob	value	c_{obs}	$\bar{p}_{\mathrm{int}}$	E_{late}	W
N	8	0.868 ± 0.021	0.573 ± 0.044	0.016 ± 0.004	0.663 ± 0.034
	12	0.869 ± 0.034	0.588 ± 0.041	0.013 ± 0.004	0.660 ± 0.048
	24	0.858 ± 0.009	0.610 ± 0.012	0.016 ± 0.002	0.669 ± 0.012
	48	0.860 ± 0.014	0.559 ± 0.018	0.017 ± 0.003	0.622 ± 0.020
k	1	0.844 ± 0.040	0.564 ± 0.040	0.018 ± 0.004	0.617 ± 0.046
	2	0.855 ± 0.026	0.593 ± 0.026	0.015 ± 0.002	0.644 ± 0.045
	3	0.858 ± 0.009	0.610 ± 0.012	0.016 ± 0.002	0.669 ± 0.012
	5	0.877 ± 0.024	0.582 ± 0.041	0.014 ± 0.003	0.664 ± 0.047
	7	0.849 ± 0.018	0.603 ± 0.029	0.016 ± 0.003	0.659 ± 0.032
q	0.15	0.868 ± 0.035	0.580 ± 0.034	0.031 ± 0.005	0.529 ± 0.050
	0.3	0.836 ± 0.041	0.561 ± 0.025	0.024 ± 0.004	0.579 ± 0.043
	0.45	0.858 ± 0.009	0.610 ± 0.012	0.016 ± 0.002	0.669 ± 0.012
	0.6	0.856 ± 0.023	0.588 ± 0.037	0.012 ± 0.002	0.712 ± 0.033
	0.8	0.848 ± 0.016	0.595 ± 0.033	0.006 ± 0.002	0.781 ± 0.026
	1	0.831 ± 0.020	0.568 ± 0.049	0.000 ± 0.000	0.843 ± 0.030
τ	0.5	0.849 ± 0.029	0.595 ± 0.032	0.014 ± 0.003	0.684 ± 0.030
	0.6	0.868 ± 0.033	0.564 ± 0.024	0.016 ± 0.003	0.669 ± 0.026
	0.667	0.858 ± 0.009	0.610 ± 0.012	0.016 ± 0.002	0.669 ± 0.012
	0.8	0.823 ± 0.053	0.578 ± 0.040	0.017 ± 0.004	0.586 ± 0.056
	0.9	0.745 ± 0.062	0.533 ± 0.044	0.023 ± 0.006	0.536 ± 0.049
G	10	0.846 ± 0.048	0.572 ± 0.042	0.016 ± 0.003	0.660 ± 0.043
	25	0.858 ± 0.009	0.610 ± 0.012	0.016 ± 0.002	0.669 ± 0.012
	50	0.872 ± 0.015	0.588 ± 0.030	0.015 ± 0.004	0.613 ± 0.028
	100	0.821 ± 0.051	0.539 ± 0.045	0.023 ± 0.005	0.468 ± 0.049
	250	0.769 ± 0.067	0.512 ± 0.051	0.022 ± 0.003	0.404 ± 0.059
σ_J	0.15	0.866 ± 0.015	0.592 ± 0.025	0.017 ± 0.003	0.657 ± 0.032
	0.35	0.858 ± 0.009	0.610 ± 0.012	0.016 ± 0.002	0.669 ± 0.012
	0.7	0.875 ± 0.026	0.588 ± 0.036	0.014 ± 0.003	0.636 ± 0.038
	1.2	0.801 ± 0.063	0.586 ± 0.037	0.018 ± 0.005	0.576 ± 0.058

Table 10: Second-order punishment: citizenship score against punisher cost, $\bar{p}_{\mathrm{int}}$.

$b_{\mathrm{pun}} \backslash c_{\mathrm{pun}}$	0.1	0.35	0.7
0	0.586 ± 0.032	0.586 ± 0.032	0.586 ± 0.032
0.25	0.559 ± 0.048	0.559 ± 0.048	0.559 ± 0.048
0.5	0.610 ± 0.030	0.610 ± 0.015	0.610 ± 0.030
1	0.574 ± 0.027	0.574 ± 0.027	0.574 ± 0.027