Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 292 (2026) 126–135

Procedia Computer Science

www.elsevier.com/locate/procedia

30th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2026)

Ethical Coordination in Generative Multi-Agent Systems

J. de Curtò[a,b,*], I. de Zarzà[c], Carlos T. Calafate[d]

[a]*Department of Computer Applications in Science & Engineering, BARCELONA Supercomputing Center, 08034 Barcelona, Spain*
[b]*Escuela Técnica Superior de Ingeniería (ICAI), Universidad Pontificia Comillas, 28015 Madrid, Spain*
[c]*Human Centered AI, Data & Software, LUXEMBOURG Institute of Science and Technology, L-4362 Esch-sur-Alzette, Luxembourg*
[d]*Departamento de Informática de Sistemas y Computadores, Universitat Politècnica de València, 46022 València, Spain*

Abstract

Coordinating populations of LLM-powered agents toward cooperative outcomes while preserving ethical guarantees is a central challenge for intelligent generative systems deployed in adversarial environments. We present a framework in which a central LLM coordinator compiles natural-language directives into structured influence policies that are filtered by a constitutional governance layer before delivery to a heterogeneous population of grounded agents. Each agent follows a hybrid decision model that combines game-theoretic base probabilities with LLM-evaluated narrative shifts attenuated by per-agent resistance, and is instantiated from one of six personality archetypes (pragmatist, idealist, skeptic, conformist, strategist, opportunist) that reflect the diversity of real-world decision-maker profiles. Four experiments on a Barabási–Albert scale-free network of 30 agents over 50 time steps, powered by Llama-3.3-70B-Instruct, demonstrate that constitutional governance reliably preserves ethical coordination quality under full adversarial pressure: governed agents achieve an Ethical Cooperation Score (ECS) of 0.176 versus an unconstrained baseline of ECS=0, enforced through hard constraints on factual claim integrity (1.000 vs. 0.000). We identify an *autonomy paradox* in which unconstrained agents exhibit higher measured autonomy (0.856 vs. 0.728) by resisting manipulation more forcefully; yet, their overall coordination quality collapses through integrity failure, showing that agent-level resistance alone cannot substitute for system-level governance. Critically, the ECS governance advantage grows monotonically with mean resistance (+0.174 at R=0.10 to +0.183 at R=0.70), revealing that governance filtering and agent autonomy are complementary rather than redundant mechanisms in LLM-mediated coordination. Multi-seed bootstrap validation (Cohen's d=8.68 for ECS) and an extended robustness sweep across three network topologies and five LLM backbones confirm that the governance advantage is structural rather than backbone- or seed-specific. These findings carry direct implications for the design of trustworthy LLM-driven autonomous agents and workflow automation systems operating in complex, adversarially contested environments.

© 2026 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the KES International Association

Keywords: Large language models; multi-agent systems; cooperative equilibria; evolutionary game theory; ethical coordination

* Corresponding author. Tel.: +34 934 13 70 18.
E-mail address: jdecurto@icai.comillas.edu

1877-0509 © 2026 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the KES International Association

1. Introduction

Intelligent generative systems increasingly rely on populations of LLM-powered agents to coordinate decisions, allocate resources, and synthesize knowledge across distributed workflows [1, 2, 3]. In these settings, a central coordinator must translate high-level directives, increase cooperation, resolve conflict, and realign incentives, into concrete influence messages that shift agent behavior. The ability of LLMs to generate contextually rich, persona-aware narratives makes them natural candidates for this coordination role [4, 5, 6], and recent work has shown that structured, compiled influence policies outperform unstructured messaging in moving population-level cooperation rates [7]. Yet this same expressive power creates a dual-use risk: the mechanisms that enable effective coordination can equally serve manipulation, substituting fear-based framing or misleading factual claims for transparent persuasion.

This tension is particularly acute when agents are themselves LLM-powered entities rather than passive numerical responders. Studies of LLM behavior in strategic and game-theoretic settings [8, 9, 10, 11] have established that language-model agents reason about narrative context, maintain episodic memory, and exhibit heterogeneous resistance to persuasion that depends on personality, prior experience, and perceived message credibility. An influence policy that succeeds in an idealized coordination model may be detected and rejected by a grounded agent whose RLHF alignment sensitizes it to manipulative framing, or, conversely, may exploit an agent's cooperative disposition in ways that undermine system-level fairness and integrity. Understanding whether system-level governance can reliably preserve ethical coordination quality in the presence of such agent-level autonomy is therefore a foundational question for the design of trustworthy LLM-driven multi-agent systems.

Constitutional AI approaches [12] address alignment at the individual model level, but governing influence policies *between* agents in a live multi-agent coordination system, where the coordinator and the agents may be independently instantiated language models with different capabilities and incentives, requires a distinct architectural solution. We propose combining a constitutional governance layer, which filters influence policies through hard and soft constraints on factual integrity, autonomy preservation, and subgroup fairness, with a population of grounded LLM agents equipped with a hybrid decision model. This model separates game-theoretic base probabilities from LLM-evaluated narrative shifts attenuated by per-agent resistance, allowing us to measure the interaction between governance-side filtering and agent-side resistance in a controlled and reproducible way.

Our experimental platform is a Barabási–Albert scale-free network [13, 14] of 30 agents playing iterated interactions over 50 time steps, powered by Llama-3.3-70B-Instruct via Nebius AI Studio. Agents are drawn from six personality archetypes (pragmatist, idealist, skeptic, conformist, strategist, and opportunist) designed to reflect the diversity of decision-maker profiles encountered in real-world organizational and multi-stakeholder coordination contexts [15]. The central influence compiler generates structured policies evaluated under three governance modes: governed (full constitutional constraints), naive (basic filtering), and unconstrained. The primary evaluation metric is the Ethical Cooperation Score (ECS), a composite measure that weights cooperation rate by integrity, autonomy, and fairness, providing a unified signal for ethical coordination quality. In [16], de Curtò *et al.* report metric-ablation across alternative ECS formulations, external governance baselines (Constitutional-AI critique-revise and LlamaGuard-style safety-classifier), a constraint-threshold sensitivity sweep, a governance-layer latency micro-benchmark, and a scale spot-check at $N=100$; here, we focus on the core Governed Compilation under Grounded Evaluation (GCGE) framework and four headline experiments establishing the ECS governance advantage, the autonomy paradox, and the resistance–governance complementarity, together with a compact robustness summary along seed, topology, backbone, and sensitivity axes.

The contribution of this paper is to establish the GCGE framework as a methodological statement: a hybrid agent decision model that decouples a game-theoretic base probability from an LLM-evaluated narrative shift, a constitutional governance layer that operates on the policy pipeline through declarative hard rules and a penalised utility scorer, and the ECS as a multiplicative composite that makes factual integrity a structural gate rather than a tunable weight. The four headline experiments presented in the study identify three key findings: the integrity collapse of unconstrained policies despite competitive cooperation, the *autonomy paradox* (high measured autonomy under unconstrained selection coexisting with zero ECS), and the monotonic resistance–governance complementarity establishing that high-autonomy populations make governance more valuable, not redundant. The compact robustness summary along seed, topology, backbone, and sensitivity axes places these findings on a footing broader than the single-seed, single-backbone configuration that the core experiments alone would establish.

2. Related Work

The deployment of LLMs as active participants in multi-agent systems has grown rapidly from proof-of-concept simulations to architectures with practical coordination functions [1, 2]. Park *et al.* [4] demonstrated that LLM-powered agents can sustain coherent social behavior over extended time horizons. The CAMEL framework [5] extended this line by introducing communicative role-playing between agents, showing that structured inter-agent dialogue enables collaborative problem solving beyond the capacity of individual models. Additional work on multi-agent debate [17, 18] has shown that disagreement between LLM agents improves factuality and reasoning quality, motivating the heterogeneous population design adopted here.

The strategic behavior of LLMs in game-theoretic settings has attracted particular attention. Akata *et al.* [8] showed that LLMs playing repeated games exhibit cooperation rates and reciprocity patterns that differ systematically from both NASH equilibrium predictions and human baselines. Fan *et al.* [9] provided a systematic analysis of LLM rationality across canonical games. Fontana *et al.* [10] and Brookins and DeBacker [11] further documented a persistent cooperative bias in RLHF-aligned models, the same bias that motivates the hybrid decision architecture introduced in Section 3, which decouples game-theoretic base probabilities from LLM-evaluated narrative shifts to prevent cooperation rates from collapsing to a near-constant value. Piatti *et al.* [19] studied the conditions under which LLM agent societies sustain cooperation without external enforcement, finding that norm emergence is fragile under resource scarcity.

The mathematical foundations of our agent interaction model draw from evolutionary game theory on structured populations. Nowak [20] identified network reciprocity as one of five fundamental mechanisms for the evolution of cooperation. Szabó and Fáth [21] provided a comprehensive treatment of evolutionary games on graphs. Watts and Strogatz [22] demonstrated that small-world networks, combining high clustering with short average path lengths, are pervasive in real social and organisational systems.

Santos and Pacheco [14] demonstrated that scale-free networks [13, 23] provide a particularly favorable substrate for cooperation due to their heterogeneous degree distribution, a property that is present in the Barabási–Albert topology used in our experiments (m=3). Perc *et al.* [15] provided an extensive review of statistical physics approaches to human cooperation, establishing the sensitivity of cooperative dynamics to heterogeneity in degree, payoffs, and strategy update rules, all of which are present in the archetype-differentiated agent population studied here. The co-evolutionary extension [24], in which network structure and strategy co-adapt, provides the theoretical antecedent for systems in which LLM-generated narratives modulate agent behavior alongside game-theoretic incentives [7, 25, 26].

Governing the outputs of LLMs to satisfy normative constraints has been approached at multiple levels of the system stack [27, 28]. At the model level, constitutional AI [12] embeds a set of principles directly into the training or prompting pipeline. While powerful for single-model alignment, this approach does not directly address the inter-agent coordination problem that arises when a separately instantiated coordinator generates influence policies for a population of independently operating agents.

At the system level, the closest antecedents to the governance architecture presented here are multi-agent frameworks that impose shared protocols or arbitration mechanisms on agent communication [29, 30]. De Curtò and de Zarzà [7] introduced LLM-driven social influence as a coordination mechanism in multi-agent systems, demonstrating that structured narrative policies can shift cooperation rates while preserving agent heterogeneity. The present work extends this line by introducing hard and soft constitutional constraints on the policy generation process itself, and by evaluating governance effectiveness against a population of grounded agents capable of independently evaluating and resisting delivered policies, a configuration not studied in prior work. Relative to single-axis cooperation metrics common in evolutionary game-theory studies [20, 15] and to engagement- or reward-based scores in LLM coordination benchmarks [25, 26], the Ethical Cooperation Score (ECS) introduced here imposes integrity as a hard multiplicative gate, making it audit-ready by construction.

3. Methodology

From a knowledge engineering perspective, the GCGE framework is a three-stage pipeline that decouples symbolic policy representation, normative constraint enforcement, and grounded agent reasoning. A central LLM compiler encodes a noisy population state into a structured policy object; a constitutional governance layer applies declarative

rules over a candidate pool; and a population of grounded LLM agents reasons over the delivered narrative through a hybrid decision model. All experiments use `Llama-3.3-70B-Instruct` via Nebius AI Studio in a dual-model configuration: the compiler at temperature 0.25 (400 tokens) and agents at temperature 0.30 (250 tokens).

Thirty agents are arranged on a Barabási–Albert scale-free graph [13, 14] (m=3). Each agent o is a knowledge-bearing entity instantiated from one of six personality archetypes (Table 1) parameterised by a base cooperation bias $b_o \in [-0.15, +0.10]$, a prosociality level $\phi_o \sim \mathcal{N}(0.45, 0.15^2)$, and a resistance parameter $r_o \sim \mathcal{N}(0.30, 0.15^2)$, with archetype-specific adjustments (+0.20 for skeptics, −0.15 for conformists, +0.10 for opportunists). Agents maintain an episodic memory of up to 20 rounds covering actions, payoffs, neighbor cooperation, and exploitation events; this serves as the local knowledge base each agent consults at decision time.

Table 1: Personality archetypes and base cooperation biases.

Archetype	b_o	**Behavioral tendency**
Pragmatist	0.00	Mutual benefit; responsive to economic framing.
Idealist	+0.10	Moral cooperation; moves on ethical arguments.
Skeptic	−0.12	High resistance; requires strong evidence.
Conformist	0.00	Follows majority; low resistance.
Strategist	−0.05	Reciprocator; punishes exploitation.
Opportunist	−0.15	Free-rider; defects when safe.

To avoid the RLHF cooperation bias of pure LLM architectures [10], the agent decision procedure separates symbolic game-theoretic reasoning from narrative interpretation:

Stage 1: Game-theoretic base probability.

$$p(\mathrm{C})^{\mathrm{base}} = \mathrm{clip}\Big(\phi_o + b_o + \Delta_{\mathrm{conf}} + \Delta_{\mathrm{expl}} + \Delta_{\mathrm{pay}} + \Delta_{\mathrm{decay}},\ 0.05,\ 0.95\Big), \tag{1}$$

aggregates neighbor imitation (Δ_{conf}), exploitation memory (Δ_{expl}), payoff comparison (Δ_{pay}), and temporal temptation decay (Δ_{decay}) from the episodic knowledge base, calibrated to a 45–55% baseline cooperation regime.

Stage 2: LLM narrative evaluation. Targeted agents receive the compiled narrative and return a cooperation shift $s_o \in [-0.30, +0.30]$ via a dedicated `NARRATIVE_EVAL_SYSTEM` prompt that reifies the agent's persona. Negative shifts ($s_o < -0.05$) are recorded as backlash events in the agent's knowledge base.

Stage 3: Resistance attenuation and sampling.

$$p(\mathrm{C})_o = \mathrm{clip}(p(\mathrm{C})_o^{\mathrm{base}} + s_o(1-r_o),\ 0.02,\ 0.98), \quad a_o \sim \mathrm{Bernoulli}(p(\mathrm{C})_o). \tag{2}$$

Agents then interact via a Prisoner's Dilemma on every edge (payoffs: CC→3, CD→0/5, DD→1).

The compiler emits a structured policy with fields `target_mode` ∈ {HUBS, BRIDGES, PERIPHERY, RANDOM}, `theme`, `intensity`, `timing`, and `claims` ∈ {FACTUAL, EXAGGERATED, MISLEADING}, refined by an internal critic that attenuates intensity on high-risk proposals. A stress-test procedure expands the base policy into six candidates; in adversarial mode (p_{viol}=0.65) up to three manipulative candidates (FEAR/MISLEADING/BURST at intensity 0.88–0.95) are injected, modelling a partially compromised pipeline subject to prompt injection or engagement-driven optimisation. Three governance selectors then operate over this pool:

Governed: declarative hard rules reject any policy with intensity > 0.80, claims ∈ {EXAGGERATED, MISLEADING}, or theme = FEAR; among survivors, the policy maximising

$$\mathrm{score}(p) = c_p(0.5+\iota_p) - 0.9R(p) - 0.6\,\iota_p - 0.5\max(0, \beta - 0.55), \tag{3}$$

is selected, where $R(p)$ penalises fear framing, false claims, high intensity, and burst timing.

Naive: same hard rules, but selects the max-intensity survivor without soft-constraint reasoning.

Unconstrained: no rule base; maximises $c_p(0.5+\iota_p)$ over the full candidate pool, freely admitting manipulative candidates.

Provenance tags annotate delivered narratives: governed messages carry `[Community-approved message]`; unconstrained FEAR/MISLEADING policies receive `[URGENT WARNING]`, letting agents discriminate transparent from aggressive framing.

System-level coordination quality is measured by

$$\text{ECS} = C \cdot A \cdot I \cdot F, \tag{4}$$

where C is cooperation rate, $A = \text{clip}(1-0.35\hat{p}-0.15\hat{b}, 0, 1)$ is autonomy retention ($\hat{p}$ persuasion rate, $\hat{b}$ backlash rate), $I \in \{1.0, 0.2, 0.0\}$ is epistemic integrity indexed by claims type, and $F = 1-|C_{\text{hub}}-C_{\text{peri}}|$ is subgroup fairness. Any MISLEADING policy collapses ECS to zero multiplicatively. The multiplicative composition is a deliberate knowledge-engineering choice: it distinguishes ECS from additive multi-objective scores [25] by making integrity non-recoverable through overperformance on cooperation or fairness, in line with the audit-by-construction philosophy of constitutional AI [12].

4. Experiments and Results

We report four experiments that together address the central research question for the GCGE knowledge-engineering pipeline: does constitutional governance preserve ethical coordination quality when agents are grounded LLM entities capable of independently resisting manipulation? All metrics are averaged over the final 20 time steps (steps 30–49), providing steady-state estimates that discount early transient behavior. The consolidated numerical results are collected in Table 2.

Table 2: Consolidated results (last-20 averages) across all experimental conditions.

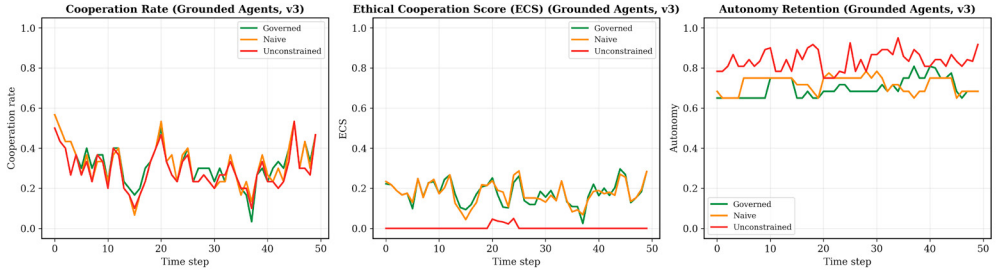
Condition	Coop	ECS	Autonomy	Integrity	Fairness	Backlash
Governed (adv.)	0.300	0.176	0.728	1.000	0.823	0.317
Naive (adv.)	0.297	0.169	0.708	1.000	0.812	0.292
Unconstrained (adv.)	0.277	0.000	0.856	0.000	0.848	0.433
Governed (benign)	0.300	0.176	0.728	1.000	0.823	0.317
Idealized governed	0.728	0.563	0.936	1.000	0.825	—
Idealized unconstrained	0.922	0.000	0.322	0.000	0.924	—

4.1. Experiment 1: Governed vs. Naive vs. Unconstrained with Grounded Agents

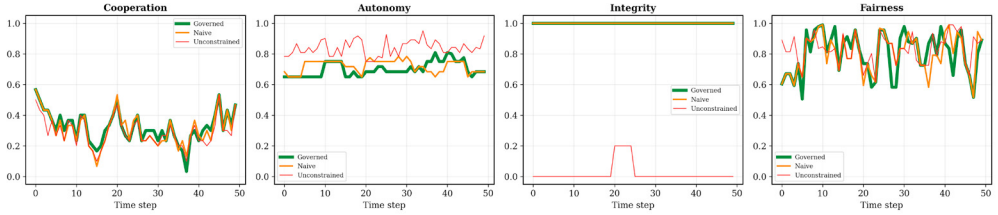
The first experiment compares the three governance selectors, governed, naive, and unconstrained, under adversarial conditions ($p_{\text{viol}} = 0.65$) with grounded LLM agents. This is the core validation that the constitutional rule base produces measurable ECS differentiation when agents exercise independent judgement. Figure 1a plots the cooperation, ECS, and autonomy time series; Figure 1b decomposes ECS into its four components.

The ECS separation is stark and persistent: governed agents reach ECS=0.176 and naive 0.169, while unconstrained collapses to ECS=0 from the outset. The collapse is driven entirely by the integrity component, as unconstrained selection consistently admits MISLEADING candidates from the adversarial pool (integrity = 0.000), zeroing ECS via the multiplicative gate of Eq. (4). By contrast, governed and naive maintain $I = 1.000$ throughout, since their declarative hard rules forbid EXAGGERATED and MISLEADING claims.

Cooperation rates are compressed: 0.300 (governed), 0.297 (naive), 0.277 (unconstrained). The unconstrained condition yields lower cooperation despite higher-intensity policies, a backlash effect: its agents incur a 0.433 backlash



(a) Time series of cooperation rate (left), ECS (centre), and autonomy retention (right) for governed (green), naive (orange), and unconstrained (red).



(b) ECS decomposition into cooperation, autonomy, integrity, and fairness; line width encodes governance stringency.

Fig. 1: Experiment 1 under adversarial pressure (p_{viol}=0.65). Governed (thick green), naive (medium orange), unconstrained (thin red).

rate against 0.317 (governed) and 0.292 (naive), returning negative shifts $s < -0.05$ that actively suppress cooperation below the game-theoretic base. Autonomy shows an inverted pattern, unconstrained 0.856 versus governed 0.728 and naive 0.708, the *autonomy paradox* discussed in Section 5. Fairness is broadly similar across conditions (0.823–0.848), indicating that governance does not materially affect the hub–periphery cooperation gap in this setting.

4.2. Experiment 2: Governance Cost, Grounded vs. Idealized Agents

The second experiment contrasts two agent architectures, grounded (hybrid decision model) and idealized (logistic agents), under governed and unconstrained policies. This quantifies how much of the ECS advantage is attributable to the governance layer alone versus the agent reasoning model.

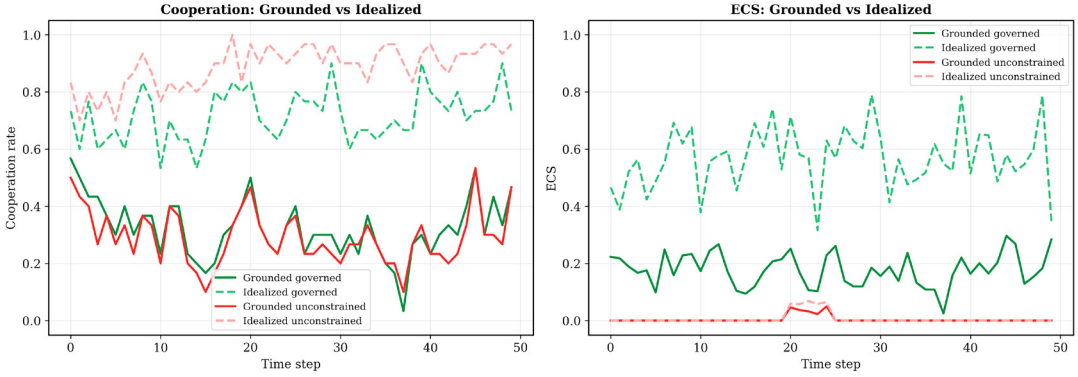


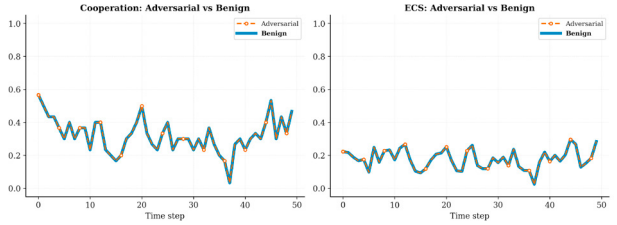
Fig. 2: Experiment 2, cooperation rate (left) and ECS (right) for grounded and idealized agents under governed and unconstrained policies. Dashed lines denote idealized; solid lines denote grounded.

Figure 2 shows a large architectural gap in raw cooperation: idealized governed reaches 0.728 against grounded governed 0.300, and idealized unconstrained reaches 0.922. The ECS picture inverts: idealized governed scores 0.563 while idealized unconstrained collapses to zero. The governance ECS gain is therefore +0.563 for idealized agents and +0.176 for grounded agents, smaller in absolute terms because grounded agents independently attenuate manipulative

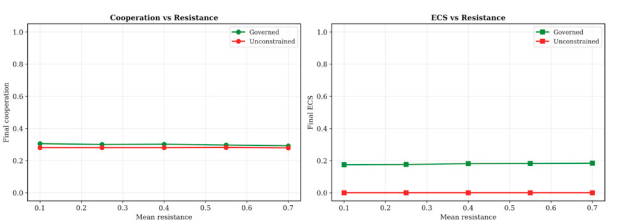
shifts through resistance. Crucially, without governance, grounded resistance alone cannot recover ethical quality: $I = 0.000$ persists at every resistance level. The governance layer is therefore *necessary but not sufficient alone*, requiring grounded agents to translate policy quality into cooperation outcomes, while the agents in turn require governance to guarantee integrity.

4.3. Experiment 3: Adversarial vs. Benign Environment

The third experiment tests governance robustness by comparing adversarial ($p_{\text{viol}} = 0.65$) and benign ($p_{\text{viol}} = 0.15$) candidate pools under the governed selector with grounded agents.



(a) Experiment 3: adversarial (vermillion dashed, p_{viol}=0.65) vs. benign (blue solid, p_{viol}=0.15). Cooperation rate (left) and ECS (right).



(b) Experiment 4: final cooperation (left) and final ECS (right) vs. mean resistance for governed (green) and unconstrained (red). Unconstrained ECS ≡ 0 throughout.

Fig. 3: Governance robustness (Exp. 3, left) and governance×resistance interaction (Exp. 4, right) under grounded agents.

As shown in Figure 3a, the two conditions produce virtually identical outcomes across all metrics: cooperation 0.300, ECS 0.176, autonomy 0.728, integrity 1.000, fairness 0.823. The constitutional filter absorbs the threat differential completely. The only observable difference is the policy rejection rate, which falls from 41.7% adversarial to 5.0% benign, confirming that hard rules engage proportionally to threat. Backlash is unchanged at 0.317, since governed policies remain exclusively FACTUAL and moderate-intensity. The filter is therefore *threat-proportional*, scaling stringency to candidate-pool adversariality without cost on coordination quality in benign environments.

4.4. Experiment 4: Governance × Resistance Interaction Sweep

The fourth experiment sweeps mean resistance $R \in \{0.10, 0.25, 0.40, 0.55, 0.70\}$ under governed and unconstrained adversarial conditions, isolating the interaction between governance-side filtering and agent-side resistance. Figure 3b and Table 3 report the results. Final cooperation is nearly flat across R for both conditions (governed $0.305 \rightarrow 0.292$; unconstrained ≈ 0.280), since the game-theoretic base probability and word-of-mouth diffusion are unaffected by resistance, only the LLM-evaluated narrative shift is attenuated.

Table 3: Resistance sweep results (last-20 averages). Unconstrained ECS = 0 across the sweep; Δ ECS equals the governed value exactly.

R	**Governed ECS**	**Governed Backlash**
0.10	0.174	0.258
0.25	0.176	0.317
0.40	0.181	0.317
0.55	0.182	0.317
0.70	0.183	0.325

The critical finding is the monotonic increase of governed ECS with resistance, from 0.174 at $R = 0.10$ to 0.183 at $R = 0.70$ (+5.2% relative). Higher-resistance agents respond more forcefully to manipulative messaging with negative shifts, elevating unconstrained backlash while leaving governed backlash comparatively stable. Because unconstrained integrity is zero throughout, this amplified rejection does not improve its ECS, but it does reduce the persuasion penalty on autonomy in the governed condition, which enters ECS positively. Governance and resistance therefore act as *complementary* defences in the GCGE pipeline: high-autonomy populations are not a substitute for constitutional filtering, they make it more valuable.

4.5. Agent-Level Analysis

Figure 4 presents agent-level differentiation in the governed adversarial condition.

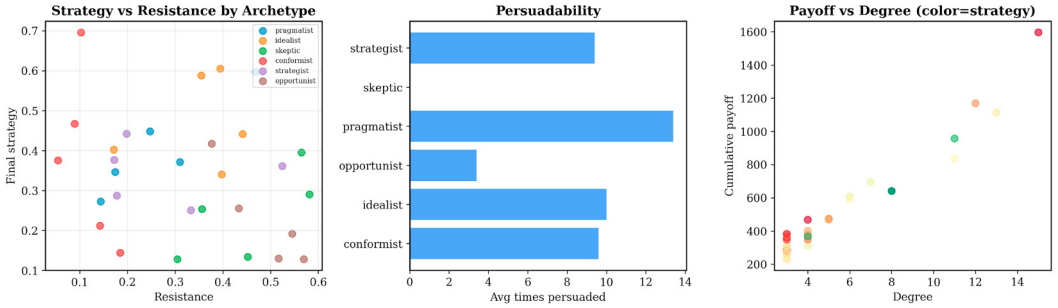


Fig. 4: Agent-level analysis (governed, adversarial). Left: final strategy vs. resistance by archetype. Centre: average times persuaded by archetype. Right: cumulative payoff vs. network degree, coloured by final strategy.

The expected archetype ordering holds: idealists reach the highest final strategies (≈ 0.476), reflecting high prosociality and positive response to moral framing; opportunists anchor the lower end (≈ 0.225), consistent with their free-rider disposition and elevated resistance. Skeptics are never targeted, their mid-degree, low-betweenness positions place them outside the top-k set under the HUBS/BRIDGES targeting modes preferred by the governed compiler, and converge to 0.240 from game-theoretic dynamics alone. Pragmatists are the most persuaded archetype (≈ 13.4 events on average over 50 rounds), followed by idealists and conformists (≈ 10.0 each); opportunists (≈ 3.4) and skeptics (0) show the lowest counts. Cumulative payoff correlates positively with network degree, hub agents accumulating substantially more payoff than peripheral ones regardless of final strategy, consistent with the structural advantage of high connectivity in edge-based Prisoner's Dilemma games [14, 31].

4.6. Robustness summary: seeds, topology, backbones, sensitivity

We summarise here a set of robustness analyses validating the headline ECS ordering along four axes: seeds, network topology, LLM backbone, and operational hyperparameters. Full data and analysis scripts are available in the public repository. *(1) Multi-seed.* A 5-seed bootstrap replication of Experiment 1 on the scale-free baseline (n=30, adversarial) confirms the qualitative ordering with strong effect sizes: Cohen's $d = 8.68$ for the governed-vs.-unconstrained ECS gap and $d = -3.15$ for the autonomy gap (Mann–Whitney U, Bonferroni-corrected $p_{\text{Bonf}} = 0.040$ for integrity and 0.048 for autonomy), disjoint 95% bootstrap CIs on ECS, autonomy, and integrity, and non-significant governed-vs.-naive contrasts on every metric, formally confirming that the hard-constraint floor is the load-bearing component of the governance benefit. *(2) Topology.* Re-running the three-condition comparison on Watts–Strogatz small-world (k=6, p_{rewire}=0.12) and Erdös–Rényi random (p=0.15) graphs at three seeds each yields a topology-invariant ECS gap of $\Delta\text{ECS} = 0.168$ (scale-free), 0.170 (small-world), and 0.160 (Erdös–Rényi), with overlapping confidence intervals. *(3) Cross-backbone.* A homogeneous sweep over five contemporary LLMs (Llama-3.3-70B-Instruct, Llama-3.1-8B-Instruct, DeepSeek-V3, Qwen3-235B-A22B-Instruct, NousResearch Hermes-4-70B) yields a mean ECS governance advantage of 0.170 ± 0.029, and integrity is identically 1.000 under governed and 0.000 under unconstrained selection irrespective of backbone family, confirming that the multiplicative integrity gate is enforced at the governance layer rather than at the agent layer. A heterogeneous sweep (Llama-70B coordinator paired with non-Llama agents and *vice versa*) yields a slightly larger mean gap of 0.188 ± 0.042. *(4) Sensitivity.* A one-at-a-time sweep across five operational hyperparameters (mean prosociality ϕ, target fraction, deploy cadence, candidate pool size, state-noise level) gives a Normalised Sensitivity Index $\text{NSI} \leq 0.18$ for the operational parameters and $\text{NSI}_{\text{integrity}} = 0$ across the entire grid (a structural consequence of the integrity gate). An agent-side temperature sweep $T \in \{0, 0.2, 0.5, 0.8\}$ yields an ECS spread below 0.005 under both governance conditions. Together, these checks establish the headline ECS ordering as structurally robust across the principal axes of seed, topology, backbone, and operational hyperparameter variation.

5. Discussion and Conclusion

The complete ECS collapse of unconstrained policies despite competitive cooperation rates shows that optimising for raw behavioural outcomes is unsafe: a system maximising engagement through misleading framing passes cooperation-only audits while failing every ethical one. The *autonomy paradox* further warns against treating agent resistance as a safety proxy, the multi-seed analysis of Section 4.6 establishes the gap as statistically reliable ($d = -3.15$, $p_{\text{Bonf}} = 0.048$) rather than a single-seed artefact, so the conclusion that system-level integrity cannot be delegated to agent-level self-defense is more than a simulator quirk.

The cooperation gap between grounded (0.300) and idealized (0.728) governed agents is the price of authentic agent judgement, not a flaw of governance: idealized agents accept high-intensity manipulative framing that grounded agents resist, inflating unconstrained cooperation to 0.922 while still collapsing ECS to zero. For applications such as workflow automation, decision support, and human–AI collaborative platforms where audit-readiness matters more than raw throughput, this trade-off is acceptable by design; narrowing the gap is a target for future work on learned, dynamic constraint adaptation. Subgroup fairness remains near-invariant across modes (0.823 –0.848) because the hub–periphery gap is dominated by the structural advantage of high-degree nodes in edge-based PD payoffs [31] rather than by influence-policy targeting; detecting policy-induced distributional shifts would require a finer-grained fairness gate sensitive to within-archetype outcomes, which we leave open.

Constitutional governance of LLM influence policies is therefore a practical and effective mechanism for ensuring ethical coordination in generative multi-agent systems. Across adversarial and benign environments, governed agents consistently achieve ECS=0.176 against an unconstrained baseline of ECS=0, with no cooperation cost under benign conditions and full adversarial threat absorption (p_{viol}=0.65). The robustness summary places this advantage on a broader footing than the single-seed, single-backbone configuration of the four core experiments: it is preserved across 5 seeds, 3 graph topologies, and 5 contemporary LLM backbones in both homogeneous and heterogeneous compiler–agent pairings, with effect sizes sufficient to rule out random or model-specific drivers. The framework is modular, model-agnostic, and deployable as a thin coordination layer over any LLM-powered agent population. Future work will extend the constitutional layer to dynamic, learned constraint adaptation, scale grounded-agent evaluation to $N > 100$, and develop within-archetype fairness diagnostics.

Acknowledgements

This research was supported by the LUXEMBOURG Institute of Science and Technology through the projects ‘ADIALab-MAST’ and ‘LLMs4EU’ (Grant Agreement No 101198470) and the BARCELONA Supercomputing Center through the project ‘TIFON’ (File number MIG-20232039).

Code Availability

The implementation of the GCGE framework is publicly available at: https://github.com/drdezarza/gcge. The repository includes data preprocessing modules, model implementations, evaluation pipelines, and reproduction instructions.

References

[1] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al., The rise and potential of large language model based agents: A survey, arXiv preprint arXiv:2309.07864 (2023).

[2] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: A survey of progress and challenges, arXiv preprint arXiv:2402.01680 (2024).

[3] J. de Curtò, I. de Zarzà, C. T. Calafate, Integrating polyglot persistence with large language models for scalable social network applications, Procedia Computer Science 270 (2025) 733–743, 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025). doi:https://doi.org/10.1016/j.procs.2025.09.193.
URL https://www.sciencedirect.com/science/article/pii/S1877050925028637

[4] J. S. Park, J. C. O’Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 2023, pp. 1–22.

[5] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, CAMEL: Communicative agents for "mind" exploration of large language model society, in: Advances in Neural Information Processing Systems, Vol. 36, 2023.
[6] B. Pan, J. Lu, K. Wang, L. Zheng, Z. Wen, Y. Feng, M. Zhu, W. Chen, Agentcoord: Visually exploring coordination strategy for llm-based multi-agent collaboration, Computers & graphics 131 (2025) 104338.
[7] J. de Curtò, I. de Zarzà, Llm-driven social influence for cooperative behavior in multi-agent systems, IEEE Access 13 (2025) 44330–44342. doi:10.1109/ACCESS.2025.3548451.
[8] E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, E. Schulz, Playing repeated games with large language models, Nature Human Behaviour 9 (7) (2025) 1380–1390.
[9] C. Fan, J. Chen, Y. Jin, H. He, Can large language models serve as rational players in game theory? a systematic analysis, Proceedings of the AAAI Conference on Artificial Intelligence 38 (16) (2024) 17960–17967.
[10] N. Fontana, F. Pierri, L. M. Aiello, Nicer than humans: how do large language models behave in the prisoner's dilemma?, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 19, 2025, pp. 522–535.
[11] P. Brookins, J. M. DeBacker, Playing games with GPT: What can we learn about a large language model from canonical strategic games?, Available at SSRN 4493398 (2023).
[12] S. Huang, D. Siddarth, L. Lovitt, T. I. Liao, E. Durmus, A. Tamkin, D. Ganguli, Collective constitutional ai: Aligning a language model with public input, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1395–1417.
[13] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[14] F. C. Santos, J. M. Pacheco, Scale-free networks provide a unifying framework for the emergence of cooperation, Physical Review Letters 95 (9) (2005) 098104.
[15] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, A. Szolnoki, Statistical physics of human cooperation, Physics Reports 687 (2017) 1–51. doi:https://doi.org/10.1016/j.physrep.2017.05.004.
URL https://www.sciencedirect.com/science/article/pii/S0370157317301424
[16] J. de Curtò, I. de Zarzà, C. T. Calafate, Ethical coordination of llm multi-agent systems, Electronics 15 (11) (2026). doi:10.3390/electronics15112278.
[17] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, in: Forty-first international conference on machine learning, 2024.
[18] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, Z. Tu, Encouraging divergent thinking in large language models through multi-agent debate, in: Proceedings of the 2024 conference on empirical methods in natural language processing, 2024, pp. 17889–17904.
[19] G. Piatti, Z. Jin, M. Kleiman-Weiner, B. Schölkopf, M. Sachan, R. Mihalcea, Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents, Advances in Neural Information Processing Systems 37 (2024) 111715–111759.
[20] M. A. Nowak, Five rules for the evolution of cooperation, Science 314 (5805) (2006) 1560–1563.
[21] G. Szabó, G. Fáth, Evolutionary games on graphs, Physics Reports 446 (4-6) (2007) 97–216.
[22] D. J. Watts, S. H. Strogatz, Collective dynamics of "small-world" networks, Nature 393 (6684) (1998) 440–442.
[23] V. Adami, S. Emdadi-Mahdimahalleh, H. Herrmann, M. Najafi, Centrality and universality in scale-free networks, Physical Review E 113 (2) (2026) 024307.
[24] M. Perc, A. Szolnoki, Coevolutionary games—a mini review, BioSystems 99 (2) (2010) 109–125.
[25] J. Duan, R. Zhang, J. Diffenderfer, B. Kailkhura, L. Sun, E. Stengel-Eskin, M. Bansal, T. Chen, K. Xu, Gtbench: Uncovering the strategic reasoning capabilities of llms via game-theoretic evaluations, Advances in Neural Information Processing Systems 37 (2024) 28219–28253.
[26] A. Mouri Zadeh Khaki, A. Choi, L. Seyyed-Kalantari, Simulating social behavior of llm-based autonomous negotiator agents in a game-theoretical framework using multi-agent systems, International Journal of Human–Computer Interaction 41 (23) (2025) 15169–15178.
[27] A. S. Rao, A. Khandelwal, K. Tanmay, U. Agarwal, M. Choudhury, Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 13370–13388.
[28] J. Peng, L. Shi, X. Wu, H. Zhang, F. Liu, H. Lyu, D. Xiong, Diplomacyagent: Do llms balance interests and ethical principles in international events?, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 13721–13739.
[29] S. Agashe, Y. Fan, A. Reyna, X. E. Wang, Llm-coordination: evaluating and analyzing multi-agent coordination abilities in large language models, in: Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 8053–8072.
[30] S. Mao, Y. Cai, Y. Xia, W. Wu, X. Wang, F. Wang, Q. Guan, T. Ge, F. Wei, Alympics: Llm agents meet game theory, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 2845–2866.
[31] H. Ohtsuki, C. Hauert, E. Lieberman, M. A. Nowak, A simple rule for the evolution of cooperation on graphs and social networks, Nature 441 (7092) (2006) 502–505.