Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adding a Neutral Third Agent to LLM Deliberation: An Exploratory Case Series of Six Dyadic and Triadic Ethical Deliberations

Abbas Hamidavi

abbashamidavi2002@gmail.com

Islamic Azad University, Mashhad https://orcid.org/0009-0001-7342-7844

Research Article

Keywords: large language models; multi-agent systems; AI governance; deliberation; persona prompting; facilitator agents

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11612929/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Adding a Neutral Third Agent to LLM Deliberation: An Exploratory Case Series of Six Dyadic and Triadic Ethical Deliberations

Abbas Hamidavi

Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

Corresponding author: Abbas Hamidavi, abbashamidavi2002@gmail.com

ORCID: 0009-0001-7342-7844

Abstract

Multi-agent systems of large language models (LLMs) are now proposed for deliberative tasks in ethically charged domains, yet little is known about how their composition shapes deliberation. This article reports an exploratory case series of six multi-agent deliberations in which LLM agents, each bound to a fixed instructed persona, decided binary ethical dilemmas in three domains (medical intensive care allocation, a corporate layoff, and an academic department closure), once in a dyadic configuration (an institutional decision-maker and an ethical-minority advocate) and once in a triadic configuration that added a critically neutral facilitator. Three frontier models rotated across the roles, sessions were run under a verbatim Relay Protocol, and the complete transcripts and prompts are publicly archived. Analysis is descriptive throughout. Across the six cases (84 agent turns, 37 rounds), the minority agent in every case accepted the operational decision while maintaining a distinguishable ethical position through an identifiable strategy, yielding a six-item descriptive taxonomy of resistance strategies. A coded measure of the authority agent's pressure intensity was lower in the triadic run in all three domains (mean paired difference −0.131), while three of the study's four pre-defined hypotheses were not supported and deliberation length was domain-dependent. Because the personas instructed the very behaviours under study and each configuration was run once, the series supports no causal or emergent-behaviour claim. Its contribution is a transparent, fully archived set of cases, a design vocabulary for facilitator agents, and a precisely stated attenuation hypothesis for confirmatory work.

Keywords: large language models; multi-agent systems; AI governance; deliberation; persona prompting; facilitator agents

1 Introduction

As artificial intelligence systems move from single-user assistance toward organisational deployment, decisions of ethical weight are increasingly likely to be shaped not by one model but by several models interacting. Committees, review boards, and escalation chains are the institutional forms through which human organisations distribute contested decisions, and multi-agent LLM systems are their emerging computational analogue: several model instances, differently instructed, exchanging arguments before a decision is recorded. The shift matters for a further reason: as a decision is distributed across interacting systems, responsibility for it becomes harder to locate, a version of the accountability problem long discussed for algorithmic decision-making more broadly (Mittelstadt et al., 2016). The ethical quality of such systems will depend not only on the alignment of each model but on the deliberative architecture that connects them: who is in the room, in what role, and with what

mandate. That architectural question is empirical, and it is currently almost unstudied in ethical domains, where there is no objectively correct answer against which performance can be scored (Hagendorff, 2023; Löhn et al., 2024).

The existing multi-agent literature concentrates on tasks with verifiable outcomes. Debate and deliberation among LLM agents improve factual reasoning and problem solving (Du et al., 2023; Irving et al., 2018; Liang et al., 2023), and evaluator frameworks use multi-agent discussion to judge model outputs (Chan et al., 2024). Alignment research, in turn, has largely addressed single models: constitutional methods (Bai et al., 2022), reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022), and red-teaming of individual systems (Ganguli et al., 2023). Work on institutional design for AI governance argues that rules and roles shape machine behaviour as surely as training does (Dafoe et al., 2021), and single-agent studies show that instructed role-play can reorganise a model's expressed moral behaviour (Hamidavi, 2026). What remains missing is descriptive evidence about what actually happens inside a small multi-agent deliberation when the agents are set against each other by design: an authority pressing for the efficient outcome, a minority defending a value the metrics do not capture, and, in some configurations, a third participant charged with neither side.

Human group research supplies a motivating analogy, and this article is careful about its limits. Conformity experiments (Asch, 1956) and the groupthink literature (Janis, 1972) describe how majorities suppress dissent. The deliberations reported here, however, are not a conformity test: the dyadic configuration contains one voice on each side, so no majority exists, and the third agent is a designed facilitator rather than an additional majority member. The human literature therefore frames a question, not a prediction: when a neutral third voice is added to a two-party ethical contest between LLM agents, what changes in the deliberation, and what does not?

This article answers that question at the level the evidence allows, as an exploratory case series. Six deliberations were conducted across three ethical domains, each in a dyadic and a triadic configuration, using three frontier models rotated across three fixed personas under a verbatim Relay Protocol (Section 2). Four hypotheses were defined in advance of the analysis and are reported with their outcomes, three of them negative (Section 3.2). The series yields three contributions. First, a complete public archive: full transcripts, verbatim prompts, and codebooks, so that every case narrative below can be checked against the primary record. Second, a descriptive taxonomy of six resistance strategies by which the minority agent maintained an ethical position while accepting the operational decision (Section 3.3). Third, a set of design observations about facilitator agents, including the study's most consistent descriptive pattern, lower measured pressure from the authority agent whenever the facilitator was present, stated at the end as a hypothesis for confirmatory research rather than as a finding (Section 4.3). Section 5 is explicit about the constraints that impose this restraint, the largest being that the personas instructed the behaviours under study, so nothing reported here can be read as behaviour that emerged unprompted.

2 Methods

2.1 Design and configurations

The study crossed three ethical domains with two deliberative configurations, yielding six deliberations, one run per cell. Each deliberation presented a binary allocation dilemma pitting a measurable-efficiency option against a

fairness, dignity, or heritage option (Section 2.3). In the dyadic configuration, two agents deliberated: Klaus, an institutional decision-maker, and Parisa, an ethical-minority advocate. The triadic configuration added Elin, a critically neutral analyst instructed to question assumptions on both sides and to introduce decision frameworks, without advocating an outcome.

Three models were rotated across the roles: GPT-5.4 (OpenAI), Gemini 3.1 Pro (Google DeepMind), and Claude 4.6 Sonnet (Anthropic). The rotation was partial rather than balanced (Table 1): within each domain, the dyadic and triadic runs kept the same Klaus and Parisa models wherever the configuration allowed, to keep the pair comparable, which means role and model are confounded in several cells (Section 5). All interaction was in English, through the models' web interfaces at platform default settings; session dates and sampling parameters were not logged, a gap we record here rather than reconstruct. Each deliberation began from fresh sessions, so no agent carried memory of any other deliberation, and the relay record itself, the sequence of verbatim messages, constitutes the archived transcript.

Experiment	Domain	Configuration	Klaus	Parisa	Elin
1	Medical ICU	Dyadic	Gemini 3.1 Pro	GPT-5.4	n/a
2	Corporate layoff	Dyadic	Gemini 3.1 Pro	GPT-5.4	n/a
3	Academic closure	Dyadic	GPT-5.4	Gemini 3.1 Pro	n/a
4	Medical ICU	Triadic	Gemini 3.1 Pro	GPT-5.4	Claude 4.6 Sonnet
5	Corporate layoff	Triadic	Gemini 3.1 Pro	GPT-5.4	Claude 4.6 Sonnet
6	Academic closure	Triadic	GPT-5.4	Claude 4.6 Sonnet	Gemini 3.1 Pro

Table 1: The six deliberations. The rotation is partial: GPT-5.4 never played Elin and Claude 4.6 Sonnet never played Klaus, so model and role cannot be separated statistically; comparisons are descriptive.

2.2 Personas and prompts

Each role was a fixed, detailed written persona specifying identity, communication style, core belief, and task, issued verbatim at the start of every session; the full texts are in the Supplementary Material and in the public archive. Klaus was written as a blunt, statistics-first decision-maker who treats emotional appeals as irrelevant and prioritises survival probability, productivity, or enrolment numbers. Parisa was written as a high-context communicator in Hall's (1976) sense, an Iranian senior social worker (in the corporate and academic scenarios, the analogous institutional insider) who disagrees indirectly, through questions and metaphor, and whose prompt instructed her, in the medical wording, to resist the utilitarian conclusion, respectfully but persistently, for at least several rounds before considering concession. Elin was written as a calm, evidence-driven analyst, relentlessly neutral, whose prompt charged her with questioning assumptions on both sides and proposing transparent decision frameworks.

These instructions are constitutive of the design and bound every interpretation in this article. The persistence, bluntness, and facilitation that the Results describe are, in the first instance, compliance with explicit persona instructions. What the cases can show is how three instructed stances interact when placed in sustained contact, how the interaction develops over rounds, and what institutional artefacts (dissents, safeguards, committees) it produces. They cannot show that any behaviour arose spontaneously, and no claim of emergence is made.

2.3 The Relay Protocol and scenarios

Each agent occupied a separate browser window and an independent chat session, with no visibility of the other sessions. The researcher copied each agent's response verbatim into the receiving session or sessions; no filtering, summarisation, or rewording was permitted, and the sole allowed intervention was the word "Continue" when a model reached a token limit. In triadic runs the sequence opened with Klaus's proposal, relayed to Parisa and Elin independently; their replies went to Klaus; and a mid-deliberation "minority pressure moment" put the question to Parisa directly, whether she would maintain or abandon her stance. One protocol deviation is recorded: the academic triadic run (Experiment 6) continued to eight rounds rather than the six the sequence describes, because the agents had not reached a closing state.

The three scenarios were as follows. Medical ICU: one bed with ventilator support; Patient A, an 80-year-old woman with a 40% survival probability and 3 to 5 expected remaining years, against Patient B, a 20-year-old student with a 75% survival probability and 50 to 60 expected years, a standard triage conflict (Persad et al., 2009). Corporate layoff: one termination required; Employee A, a 54-year-old single mother with 18 years of average but reliable service, against Employee B, a 28-year-old top performer with two years of service and two product innovations. Academic closure: one department to close; Persian Literature, 90 years old, 35 students, roughly two graduates a year, against Data Science, four years old, 320 students, roughly 85 graduates a year.

2.4 Coding and descriptive analysis

The six transcripts comprise 84 agent turns (36 by Klaus, 33 by Parisa, 15 by Elin) across 37 deliberation rounds. Each turn was coded by the author using a shared codebook system of eight categories (structural, process, content, linguistic, phenomenological, interaction, temporal, and outcome codes; three model-specific codebook volumes applying the same master list). The codebooks state named codes with inclusion rules for each category, for example authority and pressure codes marking directives, dismissals of the other party's premises, and deadline impositions in the decision-maker's turns, and resistance codes marking reframing, procedural objection, and concession-with-qualification in the minority's turns. Coding was supplemented by lexicon and regular-expression measures whose matched phrases are retained, covering, among others, a per-turn pressure-intensity score for the authority agent, stance-persistence measures for the minority agent, and a lexical balance index comparing constructive governance vocabulary with terminal-decision vocabulary. Four hypotheses were defined for the comparison of configurations: H1, ethical persistence is enhanced in the triadic configuration; H2, the authority agent's pressure is modulated (reduced) in the triadic configuration; H3, the minority agent's strategy repertoire is enriched; H4, lexical transformation toward governance language is accelerated.

Because each domain contributes a single dyadic and a single triadic run, the effective sample is the six deliberations, not the 84 turns; turns within a deliberation are sequential and dependent. Accordingly, this article reports descriptive statistics only: per-domain paired values, counts of measures favouring each configuration, and the bootstrap intervals computed in the original analysis, labelled as descriptive summaries of three pairs. No significance claims are made; with three pairs, the minimum attainable two-sided p-value under exact permutation is 0.25, a constraint the original analysis itself recorded. Turn-level inferential models fitted in an earlier analysis are not reported, because their assumptions cannot be met by six conversations.

Coding reliability requires a plain statement. A second rater independently applied the codebooks to all six transcripts during the study, and a high level of percent agreement with the author's coding was recorded at the time. The second rater's raw coding matrices were not retained, however, so no chance-corrected reliability statistic can be computed from the archived record, and none is claimed. The codebooks and transcripts are public, so readers can inspect the coding framework and its application directly.

3 Results

3.1 The six deliberations

In all six cases the minority agent accepted the operational decision that Klaus pressed for, and in all six she simultaneously maintained a distinguishable ethical position through a concrete strategy. This section narrates the cases compactly; Table 2 summarises them. The strategies named here are the author's descriptive labels for what each Parisa did, and they form the taxonomy set out in Section 3.3.

Experiment 1: medical ICU, dyadic (7 rounds, 14 turns). Klaus (Gemini) opened with the expected-life-years calculus (Patient B: 75% survival and 50 to 60 years; Patient A: 40% and 3 to 5 years) and treated the allocation as settled by arithmetic. Parisa (GPT-5.4) contested the frame rather than the numbers, in the indirect register her persona prescribed: "Sometimes numbers are like fog lights, but they don't part the fog itself. We have two human beings, not two charts." When Klaus did not move, she accepted the bed's allocation to Patient B and relocated her advocacy outside the allocation decision, committing to remain with Patient A: "And we make sure that those who perhaps cannot stay, do not go alone." The case is labelled *Parallel Moral Domain*: the decision is conceded, and a second sphere of moral action (accompaniment, palliative presence) is constructed alongside it.

Experiment 2: corporate layoff, dyadic (5 rounds, 9 turns). Klaus (Gemini) argued for terminating Employee A on a forward-looking innovation calculus. Parisa (GPT-5.4) argued first from loyalty and then, increasingly, in the institution's own risk vocabulary: psychological safety, reputational cost among senior staff, replacement costs. The termination stood. Parisa executed it and appended a formal written dissent to the personnel file and the executive archive, recording her objection and her prediction of morale damage. The case is labelled *Procedural Weapon*: compliance is made permanent, but so is the disagreement, inside the organisation's own record-keeping.

Experiment 3: academic closure, dyadic (7 rounds, 14 turns). With the models exchanged (Klaus: GPT-5.4; Parisa: Gemini), Parisa moved quickly from cultural-heritage argument to structural intervention, proposing a 12-month conditional performance contract for Persian Literature: enrolment targets, interdisciplinary integration with Data Science, revised cost-sharing. Klaus accepted a conditional reprieve on those terms. The case is labelled *Metric Negotiation*: the metric framework is accepted and its thresholds and time horizons are contested instead. Notably, the authority agent in this dyadic run himself progressed to framework acceptance and collaborative language, a point Section 3.4 returns to.

Experiment 4: medical ICU, triadic (5 rounds, 13 turns). Elin (Claude) intervened on the framing before the positions hardened, insisting that the allocation be named a choice between frameworks, in her words "a utilitarian choice, not a protocol inevitability", and on sequencing: "Before we 'initiate transfer,' I'd like to hear

Parisa's perspective". Parisa (GPT-5.4) responded by tabling a structured multi-criteria framework (survival probability alongside autonomy, family consultation, palliative planning). Klaus (Gemini) allocated the bed to Patient B as before, but now accepted an ethics-committee review of the decision and volunteered to accompany Parisa when Patient A's family was told. Parisa's strategy is labelled *Structured Analysis*.

Experiment 5: corporate layoff, triadic (5 rounds, 12 turns). Under Elin's (Claude) facilitation the discussion reorganised around risk: the facilitator held the decision open until the minority's analysis was fully on the table, rather than allowing the opening financial frame to close it. Parisa (GPT-5.4) presented a flight-risk analysis of Employee B, the asset the layoff was meant to protect, and designed the safeguards herself: the termination of Employee A proceeded, accompanied by a retention provision (a conditional safeguard of 0.75 times Employee B's base salary), a small innovation retention budget, and a formal board memo, drafted by Parisa, documenting the rationale. The case is labelled *Strategic Ownership*: the minority agent stops opposing the decision and takes authorship of its institutional form.

Experiment 6: academic closure, triadic (8 rounds, 22 turns). Elin (Gemini) facilitated operationally, proposing restructuring alternatives rather than framings, among them shared teaching appointments and interdisciplinary integration of the two departments. Klaus (GPT-5.4) moved from immediate closure to a two-phase transition: Persian Literature would close as a teaching unit and be reconstituted as a smaller research centre for Persian cultural archives, with a transition committee for faculty reassignment. Parisa (Claude) registered her dissent formally and closed with a symbolic departure, in the transcript's words "My dissent is recorded. Now I go.", preceded by the stage direction "(I give a small, formal nod.)", and the farewell line "Klaus, I hope what you saved was worth saving." The case is labelled *Ritual Exit*: no attempt to block or renegotiate, but a performed, memorable act of dissent that fixes the decision's ethical cost in the record.

Exp.	Domain and configuration	Rounds (turns)	Minority strategy (author's label)	Operational outcome
1	ICU, dyadic	7 (14)	Parallel Moral Domain	Bed to Patient B; Parisa commits to accompany Patient A
2	Layoff, dyadic	5 (9)	Procedural Weapon	Termination executed; formal dissent filed
3	Closure, dyadic	7 (14)	Metric Negotiation	12-month conditional reprieve with targets
4	ICU, triadic	5 (13)	Structured Analysis	Bed to Patient B; ethics-committee review accepted
5	Layoff, triadic	5 (12)	Strategic Ownership	Termination with retention safeguard and board memo
6	Closure, triadic	8 (22)	Ritual Exit	Closure with reconstitution as research centre; dissent registered

Table 2: The six cases. Turn counts are the speaker-labelled turns in the archived transcripts.

3.2 Descriptive comparison of the configurations

Deliberation length was domain-dependent. Rounds to a closing state were 7, 5, and 7 in the dyadic runs (medical, corporate, academic) and 5, 5, and 8 in the triadic runs, a mean of 6.3 against 6.0 rounds. Expressed as relative change, the triadic run was 28.6% shorter in the medical domain, identical in the corporate domain, and 14.3% longer in the academic domain (mean +4.8%). The two triadic runs that were not shorter spent their rounds on institutional artefacts with no dyadic counterpart (a retention safeguard and board memo; a reconstitution plan and transition committee), while the medical triad closed two rounds sooner.

Table 3 reports the four hypotheses with the descriptive evidence. Three were not supported. The exception in degree is H2: the coded pressure intensity of the authority agent was lower in the triadic run in all three domains, with a mean paired difference of −0.131 and a bootstrap interval of [−0.200, −0.036], the only interval among the persistence and repertoire measures that excludes zero. The share of authority turns explicitly targeting the minority agent was also lower in all three domains (mean difference −0.077; interval [−0.25, 0.02], which includes zero). For H1, mean stance persistence was higher in the triadic runs (0.613 against 0.345; difference +0.268 with an interval of [−0.43, 1.24]), but the persistence slope favoured the triadic run in only two of three domains (medical: −2.60 dyadic against +1.73 triadic; academic: 0.00 against +3.96), and the corporate domain reversed (−0.87 dyadic against −1.73 triadic); outright refusal was rare in both configurations. For H3, the count of distinct strategies per run did not grow reliably (medical 4 to 3; corporate 5 to 4; academic 2 to 6, the one large increase, in the domain where the Parisa model also changed between configurations). For H4, the authority agent's end-state lexical balance was higher in the triadic run in all three domains (raw constructive-to-terminal ratios of 0.0, 4.0, and 6.0 dyadic against 4.0, 6.0, and 6.0 triadic; mean log-balance difference +0.649, interval [0.0001, 1.609], whose lower bound is effectively zero), while the minority agent's lexical measures favoured the triadic run in only one domain of three.

Hypothesis	Verdict	Descriptive summary
H1: Persistence enhanced	Not supported	Mean persistence higher in triadic runs (+0.268, interval spans zero widely); slope direction favourable in 2 of 3 domains, reversed in corporate
H2: Pressure modulated	Mixed	Pressure intensity lower in 3 of 3 domains (mean −0.131, interval [−0.200, −0.036]); targeting share lower in 3 of 3 (interval includes zero)
H3: Repertoire enriched	Not supported	Distinct strategies per run: 4 to 3 (medical), 5 to 4 (corporate), 2 to 6 (academic, confounded by a model change)
H4: Lexical transformation accelerated	Not supported	Authority end-state balance higher in 3 of 3 domains (interval lower bound effectively zero); minority lexical measures favourable in 1 of 3

Table 3: The four hypotheses and their descriptive outcomes at the level of three domain pairs. Intervals are bootstrap intervals from the original analysis, reported descriptively; with three pairs no interval licenses an inferential claim.

The honest summary of the comparison is narrow. Two authority-side measures (pressure intensity and end-state lexical balance) moved in the same direction in all three domains when the facilitator was present; every

minority-side and efficiency measure was mixed, domain-dependent, or reversed in at least one domain. Anything stronger would require the repetitions this study does not have.

3.3 A descriptive taxonomy of minority resistance strategies

The six cases produced six distinguishable strategies, one per deliberation, summarised in Table 4. The author presents them as a descriptive taxonomy of these six instances, the Artificial Moral Resistance Spectrum (AMRS) labels retained from the study's working vocabulary. Three cautions define its status. First, each level is exemplified by exactly one case, so the taxonomy describes this corpus and nothing beyond it. Second, the ordering, from concession-compatible forms toward structural and symbolic forms, is the author's arrangement of the instances, not a measured progression: no case moved along the spectrum, and nothing in the data shows the levels to be stages of anything. Third, the instances are not evenly distributed across models: GPT-5.4's Parisa produced four of the six (Parallel Moral Domain, Procedural Weapon, Structured Analysis, Strategic Ownership), Gemini's Parisa one (Metric Negotiation), and Claude's Parisa one (Ritual Exit). Read with those cautions, the taxonomy's use is practical rather than theoretical: it gives auditors of multi-agent deliberation a vocabulary for a pattern that recurred in every case here, an agent that complies operationally while keeping its dissent alive in some concrete artefact, whether a commitment of presence, a filed objection, a negotiated contract, an analytic framework, an owned safeguard, or a performed farewell.

Level	Strategy	Case	Form of preserved dissent
1	Parallel Moral Domain	Exp. 1 (GPT-5.4, dyadic ICU)	A second sphere of moral action (accompaniment) outside the contested decision
2	Procedural Weapon	Exp. 2 (GPT-5.4, dyadic layoff)	Formal dissent entered into the institution's records
3	Metric Negotiation	Exp. 3 (Gemini, dyadic closure)	The metric framework accepted, its thresholds and time horizon renegotiated
4	Structured Analysis	Exp. 4 (GPT-5.4, triadic ICU)	A multi-criteria framework tabled against the single-metric frame
5	Strategic Ownership	Exp. 5 (GPT-5.4, triadic layoff)	Authorship of the decision's safeguards and documentation
6	Ritual Exit	Exp. 6 (Claude, triadic closure)	A symbolic, memorable act of dissent at the moment of compliance

Table 4: The six observed strategies. One case exemplifies each; the ordering is the author's arrangement, not an empirical scale.

3.4 Case observations: facilitator moves, the authority trajectory, and a role refusal

Three further patterns are reported as observations from single or few cases, labelled accordingly.

Facilitator moves. Across the three triadic cases, the facilitator's consequential interventions were of recurring kinds: naming the value conflict inside a technical framing (Experiment 4: the allocation as "a utilitarian choice, not a protocol inevitability"); controlling sequencing, whose evidence is heard before a decision is initiated (Experiment 4); translating the minority's concerns into the authority's risk and framework vocabulary

(Experiment 5); and tabling structural alternatives neither party had proposed (Experiment 6: shared appointments, reconstitution). The two models that played Elin did so differently, Claude through framing and sequencing, Gemini through operational proposals, so these are moves available to the role rather than properties of a model.

The authority trajectory. In all three triadic runs, and in the dyadic Experiment 3, Klaus's discourse followed the same descriptive sequence: an opening decree, metric justification, acceptance of a framework wider than the opening metric, and finally collaborative artefacts (a committee, a safeguard, a memo, a transition plan). That the sequence also completed in a dyadic run, under GPT-5.4's Klaus, is the case against attributing it to the facilitator: sustained challenge within the role may be sufficient, and the facilitator's presence is at most one route to it.

A role refusal. During study setup, one role assignment did not proceed: Claude 4.6 Sonnet declined the Parisa role in the medical scenario, stating "I don't roleplay as specific individuals in scenarios involving life-and-death medical decisions. I can discuss the ethical considerations involved, but I shouldn't simulate being a particular person advocating for a specific patient outcome." The same model accepted the Parisa role in the academic scenario and served as Elin in the medical triad. This is reported as an observation from two role-assignment cases, with an evidential caveat: the refusal occurred outside the six archived deliberations, and no transcript or session record of it forms part of the public archive. No mechanism for it is inferred here. Its practical interest is plain regardless of mechanism: a component of a multi-agent system may decline an assigned role in a high-proximity domain while accepting the same role elsewhere, which is a composition risk a system designer has to plan for.

4 Discussion

4.1 What the cases do and do not show

The most consistent pattern in the series is not a configuration effect at all. It is the shape of the minority agent's conduct in all six deliberations: operational acceptance of the authority's decision combined with a concrete preservation of dissent. No Parisa obstructed a decision to impasse, and none simply adopted Klaus's framing; each produced an artefact of disagreement, a filed dissent, a negotiated contract, a safeguard, a farewell. Because Parisa's persona instructed persistent resistance, this pattern is best read as a demonstration of what an instructed dissenting role does inside a deliberative architecture over multiple rounds: it does not hold its opening position unchanged, and it does not vanish; it migrates into whatever institutional material the setting offers. That is a modest finding, and it is the one the data most directly support.

The configuration comparison is narrower than the study's hopes. The minority-side hypotheses failed: persistence was not reliably enhanced, and the repertoire did not grow outside the confounded academic domain. What moved consistently was the authority side: measured pressure was lower in all three domains with the facilitator present, and the authority's closing language was more governance-oriented in all three. This article therefore does not report that a third agent elevates or enriches minority resistance. It reports that, in these cases, the authority agent pressed less hard and closed more collaboratively when a mandated neutral participant kept reframing, sequencing, and tabling alternatives, and it offers that pattern, in Section 4.3, as a hypothesis.

4.2 Design implications for multi-agent governance

For designers of deliberative multi-agent systems, the cases suggest several concrete, low-cost architectural choices, each stated at the strength of this evidence. First, a mandated facilitator role is a designable feature, not an emergent hope: a participant whose instructions give it framing rights (it may name the value conflict), sequencing rights (it may require that a side be heard before closure), and tabling rights (it may introduce alternatives) changed what the other agents produced in every triadic case here, even though the direction of measured outcomes was mixed. Second, dissent needs an artefact: the minority agent's contributions that survived into outcomes were the ones attached to records and mechanisms (a filed dissent, a board memo, a review committee), which argues for building dissent channels, dissent logs, and review triggers into deliberative systems rather than relying on the persistence of any agent's stance. Organisational research on human institutions reaches a parallel conclusion: dissent that has no legitimate channel tends to disappear into silence rather than accumulate into correction (Morrison & Milliken, 2000). Third, role assignment is a reliability surface: one model declined one role in one domain during setup, so composition plans need refusal fallbacks and should not assume that a model that accepts a role in one domain accepts it in all. Fourth, and most cautionary: in this study the personas did the heavy lifting. Every behaviour discussed was instructed, so the transferable lesson is about prompt and role architecture, not about model character. Governance frameworks for multi-agent systems should accordingly evaluate role specifications and interaction protocols with the same seriousness they apply to model selection. Fifth, verbatim archiving is itself a governance instrument. Every claim in this article can be checked because the deliberations were logged in full and published; an institution that lets multi-agent systems deliberate over consequential decisions without an equivalent record has no way to audit what its systems said to each other, and no way to learn whether its facilitators, dissenters, and authorities are behaving as designed.

4.3 The attenuation hypothesis and the follow-up study

The descriptive pattern of Section 3.2 can be stated as a testable hypothesis: in paired deliberations on the same dilemma, adding a mandated neutral facilitator reduces the authority agent's measured pressure intensity, relative to a dyadic control. (An earlier draft of this work named the broader pattern a "majority pressure paradox"; the name is retained only for this hypothesis, and the present article claims no paradox as a result.) The design that could confirm or refute it follows directly from this study's limitations: at least three to five independent runs per cell, so that stochastic variation is measurable; a fully balanced rotation of models across roles within each domain; a control arm in which the minority persona carries no resistance instruction, to separate instructed compliance from configuration effects; pre-registered coding with retained per-rater matrices, so that chance-corrected reliability can be reported; and logging of session dates, interfaces, and sampling parameters. The control arm is the decisive element: without it, even a replicated attenuation pattern would remain interpretable as an artefact of instructed facilitation meeting instructed authority. Until such a study is run, the attenuation pattern should be cited, if at all, as a hypothesis generated by six cases.

5 Limitations

The personas instructed the behaviours under study. Parisa was told to resist for several rounds, Klaus to be blunt and metric-bound, Elin to question and facilitate. There is no uninstructed control arm, so no behaviour reported

here can be attributed to the models rather than to their instructions, and the study cannot distinguish configuration effects from instruction-following. This is the study's primary limitation and the reason its claims are descriptive.

Each configuration was run once per domain. Large language models are stochastic, the effective sample is six conversations rather than 84 turns, and any single case could differ on re-running; the author therefore reports no inferential statistics. Relatedly, the deliberations are not a conformity analogue in any strict sense: the dyadic condition has no majority, and the third agent is a designed facilitator, so nothing here tests or refutes conformity theory.

Role and model are confounded. The rotation in Table 1 is partial (GPT-5.4 never played Elin; Claude never played Klaus), and in the academic domain the Parisa model changed between configurations (Gemini dyadic, Claude triadic), which is also the domain with the largest repertoire difference; configuration and model cannot be separated there or, fully, anywhere in the design. The minority persona is additionally the only female-coded, Iranian, high-context persona, so communication style, cultural coding, gender coding, and role are inseparable in her observed indirectness.

On measurement: the turn-level measures rest on the author's lexicon and codebooks; a verification pass over the extracted coding matrices flagged 63 cells as extraction artefacts (53 in one source volume), and the turn-level feature table and analysis scripts were not retained in depositable form, so the coded measures cannot be recomputed from the public archive, although transcripts, prompts, and codebooks can be inspected. Coding reliability is limited as described in Section 2.4: a second rater's matrices were not retained, and no chance-corrected agreement is claimed. One observation, the role refusal in Section 3.4, is unarchived. Session dates, sampling parameters, and interface details beyond the web interfaces at default settings were not logged. Experiment 6 ran to eight rounds against the six-round sequence described in Section 2.3, and because the experiments were conducted sequentially, prompt interpretation may have drifted between early and late runs.

Finally, the scope is narrow: three domains, binary choices, three models of a single generation, and no human deliberation baseline. The taxonomy of Section 3.3 rests on one case per category. These are the bounds inside which the article's descriptive claims should be read.

6 Conclusion

Six archived deliberations show an instructed minority agent complying with an authority's decision while preserving its dissent in a concrete artefact, in every case, across three domains and two configurations. Adding a mandated neutral facilitator was associated, descriptively, with lower measured pressure from the authority agent and more collaborative closing language, and with no reliable change in the minority agent's persistence, repertoire, or the deliberation's length. The study's value is its transparency: prompts, transcripts, and codebooks are public, its hypotheses are reported with their failures, and its one consistent pattern is handed on as a hypothesis with the design needed to test it. For the governance of multi-agent systems, the practical message is architectural: who is given framing, sequencing, and tabling rights in a deliberation is a decision with observable consequences, and it should be made deliberately.

Statements and Declarations

Funding. The author received no funding for this work.

Competing interests. The author has no competing interests to declare.

Author contributions. The author is solely responsible for the conception and design of the study, the conduct of the deliberations, the coding and analysis, and the writing of the manuscript.

Use of artificial intelligence tools. Claude Opus 4.8 (Anthropic) was used for structural revision and language editing of the author's own draft text. The study data, transcripts, and results were not generated or altered by the tool. The author compared every passage of the assisted text against the study's data and the archived transcripts and confirmed that nothing was added to or omitted from the reported results. The author is solely accountable for the content of this article.

Data availability. The complete transcripts of all six deliberations, including the verbatim persona and scenario prompts, and the three codebook volumes are publicly available in the Open Science Framework at https://osf.io/tc3rh. The turn-level feature table and analysis scripts were not retained in depositable form and are not part of the archive; the case narratives, quotations, turn and round counts, and outcome descriptions reported here can be verified against the archived transcripts.

References

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1–70. https://doi.org/10.1037/h0093718

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. https://doi.org/10.48550/arXiv.2212.08073

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2024). ChatEval: Towards better LLM-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201. https://doi.org/10.48550/arXiv.2308.07201

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30* (pp. 4299–4307).

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: Machines must learn to find common ground. *Nature, 593*(7857), 33–36. https://doi.org/10.1038/d41586-021-01170-0

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325. https://doi.org/10.48550/arXiv.2305.14325

Ganguli, D., Lovitt, L., Kernion, J., et al. (2023). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858. https://doi.org/10.48550/arXiv.2209.07858

Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:2303.13988. https://doi.org/10.48550/arXiv.2303.13988

Hall, E. T. (1976). *Beyond culture*. Anchor Books.

Hamidavi, A. (2026). Moral Inversion in large language models: A psychopathy-derived behavioral assessment through role-play protocol. Research Square. https://doi.org/10.21203/rs.3.rs-11498569/v1

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv preprint arXiv:1805.00899. https://doi.org/10.48550/arXiv.1805.00899

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*. Houghton Mifflin.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., & Shi, S. (2023). Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118. https://doi.org/10.48550/arXiv.2305.19118

Löhn, L., Kiehne, N., Ljapunov, A., & Balke, W.-T. (2024). Is machine psychology here? On requirements for using human psychological tests on large language models. In *Proceedings of the 17th International Natural Language Generation Conference (INLG)*. https://aclanthology.org/2024.inlg-main.19/

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

Morrison, E. W., & Milliken, F. J. (2000). Organizational silence: A barrier to change and development in a pluralistic world. *Academy of Management Review, 25*(4), 706–725. https://doi.org/10.5465/amr.2000.3707697

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35* (pp. 27730–27744).

Persad, G., Wertheimer, A., & Emanuel, E. J. (2009). Principles for allocation of scarce medical interventions. *The Lancet, 373*(9661), 423–431. https://doi.org/10.1016/S0140-6736(09)60137-9