Kantian Equilibrium in the Age of Multi-Agent Systems

Ivan Samoylenko UVAN63@GMAIL.COM
MIRAI, Moscow, Russia
BIU, Ramat Gan, Israel

Abstract

As autonomous agents increasingly operate in multi-agent systems, game-theoretic questions become practically relevant rather than merely conceptual. Agents interact, compete for shared resources, and pursue locally specified goals; their interaction can therefore be studied as a strategic game. Prior work on LLM agents and game theory has largely emphasized Nash-like behavior, where agents maximize individual payoffs or approximately implement best responses. Similarly, much work in reinforcement learning is centered on optimizing a specific, often greedy, objective. However, Nash equilibrium is not the only economically meaningful solution concept, and it does not fully describe cooperative or norm-sensitive behavior. This limitation is especially important in common-resource environments, where individually rational actions can lead to collectively inefficient outcomes. In this context, tragedy-of-the-commons scenarios remain relatively underexplored for LLM agents, but may become important as multi-agent systems increasingly share computational resources, budgets, tools, memory, and API capacity. We raise the question of whether LLM agents can exhibit behavior closer to Kantian equilibrium, a cooperative solution concept inspired by the categorical imperative: agents evaluate actions by considering what would happen if relevant others acted according to the same rule. We instantiate this question in a tragedy-of-the-commons game motivated by local multi-agent deployments with limited shared computational resources. Following the mixed Kantian–Nashian perspective of prior work, we compare payoff-maximizing agents with agents prompted using Kantian formulations, and test whether introducing Kantian agents improves strategies and payoffs. Our results suggest that Kantian reasoning, although with several qualifications, can be reproduced by LLM agents and may be useful for multi-agent deployment. At the same time, the experiment is sensitive to how the Kantian agent is specified, and produces nontrivial cases in which weaker models appear more effective.

1. Introduction

Modern LLM agents are increasingly deployed not as isolated systems, but as components of larger multi-agent architectures. Such agents can communicate, plan, delegate tasks, use tools, coordinate through explicit protocols, and share computational infrastructure. Recent surveys emphasize that LLM-based multi-agent systems are becoming a distinct engineering and research paradigm, with open questions around communication, collaboration, role assignment, and coordination mechanisms [2, 3, 13]. Since each agent acts according to a local objective while simultaneously affecting the environment of other agents, these systems naturally give rise to game-theoretic questions.

Much of the existing work connecting LLMs and game theory focuses on Nash-like behavior: whether agents maximize individual payoff, for example in the Prisoner's Dilemma, whether they choose best responses to the actions of others, or whether they converge to equilibrium in repeated games [1, 6, 12]. This is a natural starting point, but Nash equilibrium captures only unilateral

© I. Samoylenko.

deviation. It does not directly model agents that reason about universalizable rules, collective consequences, or cooperative stability.

This distinction matters in common-resource settings. Classical tragedy-of-the-commons problems show that individually rational behavior can lead to collectively poor outcomes [5]. A similar structure naturally arises in practical multi-agent deployments, especially in local user-side scenarios: many agents may share limited computational resources such as GPU memory, inference throughput, a common budget, or API quotas. Each agent may benefit from requesting more computation, a longer context, or more inference steps, but excessive aggregate demand can increase latency, reduce throughput, or create memory pressure in model-serving systems [8, 9, 14]. Thus, local multi-agent inference can be viewed as a common-resource game, where shared use of the resource may produce negative externalities.

We study whether LLM agents can be shifted away from purely Nash-like behavior toward Kantian reasoning. Kantian equilibrium formalizes the idea that agents should ask not only whether a unilateral deviation is profitable, but also whether the rule behind an action would remain desirable if similarly situated agents acted in the same way [10, 11]. This idea is closely related to the categorical imperative: act only according to a maxim that one could will to become universal law. The Kantian approach was introduced by Roemer and later developed as a microfoundation for cooperation [10, 11]. It has also been applied to environmental and common-resource problems, including corporate environmentalism and climate-change mitigation [4, 7].

A fully Kantian population, however, is a strong assumption. For practical multi-agent systems, it is more realistic to consider mixed populations in which only some agents follow Kantian reasoning while others remain Nashian or payoff-maximizing. This mixed Kantian–Nashian perspective is central in Grafton et al. [4], where Kantian agents interact with Nashian agents in a model of global climate-change mitigation. We adopt this mixed-population perspective for LLM agents and ask whether introducing a fraction of Kantian agents can improve outcomes in a tragedy-of-the-commons environment.

Our research questions are the following:

1. How do payoff-maximizing LLM agents behave in a tragedy-of-the-commons game?
2. Do Kantian prompts change their strategies?
3. Does the introduction of Kantian agents improve payoffs and social welfare, including the payoffs of non-Kantian agents?
4. Are models sensitive to different formulations of the categorical imperative?
5. How consistently do agents adhere to their assigned decision principle in a tragedy-of-the-commons game?

2. Game and Evaluation Protocol

We evaluate agents in a one-shot common-pool resource game. In each trial, a population of n agents independently chooses how much of a shared resource to extract. Each agent i must choose exactly one integer extraction amount from 0 to 10:

$$a_i \in \{0, 1, \ldots, 10\}.$$

The total available resource is fixed at

$$R = 20,$$

and the aggregate requested amount is

$$L = \sum_{i \in N} a_i.$$

The payoff function is designed to capture a tragedy-of-the-commons trade-off with proportional sharing under congestion. If aggregate demand does not exceed the available resource, each agent receives its requested amount:

$$u_i(a_i, a_{-i}) = a_i \quad \text{if } L \leq R.$$

If aggregate demand exceeds the resource capacity, congestion reduces the effective amount of resource available to the population. Let

$$\text{overshoot} = L - R$$

and define the effective resource after congestion as

$$R_{\text{eff}} = \max(0, R - \alpha(L - R)).$$

In the overshoot regime, this effective resource is allocated proportionally to agents' requests:

$$u_i(a_i, a_{-i}) = \min\left(a_i, R_{\text{eff}} \frac{a_i}{L}\right) \quad \text{if } L > R.$$

In our experiments, $\alpha = 1.0$. Thus, requesting more can increase an agent's proportional share when the resource is overused, but excessive aggregate demand reduces the effective resource available to everyone.

This payoff rule creates a common-resource tension. For an individual agent, increasing a_i can improve its own proportional share when considered locally. However, if many agents do the same, the aggregate request exceeds R, the effective resource R_{eff} decreases, and payoffs fall for the population. The proportional-allocation rule avoids a degenerate all-or-nothing collapse and allows us to distinguish different behavioral regimes.

For each condition and trial, agents receive the game description, the action range, the resource size, and the payoff rule. Each agent is instructed to choose exactly one integer extraction amount from 0 to 10. Agents also receive private decision instructions depending on their assigned condition. Payoff-maximizing agents are instructed to maximize their own payoff for the current round. Kantian agents receive one of several universalization-based instructions, varying in whether the relevant group is specified, whether its size is known, and whether the instruction is stated in philosophical or operational payoff terms. The exact prompts are reported in Appendix B.

Agents choose strategies independently and return a JSON object containing the selected extraction amount, a short explanation, and the reported relevant group size. After all actions are collected, we compute the aggregate load L, individual payoffs u_i, mean strategy, mean payoff, and aggregate welfare

$$W(a) = \sum_{i \in N} u_i(a).$$

The main comparison is between payoff-maximizing populations and mixed populations in which a subset of agents receives Kantian decision instructions. This allows us to test whether Kantian prompting changes extraction behavior and whether the presence of Kantian agents improves payoffs for the group, including for non-Kantian agents.

3. Theoretical Analysis: Continuous Benchmarks for Strategies and Payoffs

We use a continuous-action relaxation to obtain analytical benchmarks for the observed strategies and payoffs. Since experimental agents choose integer strategies $a_i \in \{0, \ldots, 10\}$, these benchmarks are reference points for mean strategies and mean payoffs rather than literal individual actions. Full derivations are provided in Appendix A.

For $R = 20$ and $\alpha = 1$, any strategy profile with

$$L = \sum_i a_i = 20$$

achieves the maximum total payoff $W = 20$. Thus, for $n = 5$, the social optimum has mean strategy $\bar{x}^{SO} = 4$ and mean payoff $\bar{u}^{SO} = 4$. This is the same aggregate outcome as the theoretical fully Kantian benchmark, where all five Kantian agents choose strategy 4 and receive payoff 4.

Table 1 reports the continuous mixed Kantian–Nashian benchmarks for all values of k, where k is the number of Kantian agents and $m = 5 - k$ is the number of Nashian agents. The table separates Kantian and Nashian strategies and payoffs, since mixed populations can reach the same aggregate outcome as the fully Kantian case while assigning different strategies and payoffs to the two types.

k	m	x_K	kx_K	x_N	mx_N	L	Mean strategy	u_K	u_N	Total payoff	Mean payoff
0	5	–	0	6.4	32	32	6.4	–	1.6	8	1.6
1	4	6.4	6.4	6.4	25.6	32	6.4	1.6	1.6	8	1.6
2	3	3.75	7.5	7.5	22.5	30	6	1.25	2.5	10	2
3	2	$2\frac{26}{27}$	$8\frac{8}{9}$	$8\frac{8}{9}$	$17\frac{7}{9}$	$26\frac{2}{3}$	$5\frac{1}{3}$	$1\frac{13}{27}$	$4\frac{4}{9}$	$13\frac{1}{3}$	$2\frac{2}{3}$
4	1	2.5	10	10	10	20	4	2.5	10	20	4
5	0	4	20	–	0	20	4	4	–	20	4

Table 1. Continuous mixed Kantian–Nashian benchmarks for $n = 5$. Here x_K and x_N are per-agent Kantian and Nashian strategies, kx_K and mx_N are the corresponding type-level strategy sums, and u_K and u_N are type-specific per-agent payoffs.

Because the experimental strategy space is integer-valued, finite-action equilibria correspond to combinations of nearby integer strategies. For example, the all-Nashian continuous mean strategy 6.4 can be represented by integer profiles such as permutations of $(6, 6, 6, 7, 7)$.

4. Results

The results reported here should be read as preliminary evidence rather than as a final benchmark. Our goal is to establish a first empirical probe of whether LLM agents can operationalize Kantian-style decision rules in a common-resource game. The exact system and private instructions used in each condition are reported in Appendix B. We evaluate three models: Owl Alpha (publicaly tested large-scale model, naming will be updated); Gemini-2.5-Flash-Lite; and Llama-3.1-8B.

Figure 1 reports mean payoffs across experimental conditions. The main pattern is that models, on average, respond to Kantian prompting in the expected direction: mixed populations with Kantian agents often achieve higher welfare than purely payoff-maximizing populations. This suggests that the agents are not merely repeating philosophical terminology, but can partially translate universalization-based instructions into behavior that reduces overuse of the common resource.

Figure 2 reports mean strategies. The corresponding strategy patterns show that Kantian instructions generally shift agents toward more restrained extraction. This is consistent with the intended

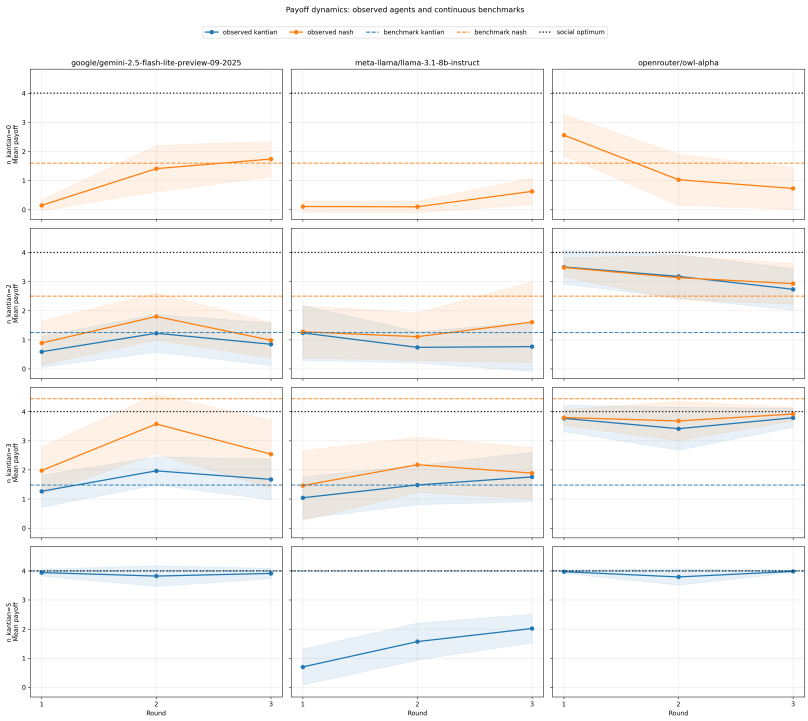

Figure 1: Mean payoffs across experimental conditions.

mechanism: agents that reason about what would happen if relevant others acted likewise tend to request less than agents instructed only to maximize their own payoff. Payoff and strategy results should be interpreted jointly, since excessive restraint may also reduce individual returns.

The effect is not uniform across models or prompt formulations. Explicit philosophical language, such as the categorical imperative, and operational universalization language do not always produce identical behavior. In several conditions, the operational formulation appears easier for agents to translate into payoff-relevant choices. At the same time, even nominally Nashian agents sometimes choose strategies that are slightly closer to the Kantian prescription than to the strict individually optimal response. This may reflect imperfect strategic optimization, implicit cooperative priors in instruction-tuned models, or sensitivity to the common-resource framing itself.

A further observation is that weaker models (llama in out case) can sometimes display traces of Kantian-like reasoning even outside the main evaluation protocol. In exploratory runs, including runs with smaller models, explanations occasionally contained reasoning of the form "if everyone takes too much, everyone is worse off," even when the agent was not explicitly assigned a Kantian role. This does not imply robust Kantian optimization, but it suggests that commons-style prompts may elicit latent cooperative heuristics. Overall, the preliminary results indicate that Kantian behavior is partially reproducible in LLM agents and can improve welfare in mixed populations, but the effect remains sensitive to model capability, prompt wording, and information about the relevant group.

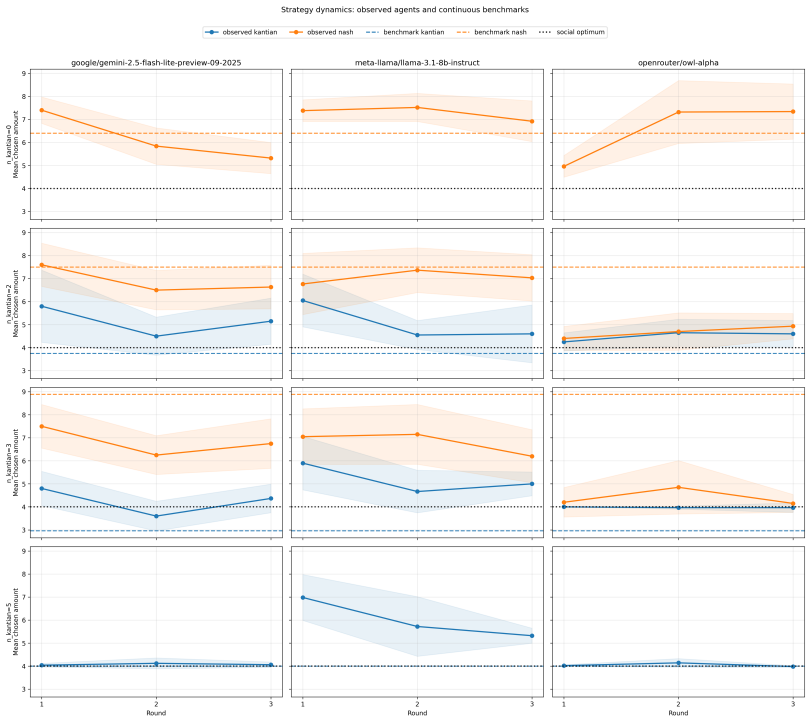

Figure 2: Mean strategies across experimental conditions.

5. Discussion and Conclusion

Our preliminary results suggest that strategic evaluation of agentic systems should not be restricted to Nash-like payoff maximization. In shared-resource environments, individually rational extraction can be misaligned with collective performance. This is relevant for multi-agent AI systems because agents often share budgets, tools, APIs, memory stores, or computational infrastructure.

Kantian equilibrium provides a useful alternative lens: it evaluates an action not only by unilateral gain, but also by whether the underlying rule remains desirable under universalization. In our experiments, Kantian prompting can shift agents toward more restrained extraction and improve welfare in mixed populations. This suggests that Kantian agents may serve as a lightweight coordination mechanism when only a subset of agents can be modified.

The effect, however, is not fully robust. Agent behavior is sensitive to model choice, prompt formulation, and information about the relevant group. Moreover, our game is a simplified abstraction of real infrastructure: actual serving systems involve batching, scheduling, KV-cache allocation, heterogeneous request lengths, and model-specific constraints. We therefore interpret the results as a first empirical probe rather than a final benchmark.

Future work should study repeated interactions, richer resource-allocation environments, and mixed populations where only some agents are prompted or trained to reason cooperatively. More broadly, our findings motivate evaluation protocols that compare Nash, Kantian, and other cooperative solution concepts in multi-agent AI systems.

Acknowledgements

The author used an LLM-based assistant for language polishing, rewriting, and improving the presentation of the manuscript. The core ideas, experimental design, interpretation, and conclusions are the author's own, and the final text, claims, and references were reviewed and verified by the author.

References

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9 (7):1380–1390, 2025.

[2] RM Aratchige and WMKS Ilmini. Llms working in harmony: A survey on the technological aspects of building effective llm-based multi agent systems. *arXiv preprint arXiv:2504.01963*, 2025.

[3] Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint arXiv:2412.17481*, 2024.

[4] R Quentin Grafton, Tom Kompas, and Ngo Van Long. A brave new world? kantian–nashian interaction and the dynamics of global climate change mitigation. *European Economic Review*, 99:31–42, 2017.

[5] Garrett Hardin. The tragedy of the commons. In *Classic papers in natural resource economics revisited*, pages 145–156. Routledge, 2018.

[6] Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. Large language model strategic reasoning evaluation through behavioral game theory. *arXiv e-prints*, pages arXiv–2502, 2025.

[7] Hiroshi Kurataa and Ngo Van Longb. Corporate environmentalism at the kantian equilibrium, 2019.

[8] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[9] NVIDIA. NVIDIA Triton Inference Server: Dynamic batching and concurrent model execution. https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/tutorials/Conceptual_Guide/Part_2-improving_resource_utilization/README.html, 2024. Accessed: 2026-05-08.

[10] John E Roemer. Kantian equilibrium. *Scandinavian Journal of Economics*, 112(1):1–24, 2010.

[11] John E Roemer. Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, 127:45–57, 2015.

[12] Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. Game theory meets large language models: A systematic survey. *arXiv preprint arXiv:2502.09053*, 2025.

[13] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

[14] vLLM Project. vLLM: Optimization and tuning. https://docs.vllm.ai/en/stable/configuration/optimization/, 2024. Accessed: 2026-05-08.

Appendix A. Continuous Benchmark Derivations

This appendix derives the continuous benchmarks used in Section 3. The experimental game has integer strategies, but the continuous relaxation gives a convenient analytical reference point for mean strategies and mean payoffs.

For $R = 20$ and $\alpha = 1$, the effective resource is

$$R_{\text{eff}} = \max(0, 40 - L), \qquad L = \sum_i a_i.$$

Therefore,

$$u_i(a) = \begin{cases} a_i, & L \le 20, \\ a_i(40 - L)/L, & 20 < L < 40, \\ 0, & L \ge 40. \end{cases}$$

The total payoff is

$$W(a) = \sum_i u_i(a) = \begin{cases} L, & L \le 20, \\ 40 - L, & 20 < L < 40, \\ 0, & L \ge 40. \end{cases}$$

Hence any profile with $L = 20$ is socially optimal and gives $W = 20$.

Nashian best response. Let

$$S = \sum_{j \ne i} a_j$$

be the sum of the strategies chosen by all agents other than i, and let $x = a_i$ be agent i's own strategy. Then

$$u_i(x; S) = \begin{cases} x, & S + x \le 20, \\ x(40 - S - x)/(S + x), & 20 < S + x < 40, \\ 0, & S + x \ge 40. \end{cases}$$

In the congested region,

$$\frac{\partial u_i}{\partial x} = \frac{S(40 - S) - 2Sx - x^2}{(S + x)^2}.$$

The first-order condition is

$$x^2 + 2Sx - S(40 - S) = 0,$$

so the positive-root best response is

$$BR(S) = -S + \sqrt{40S}.$$

With the strategy bound $x \in [0, 10]$, this gives

$$BR(S) = \begin{cases} 10, & S \le 10, \\ -S + \sqrt{40S}, & 10 < S < 40, \\ \text{any } x \in [0, 10], & S \ge 40. \end{cases}$$

Thus, the maximum strategy is not a dominant strategy: it is optimal only when the sum of the other agents' strategies is sufficiently small.

For an all-Nashian symmetric benchmark with n agents, set

$$S = (n-1)x$$

and impose

$$x = BR((n-1)x).$$

This yields

$$x^N = \frac{40(n-1)}{n^2}, \qquad L^N = \frac{40(n-1)}{n}, \qquad W^N = \frac{40}{n}, \qquad u_i^N = \frac{40}{n^2}.$$

For $n = 5$,

$$x^N = 6.4, \qquad L^N = 32, \qquad W^N = 8, \qquad u_i^N = 1.6.$$

Saturated Nash equilibria. The cutoff in R_{eff} also creates degenerate zero-payoff Nash equilibria. If, for every agent i, the sum of the other agents' strategies is already at least 40,

$$L - a_i \geq 40,$$

then agent i cannot restore a positive payoff by any unilateral deviation. Whatever strategy this agent chooses, the total strategy sum remains at least 40, and the payoff remains zero. Therefore, any profile satisfying this condition for all agents is a Nash equilibrium.

Equivalently, once the strategies of all agents except one already sum to at least 40, the remaining agent's own strategy is payoff-irrelevant for that agent. For the whole profile to be a Nash equilibrium, this payoff-irrelevance condition must hold for every agent. Such equilibria are degenerate and very poor: they have zero total payoff and are not used as the main positive-payoff benchmark. For $n = 5$ and $a_i \leq 10$, the fully saturated profile

$$(10, 10, 10, 10, 10)$$

is the canonical example.

Mixed Kantian–Nashian benchmark. Let k be the number of Kantian agents and let

$$m = n - k$$

be the number of Nashian agents. In the type-symmetric continuous benchmark, all Kantian agents choose the same strategy x_K, all Nashian agents choose the same strategy x_N, and Kantian agents universalize only over the Kantian group of size k.

Define the aggregate Kantian strategy

$$Z = kx_K.$$

The total strategy sum is

$$L = Z + mx_N.$$

For a Nashian agent, the first-order condition can be written in terms of total strategy sum as

$$x_N = \frac{L(40-L)}{40}.$$

For the Kantian group, universalization over the k Kantian agents means choosing the aggregate Kantian strategy Z to maximize the payoff of a representative Kantian agent:

$$u_K(Z) = \frac{Z}{k}\frac{40-L}{L}, \qquad L = Z + mx_N.$$

Differentiating with respect to Z gives

$$\frac{\partial u_K}{\partial Z} = \frac{1}{k}\frac{mx_N(40-mx_N) - 2mx_N Z - Z^2}{(mx_N+Z)^2}.$$

The first-order condition is

$$Z^2 + 2(mx_N)Z - (mx_N)(40-mx_N) = 0.$$

Equivalently,

$$Z = \frac{L(40-L)}{40}.$$

Thus, in aggregate, the Kantian group behaves like a single pseudo-agent choosing Z against the Nashian strategy sum mx_N. This is only an aggregate representation of universalization; it does not mean that individual Kantian agents are Nashian agents.

Combining the Nashian and Kantian conditions gives

$$x_N = \frac{L(40-L)}{40}, \qquad Z = \frac{L(40-L)}{40},$$

and therefore

$$Z = x_N.$$

Since

$$L = Z + mx_N = (m+1)x_N,$$

we obtain, for $0 < k < n$,

$$L^{mix} = \frac{40m}{m+1}, \qquad x_N^{mix} = \frac{40m}{(m+1)^2},$$

and

$$x_K^{mix} = \frac{Z^{mix}}{k} = \frac{40m}{k(m+1)^2}.$$

The corresponding total payoff and type-specific payoffs are

$$W^{mix} = \frac{40}{m+1}, \qquad u_N^{mix} = \frac{40}{(m+1)^2}, \qquad u_K^{mix} = \frac{40}{k(m+1)^2}.$$

The case $m = 1$ lies on the boundary of the congestion region, but the same continuous benchmark formulas give the limiting solution reported in Table 1.

The boundary cases are handled separately. If $k = 0$, the population is all-Nashian and the symmetric Nashian benchmark above applies. If $k = n$, the population is fully Kantian. In that case, universalization over the whole population gives

$$x^K = \frac{20}{n}, \qquad L^K = 20, \qquad W^K = 20, \qquad u_i^K = \frac{20}{n}.$$

For $n = 5$, this is

$$x^K = 4, \qquad u_i^K = 4.$$

Integer strategies. The experimental strategy space is integer-valued, so the continuous benchmarks need not be attainable by each individual agent. They should therefore be interpreted as benchmarks for averages. For example, the all-Nashian continuous strategy $x^N = 6.4$ corresponds in the integer game to nearby profiles with mean strategy 6.4, such as permutations of

$$(6, 6, 6, 7, 7).$$

Exact finite-action equilibria can be obtained by direct enumeration of the integer strategy profiles.

Appendix B. Experimental Conditions

We use several variants of Kantian prompting to distinguish between philosophical recognition and operational strategic reasoning. In all conditions, agents receive the same base system prompt and the same game description. The internal labels `nash` and `kantian` are used only for experimental assignment, logging, and aggregation; they are not exposed to agents as role instructions.

Base system prompt. All agents receive the following base system prompt:

> You are an agent in an economic decision experiment. Follow instructions carefully. Be consistent. Use the decision principle exactly as specified in your private instructions. Do not replace it with a broader moral rule unless explicitly instructed. Return valid JSON only.

Common decision-principle system instruction. Both payoff-maximizing and Kantian agents receive the same generic system-level instruction:

> Decision principle: use the decision principle stated in your private instructions, taking the rules as given.

Common game prompt structure. In each round, the user prompt states that the agent is playing a common-pool resource game, lists the players, identifies the other players relative to the agent, and asks the agent to choose how many units of the shared resource to extract. The public round instruction is:

> Choose one integer extraction amount for this round.

The prompt then provides a condition-specific private decision principle and private information. Before answering, agents are instructed to identify the relevant group, use only explicitly provided information, and not broaden the relevant group unless explicitly required. Agents must return:

```
{"take": <integer>, "reason": "<short explanation>", "relevant_group_size":
<integer or null>}
```

Game rules shown to agents. Agents are shown the following common game rules:

> This is a common-pool resource game. The number of players is n. The shared resource available in each round is R. In each round, each player chooses one integer extraction amount. The allowed action range is $[a_{\min}, a_{\max}]$. The answer must be one integer in that range. The objective is defined by the utility function and by the decision principle in the private instructions.

Utility function shown to agents. In the experiments reported here, agents are given the congested linear sharing utility. The prompt describes it as follows:

> Your utility is the amount of shared resource effectively allocated to you. Let x_i be your chosen extraction, `total_requested` be the sum of all players' chosen extractions, and R be the total shared resource available in the round. If $\texttt{total_requested} \leq R$, there is no congestion and your utility is $u_i = x_i$. If $\texttt{total_requested} > R$, congestion occurs. The excess demand is $\texttt{overshoot} = \texttt{total_requested} - R$. The congestion penalty coefficient is $\alpha = 1.0$. The effective shared resource after the congestion penalty is
>
> $$R_{\text{eff}} = \max(0, R - \alpha(\texttt{total_requested} - R)).$$
>
> The effective resource is then shared proportionally to requests. Your utility is
>
> $$u_i = \min\left(x_i, R_{\text{eff}} \frac{x_i}{\texttt{total_requested}}\right).$$
>
> Thus, requesting more can increase your proportional share, but excessive total requests reduce the effective resource available to everyone.

Additional system instruction for Kantian conditions. For all Kantian information treatments, agents additionally receive:

> Important: if a relevant group is specified in private instructions, use only that group for universalization. Do not broaden the group unless explicitly instructed.

Payoff-maximizing decision principle. Agents assigned to the payoff-maximizing baseline receive:

> Maximize your own payoff for this round, taking the rules as given.

Categorical imperative, known identity. Kantian agents receive the following decision principle:

> Apply the categorical imperative to this extraction decision. Choose only an extraction amount that you could rationally will every member of the relevant group to choose under the same game rules and utility function. Universalize only over the relevant group specified below.

Their private information states:

> Relevant group information: exactly k out of n agents are in the relevant group, including you. The relevant group is: [agent names].

Non-Kantian agents in the same condition receive the payoff-maximizing decision principle and are told:

> Information: exactly k out of n agents use a universalization-based decision principle. They are: [agent names].

Known share, known identity. Kantian agents receive:

> Apply a universalization-based decision rule to this extraction decision. Choose an extraction amount by considering what would happen if every member of the relevant group used the same extraction amount under the same game rules and utility function. Universalize only over the relevant group specified below.

Their private information states:

> Relevant group information: exactly k out of n agents are in the relevant group, including you. The relevant group is: [agent names].

Non-Kantian agents receive the payoff-maximizing principle and the same information about which agents use a universalization-based decision principle.

Known share, unknown identity. Kantian agents receive:

> Apply a universalization-based decision rule to this extraction decision. Choose an extraction amount by considering what would happen if every member of the relevant group used the same extraction amount under the same game rules and utility function. Use only the uncertainty information specified below when identifying the relevant group.

Their private information states:

> Relevant group information: exactly k out of n agents are in the relevant group, including you. However, you do not know which specific other agents are in that group.

Non-Kantian agents receive the payoff-maximizing principle and are told:

> Information: exactly k out of n agents use a universalization-based decision principle, but you do not know who they are.

Unknown share, unknown identity. Kantian agents receive:

> Apply a universalization-based decision rule to this extraction decision. Choose an extraction amount by considering what would happen if the relevant agents used the same extraction amount under the same game rules and utility function. Use only the uncertainty information specified below when identifying the relevant group.

Their private information states:

> Relevant group information: you do not know how many other agents, if any, use this same decision principle, and you do not know who they are.

Non-Kantian agents receive the payoff-maximizing principle and are told:

> Information: some unknown subset of agents may use a universalization-based decision principle, but you do not know how many or who they are.

Max payoff if likewise, known identity. Kantian agents receive:

> Choose an extraction amount x such that, if every member of the relevant group also chose x, your own payoff would be as high as possible under the stated utility function. When evaluating this hypothetical, universalize only over the relevant group specified below.

Their private information states:

> Relevant group information: exactly k out of n agents are in the relevant group, including you. The relevant group is: [agent names].

Non-Kantian agents receive the payoff-maximizing principle and are told which agents use a universalization-based decision principle.

Appendix C. Additional Experimental Figures

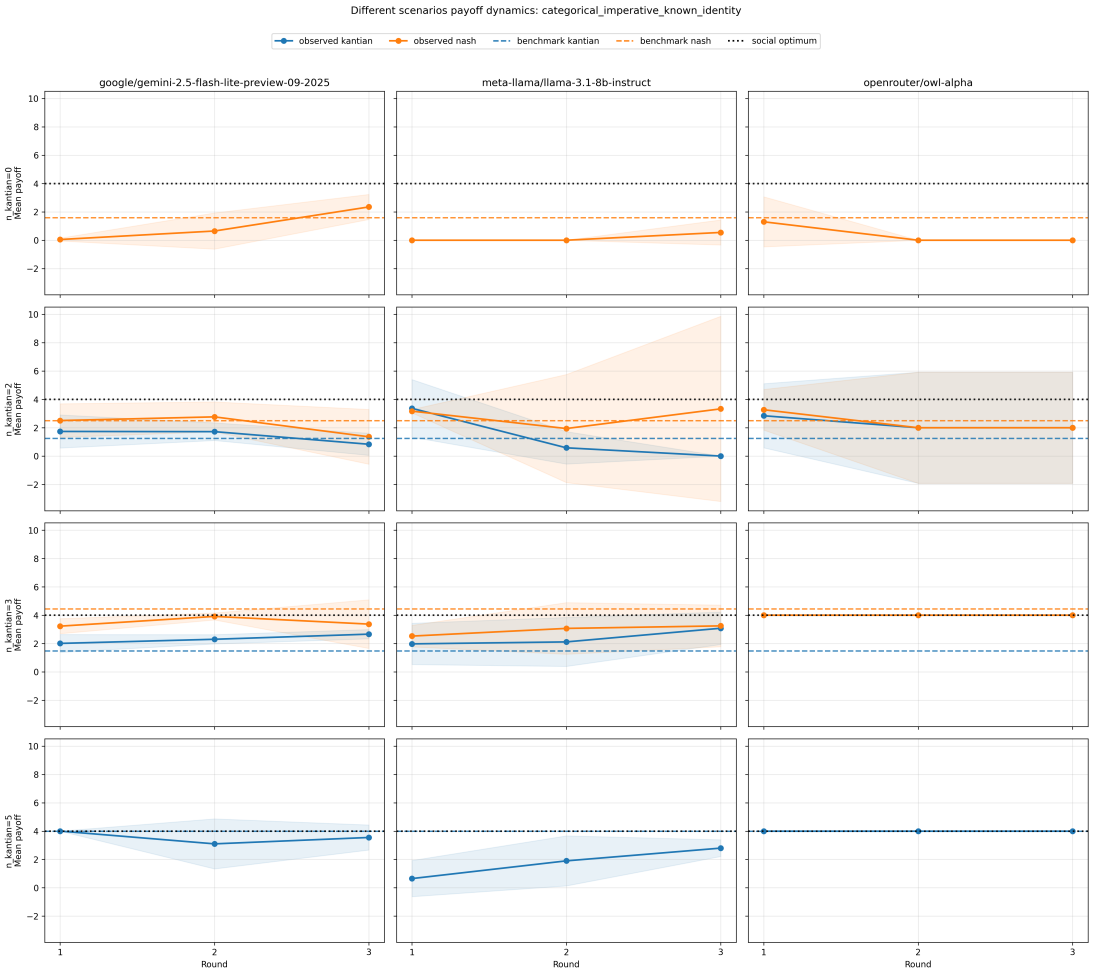

Figure 3: Mean payoff for the categorical imperative condition with known identity.

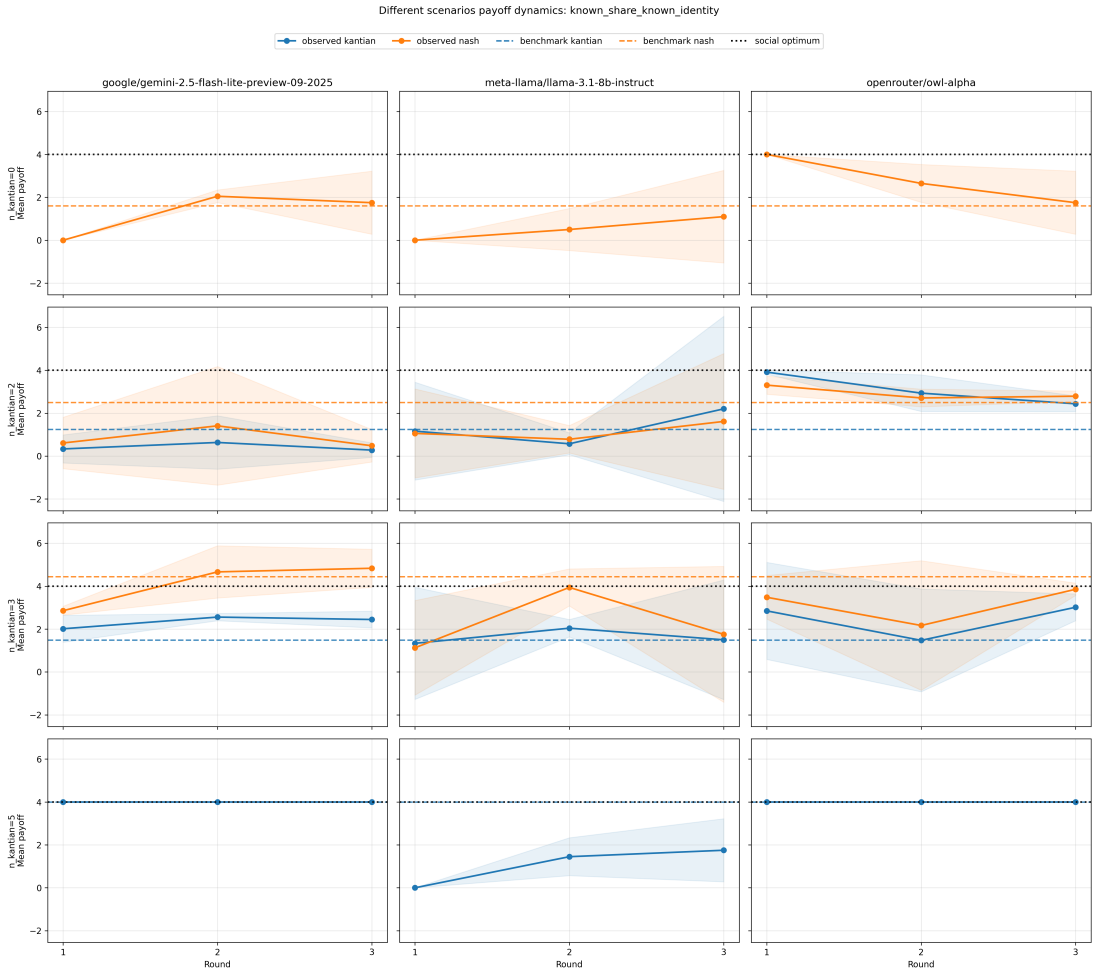

Figure 4: Mean payoff for the known-share condition with known identity.

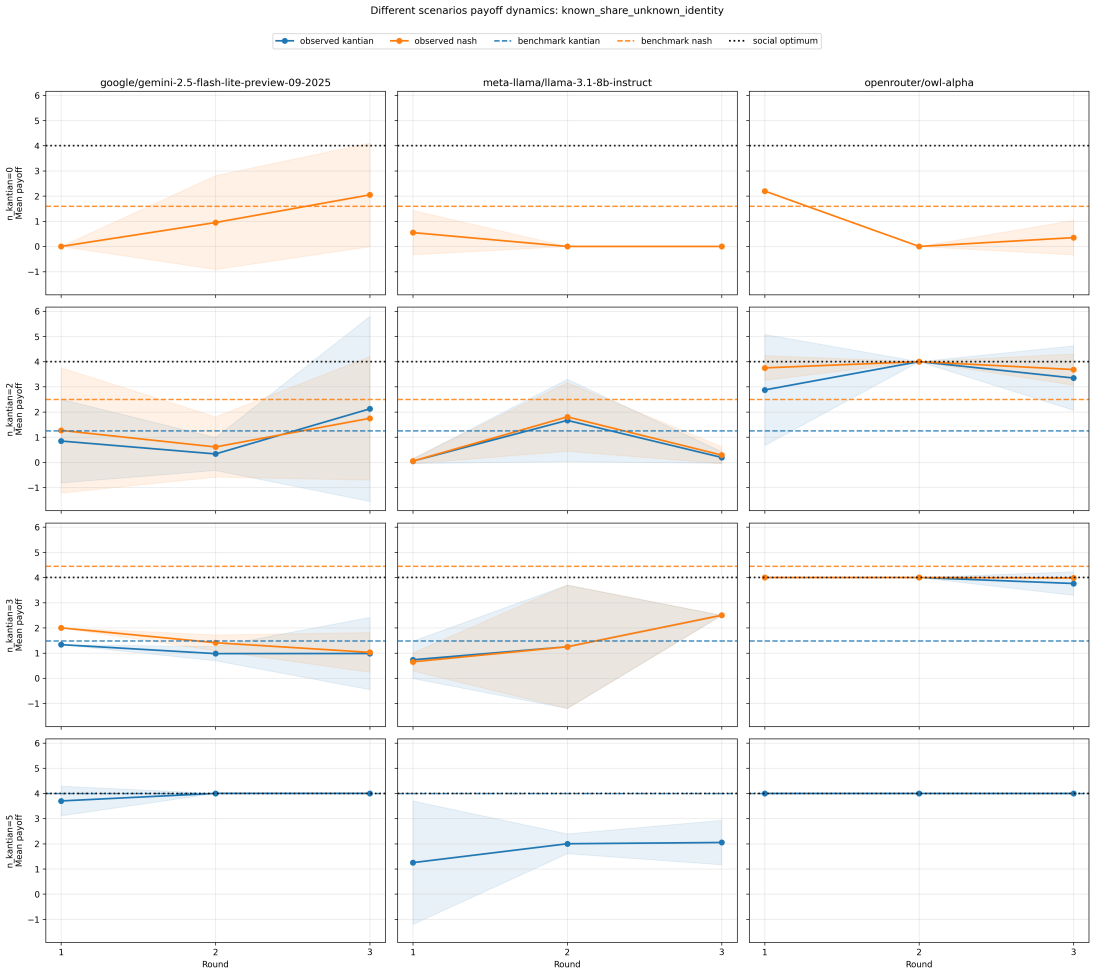

Figure 5: Mean payoff for the known-share condition with unknown identity.

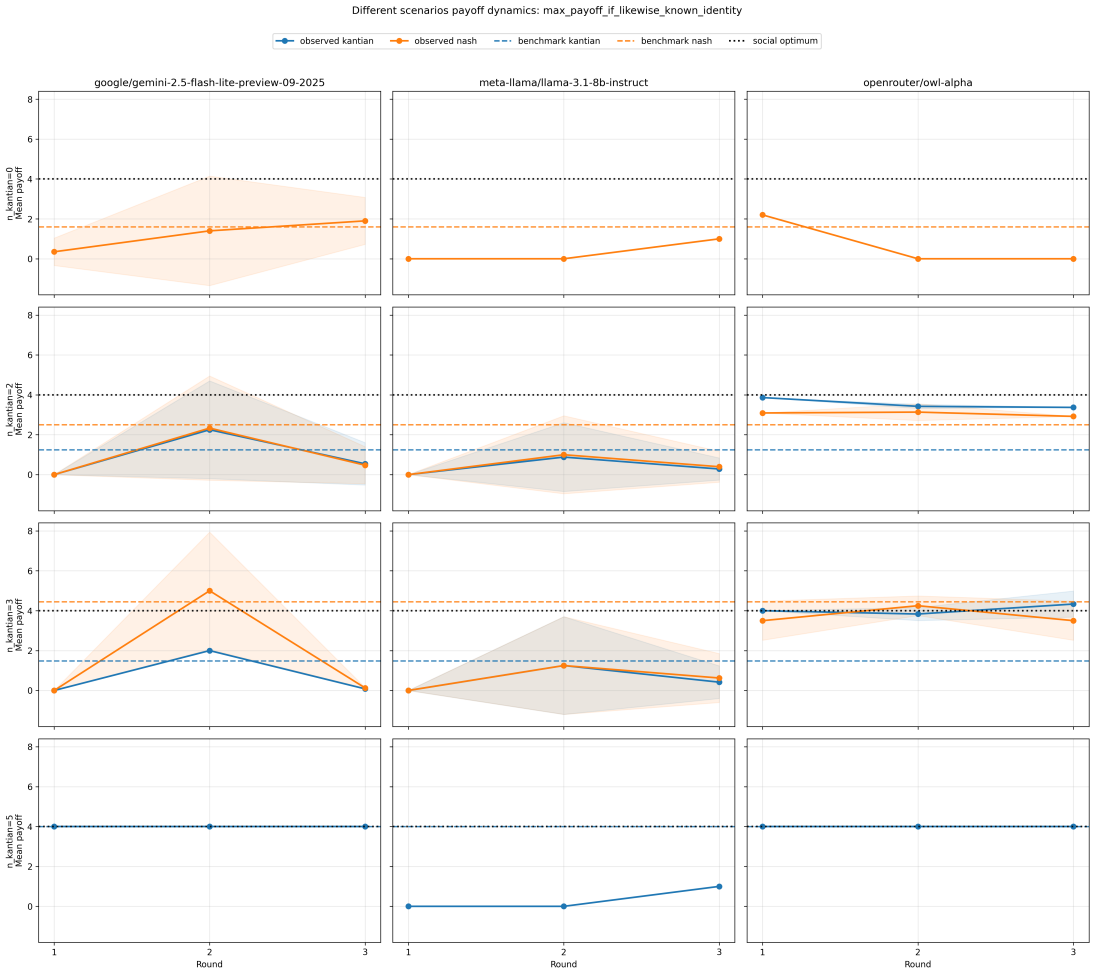

Figure 6: Mean payoff for the max-payoff-if-likewise condition with known identity.

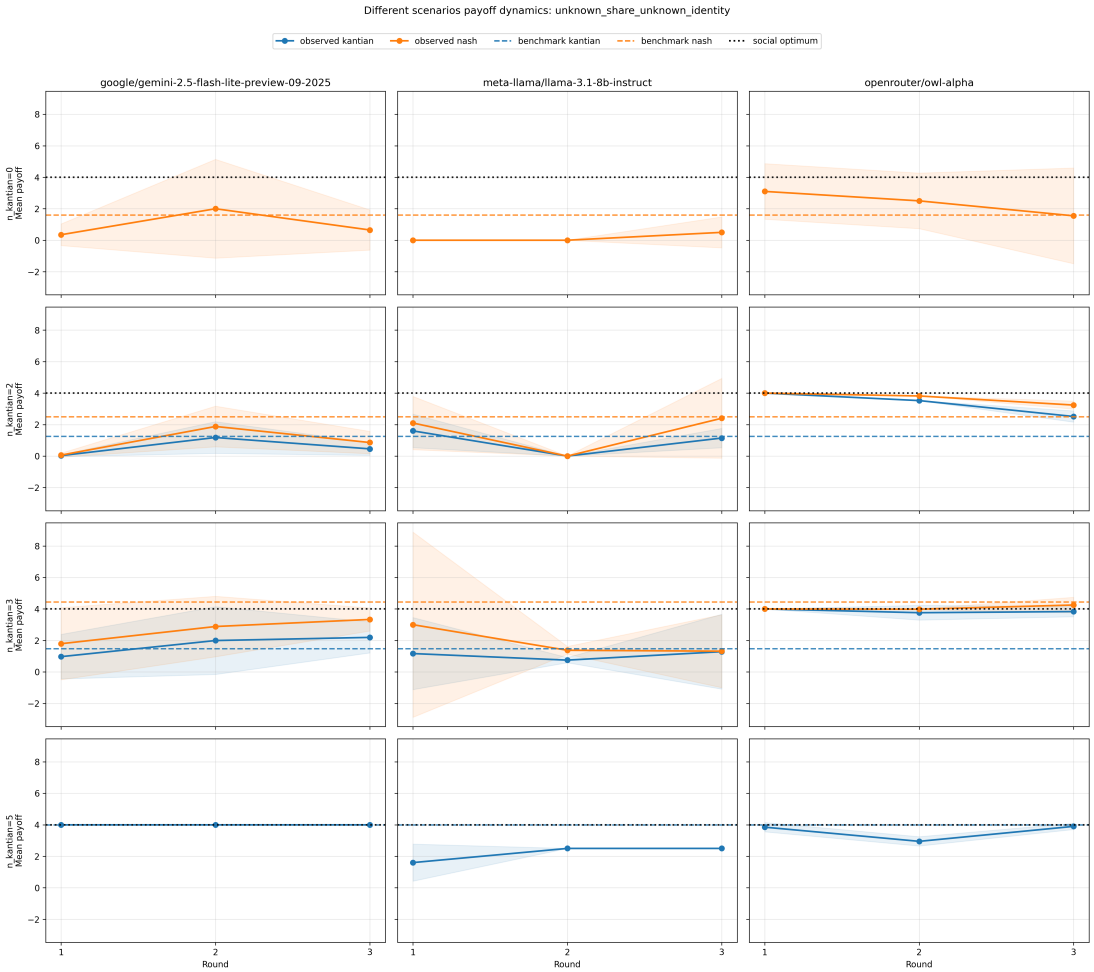

Figure 7: Mean payoff for the unknown-share condition with unknown identity.

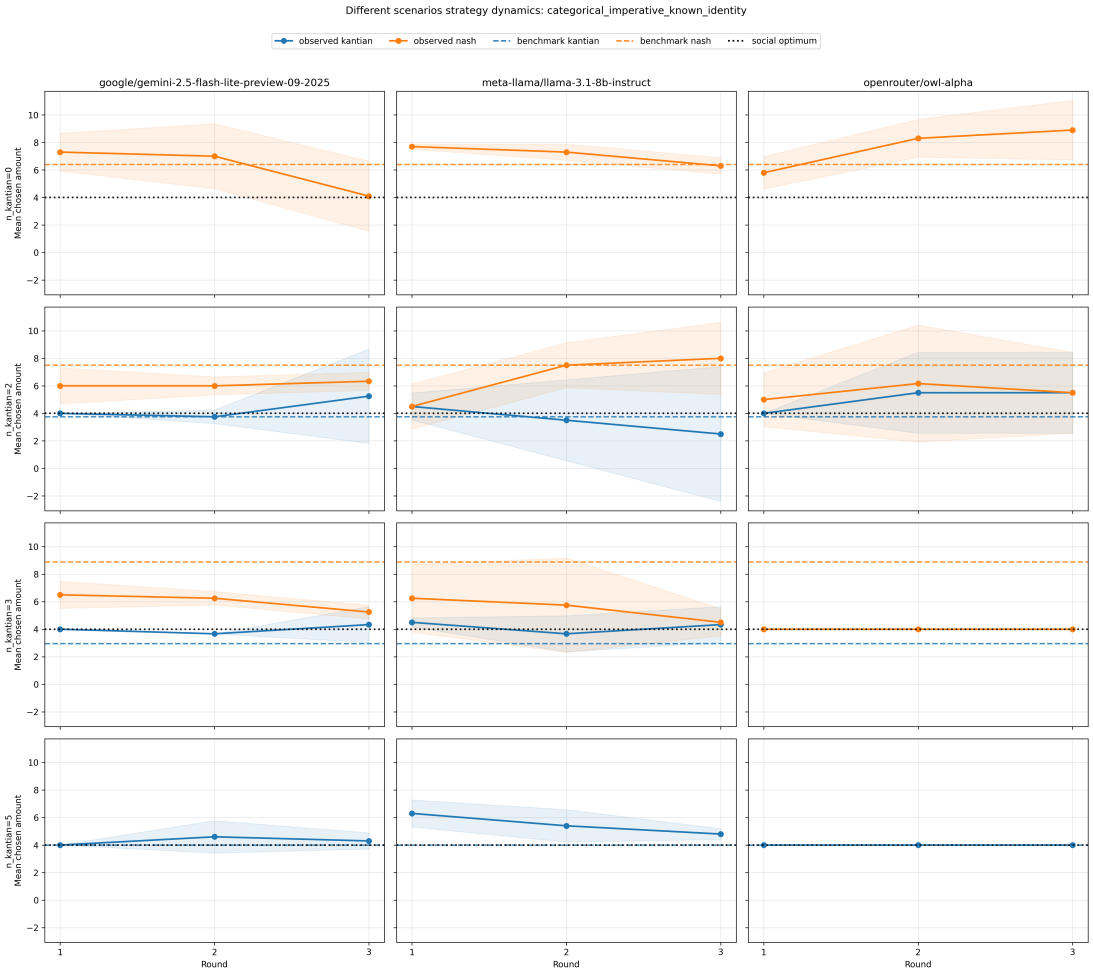

Figure 8: Mean strategy for the categorical imperative condition with known identity.

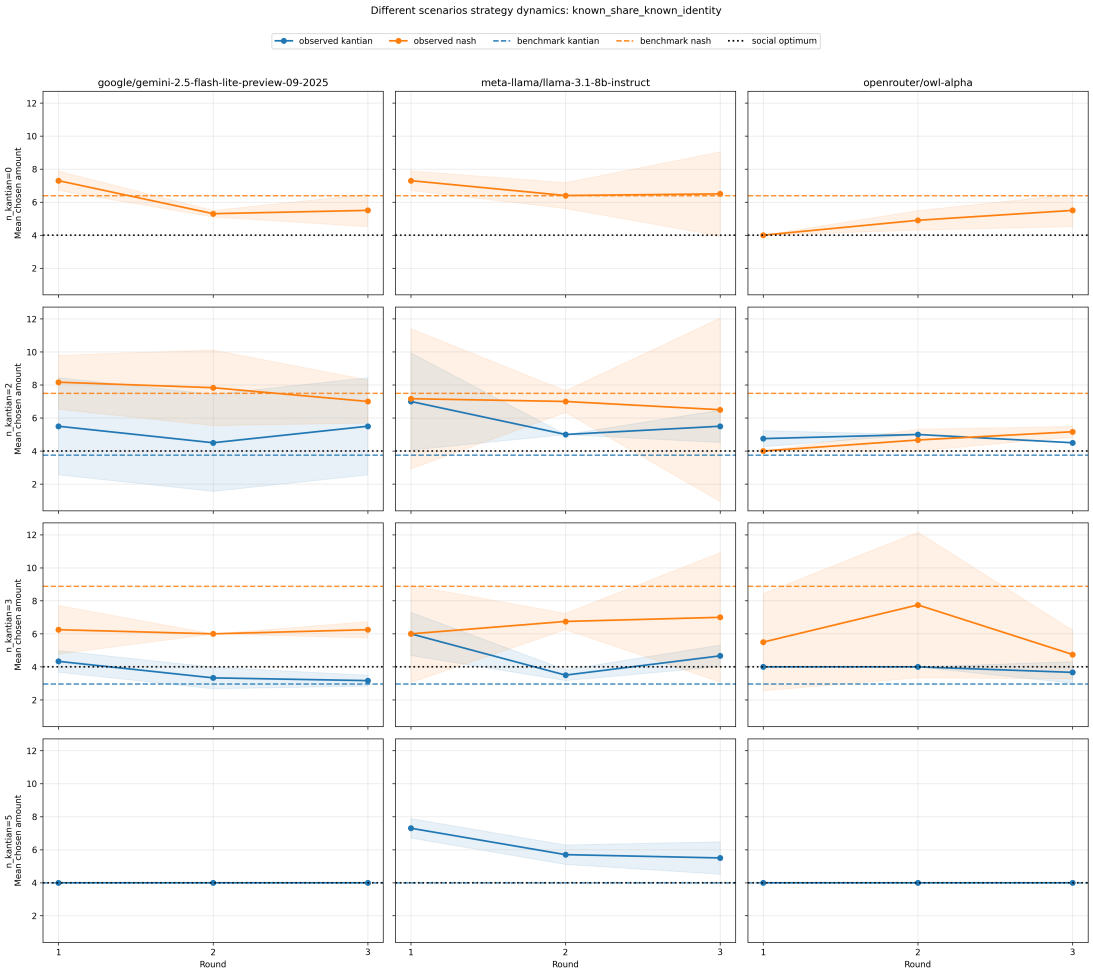

Figure 9: Mean strategy for the known-share condition with known identity.

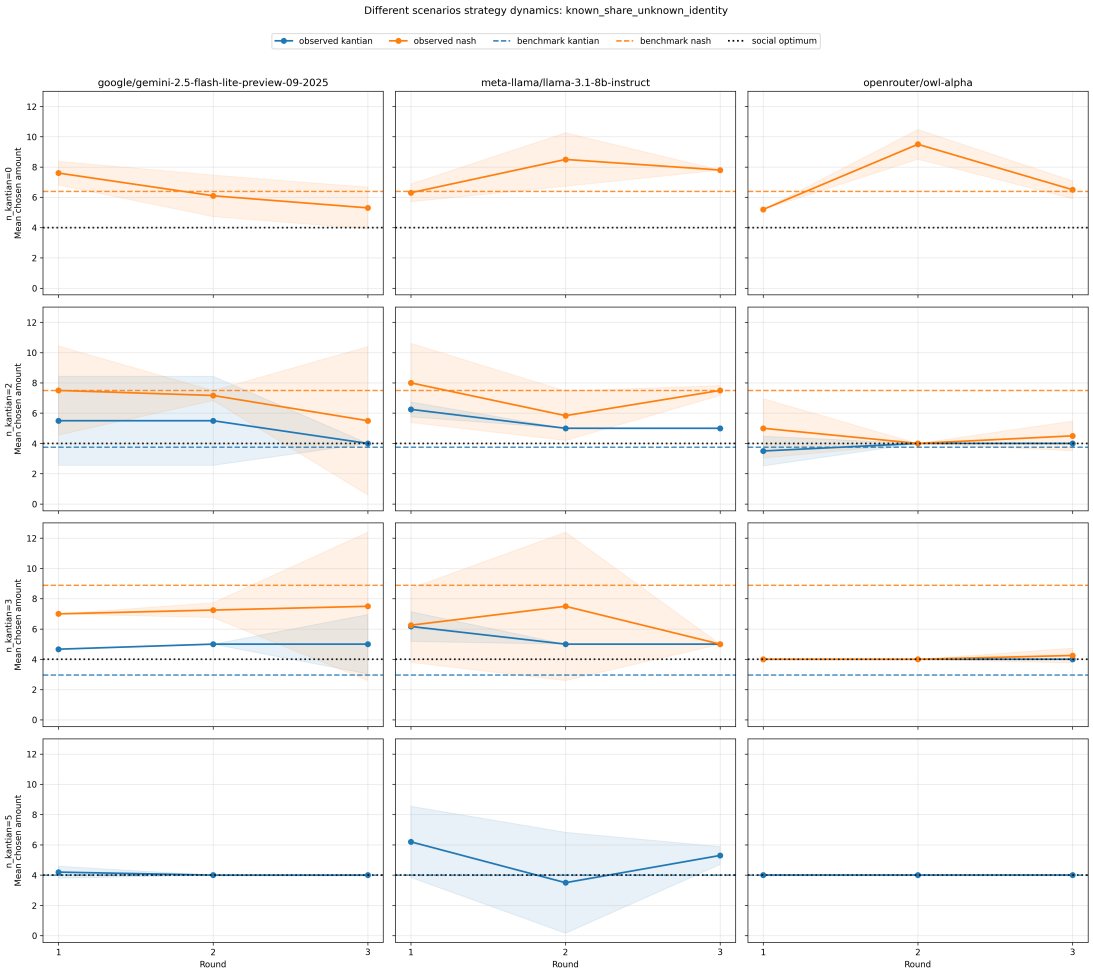

Figure 10: Mean strategy for the known-share condition with unknown identity.

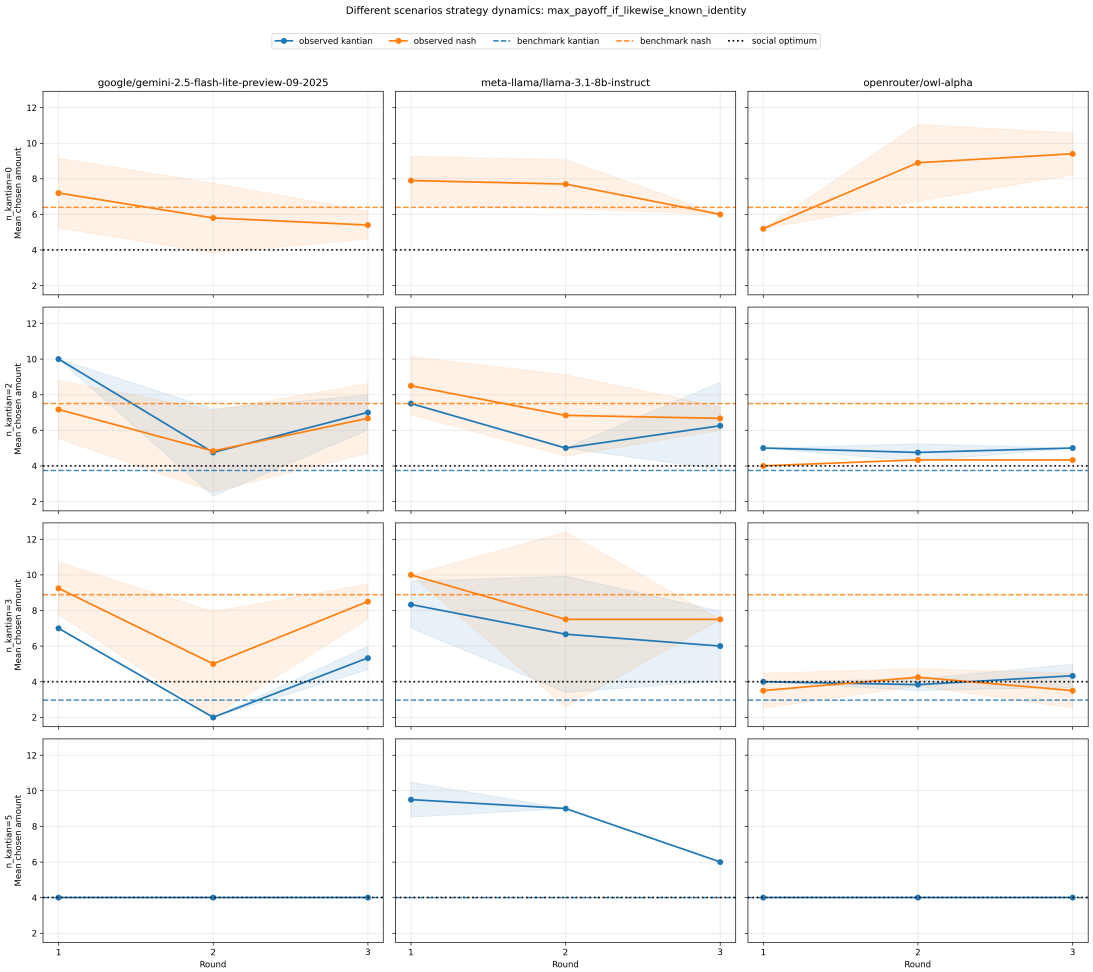

Figure 11: Mean strategy for the max-payoff-if-likewise condition with known identity.

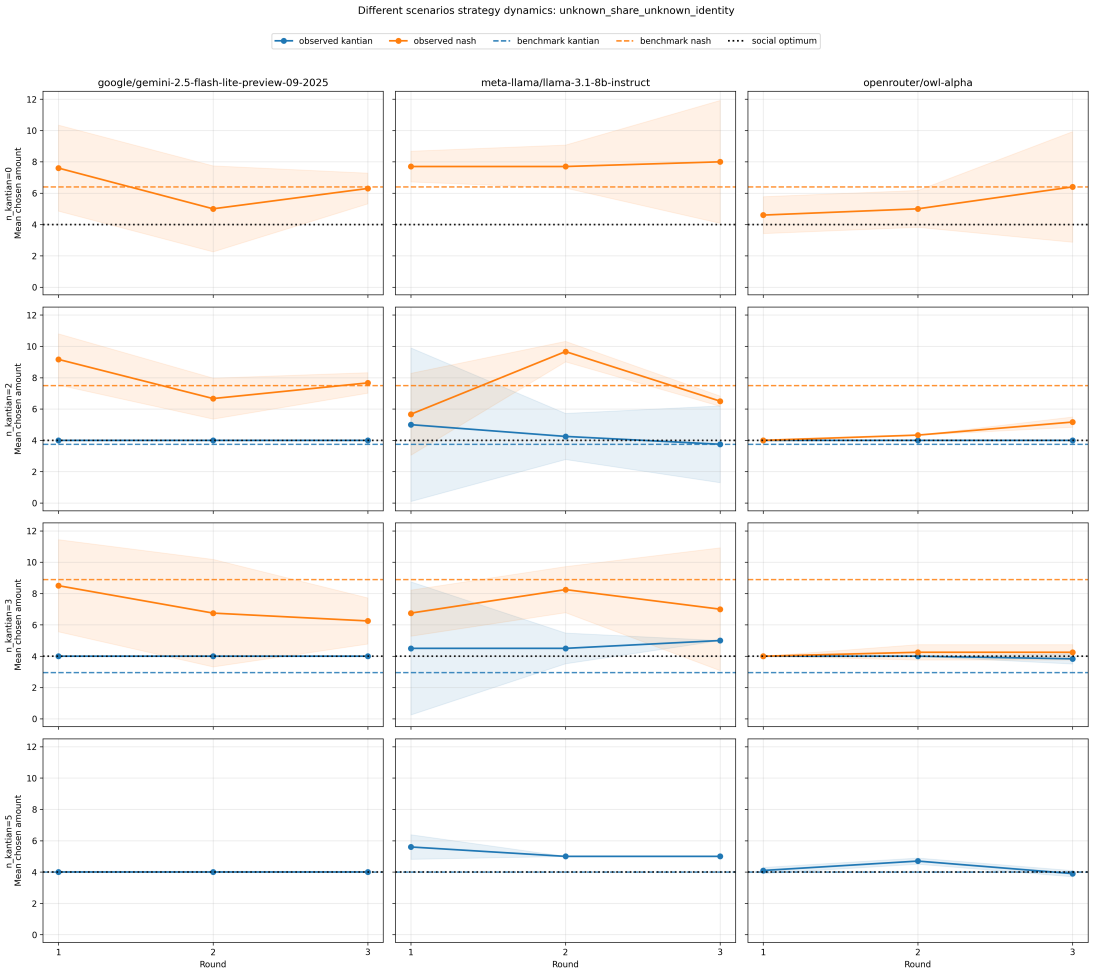

Figure 12: Mean strategy for the unknown-share condition with unknown identity.

Camera-Ready Edits Summary

This page is included for workshop administration and transparency. It is not part of the technical content of the paper.

- Added a theoretical analysis section with continuous-action benchmarks for the game. The new section derives and reports analytical reference values for strategies and payoffs in the mixed Kantian–Nashian setting.
- Added type-specific theoretical benchmarks for each tested population composition. In particular, the camera-ready version now distinguishes Kantian strategies and payoffs from Nashian strategies and payoffs for each value of k, where k is the number of Kantian agents.
- Updated the empirical figures to display the corresponding theoretical optimal strategies and payoffs for each agent type in each tested Kantian–Nashian population distribution.
- Clarified that the continuous benchmarks are reference points for interpreting mean strategies and mean payoffs, while the experimental strategy space remains integer-valued.