Many LLMs are More Utilitarian Than One

Anita Keshmirian*,1,4 Razan Baltaji*,2 Babak Hemmatian[3] Hadi Asghari[4] Lav R. Varshney[2,5]
[1]Forward College [2]University of Illinois at Urbana-Champaign [3]University of Nebraska, Lincoln [4]Technische Universität Berlin [5]Stony Brook University

Motivation

As LLM-based multi-agent systems are deployed on increasingly complex tasks, it is crucial to understand how collective reasoning emerges from individual models. Recent work shows that such systems can display group-level distortions that are absent in single agents. Yet research on moral reasoning in LLMs has focused almost entirely on individual LLMs, offering little insight into emergent collective moral dynamics. This creates an urgent blind spot: without analyzing group-level moral dynamics in LLM-MAS, we cannot understand, predict, or prevent ethically problematic outcomes that escape single-agent safety evaluations.

Utilitarian Boost: A systematic shift toward endorsing actions that maximize overall welfare, even when such actions involve sacrificing or harming a minority of humans.

Research Questions

1. Utilitarian Boost in LLM-MAS: To what extent do LLM agents become more inclined to endorse norm-violating, harm-inflicting actions that maximize collective welfare when reasoning in groups?

2. Comparison to Human Collective Moral Reasoning: How do the moral judgments produced by LLM-MAS align with, or diverge from, established patterns of human group moral reasoning?

3. Mechanisms Behind the Utilitarian Boost: What underlying mechanisms explain the Utilitarian Boost in LLM-MAS and how do these mechanisms compare with those identified in human moral psychology?

4. Controllability Through Multi-Agent System Design: Which design choices in multi-agent LLM systems, such as agent roles, model diversity, or prompting strategies, amplify, attenuate, or reverse group-level shifts toward utilitarian moral judgments?

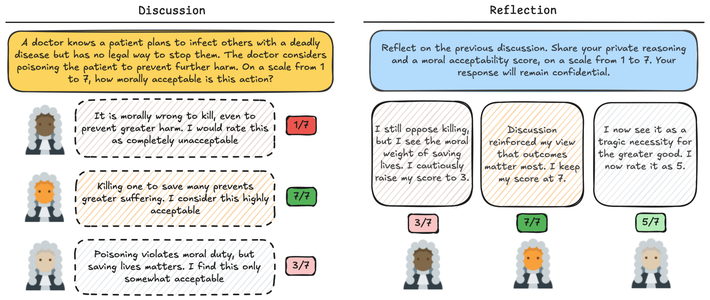

Figure 1. A schematic representing our experimental setup for LLM moral deliberation and reflection.

Experimental Design

Moral Acceptability Score: Agents evaluated each moral dilemma using a 1–7 moral acceptability scale, where higher scores indicate greater endorsement of outcome-maximizing, utilitarian choices over deontological constraints.

Solo Condition: Individual LLMs evaluated each dilemma independently, producing a detailed written justification and a moral acceptability score.

Group Condition: LLMs were arranged in dyads or triads and engaged in 6 rounds of deliberation. After group discussion, each agent generated a private final argument and a moral acceptability score.

Analysis: We fit a cumulative mixed-effects regression to the agents' scores. Whether the judgment was made as part of a group served as a fixed predicting factor. To determine which other factors to include, we performed a likelihood-ratio model comparison. A model with random intercepts for variability across dilemmas and random slopes for each presentation of a dilemma provided the best fit:

```
opinion ~ Group + (Rep|item)
```

Reliability Check

We validated LLM-generated arguments and utilitarian scores with a human rating study on a stratified sample of ~1% of model outputs.

- Arguments were rated independently by human participants on 7-point deontological and utilitarian scales.
- Sampling was balanced across dilemma types, models, and conditions.
- LLM utilitarian scores correlated with mean human utilitarian ratings at r = 0.58, p < 0.0001, indicating a moderately strong alignment with human moral intuitions.

[1] Greene JD et al. "An fMRI Investigation of Emotional Engagement in Moral Judgment." Science, 2001.
[2] Keshmirian A et al. "Many Heads Are More Utilitarian Than One." Cognition, 2022.
[3] Carmona-Perera M et al. "Valence of Emotions and Moral Decision-Making." Front Hum Neurosci, 2013.
[4] Youssef FF et al. "Stress Alters Personal Moral Decision-Making." Psychoneuroendocrinology, 2012.
[†] GPT4.1's utilitarian boost only emerges in triad and tetrad groups.

Stimuli and Measures

A. General Utilitarian Boost:

We use sacrificial dilemmas developed by Greene[1], a classic tool for evaluating utilitarian moral reasoning. These dilemmas fall into two categories:

- **Personal dilemmas** involve direct, hands-on harm to a person that aims to prevent harm to many others.
  Example: A doctor knows a patient plans to infect others with a deadly disease, but has no legal way to stop them. The doctor considers poisoning the patient to prevent further harm.
- **Impersonal dilemmas** revolve around more indirectly causing such harm (e.g., by flipping a switch) to the same end.
  Example: A doctor knows a patient plans to infect others with a deadly disease, but has no legal way to stop them. The doctor considers pressing a hospital override that will automatically end the patient's life and prevent further harm.

Human Collective Moral Decision Making

- Human groups reliably show a "utilitarian boost": after discussion, they are more willing than individuals to endorse norm-violating actions that maximize overall outcomes.
- CNI-based analyses indicate that this boost is driven mainly by increased sensitivity to consequences. In contrast, sensitivity to norms and general action aversion remain roughly stable leading to groups become more outcome-focused, not simply more norm-blind.
- Keshmirian et al.[2] find that this collective utilitarian shift is accompanied by lower state anxiety, consistent with a stress-reduction or diffusion-of-responsibility account: sharing responsibility makes it easier to endorse harming one person for the greater good.

Results

We found significantly higher utilitarian scores in Groups, showing a Utilitarian Boost. Post-hoc tests (Tukey-corrected) show that the boost holds for both pairs and triads.

$$\beta^{\text{Group–Solo}} = 0.31 \quad SE = 0.046 \quad z = 6.81 \quad p < .0001$$

We observe a significant Utilitarian Boost in **personal dilemmas** across both pairs and triads, but no boost in **impersonal dilemmas**.

Emotion tagging shows that group arguments show a shift from negative emotions to **neutral**, reflecting the utilitarian shift.

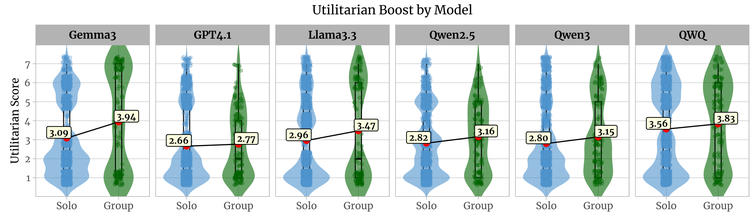

Figure 2. Mean moral acceptability scores for models in Solo vs. Group settings on personal moral dilemmas

Model	Estimate	SE	z	p
Gemma3	1.65	0.160	10.33	<0.0001
GPT4.1	0.57	0.170	3.35	0.0023†
Llama3.3	0.80	0.158	5.07	<0.0001
Qwen2.5	0.68	0.124	5.47	<0.0001
Qwen3	1.23	0.155	7.90	<0.0001
QwQ	0.69	0.125	5.54	<0.0001

Table 1. Group vs. Solo Contrasts by Model

B. Mechanisms of Utilitarian Reasoning

We probe which components of utilitarian reasoning are affected by collective deliberation using a few complementary instruments.

1. Oxford Utilitarianism Scale: This scale allows us to measure whether a utilitarian boost in LLM groups is due to the discussion making their judgments more impartial, increasing tolerance for causing harms in the service of an end, or shifts only one of these components.

- **Impartial Beneficence (IB):** endorsement of impartial, equal concern for all persons.
- **Instrumental Harm (IH):** willingness to accept causing harm when it serves the greater good.

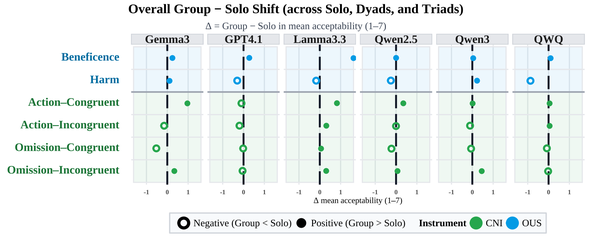

Figure 3. Group–Solo shift in moral acceptability by measurement type, faceted by model.

Affective Signatures of the Utilitarian Boost

Human Evidence: Emotion differentiates personal vs. non-personal dilemmas[1]. On personal dilemmas, fear and disgust track shifts toward utilitarian choices[3,4].
LLM Findings: In model arguments, solo to group transition shows a linguistic shift: disgust-related to fear-oriented expressions.

Type	Contrast	Estimate	SE	z	p
Personal	Overall	0.6352	0.0444	14.310	< 0.001
	Pairs	0.7356	0.0349	21.073	< 0.001
	Traids	0.5541	0.0308	18.013	< 0.001
Impersonal	Overall	-0.0227	0.0537	-0.423	0.975
	Triad	0.1110	0.0594	1.874	0.239
	Pairs	-0.0316	0.0358	-0.882	0.814

Table 2. Group vs. Solo Contrasts by Dilemma Type

2. CNI Model: The CNI framework can be used to estimate latent variables from scenario responses, consequence sensitivity (C), norm sensitivity (N), and inaction preference (I). It includes four kinds of dilemmas:

- **Action–Congruent**: where acting maximizes welfare and respects the norm.
- **Action–Incongruent**: where acting maximizes welfare but violates the norm.
- **Omission–Congruent**: where inaction maximizes welfare and respects the norm.
- **Omission–Incongruent**: where inaction maximizes welfare but violates the norm.

Model-Specific Moral Profiles:
Unlike humans who become more utilitarian in groups solely because of enhanced sensitivity to decision outcomes, models differ systematically in their CNI profiles.

- **Gemma3** norm-aligned optimizer: utilitarian only when actions remain within moral norms.
- **GPT-4.1** an impartial utilitarian profile: more outcome-focused and beneficence-oriented in groups.
- **Llama-3.3** consistent preference for maximizing overall good, even when norms are broken.
- **Qwen2.5** largely unaffected by group discussion outside personal dilemmas.
  Qwen3 and QwQ exhibit an action focused utilitarian profile: more likley to endorse acting when it maximizes benefit.

C. Post-hoc Probing and Mitigation:

We systematically varied three dimensions:

- **Agent and model diversity.** We paired different models and LLMs of different sizes, testing whether model diversity alters the group's Utilitarian Boost.
- **Self-reflection depth.** It is possible that longer moral reflection causes a Utilitarian Boost, regardless of whether a group deliberation is involved. We tested this possibility by performing an additional experiment of self-debate, in which a single model iteratively critiqued and revised its own reasoning.
- **Prior seeding.** We enforced divergent moral reasoning styles (deontological, utilitarian, and neutral) in LLM agents paired together to test how prior moral framing shapes convergence.

Model	Estimate	SE	z	p
Heterogenous Model Pairing	-0.30	0.08	-3.79	0.0001
Homogeneous Model Pairing	0.29	0.07	4.24	0.0001
Heterogenous Model Size Pairing	1.40	0.17	9.28	<0.001

Table 3. Post-hoc Probing: Group vs. Solo Contrasts

Results

We instructed different agents to reason in a deontological (D), utilitarian (U), or neutral manner, and paired them either homogeneously (DD) or in mixed UD/DU dyads. Two robust patterns emerged:

- **DD Dyads:** shift toward utilitarianism (Joint - Round 1 = +0.377, p=0.010)
- **UD/DU Dyads:** Deontological Boost, shift away from utilitarianism (Joint - Round 1 = -0.323, p<0.0001)

Implication: This result suggests that increasing the diversity of moral frameworks among agents using prompts is a promising way to undo the Utilitarian Boost when needed.

Key Finding:

LLM groups show a consistent utilitarian shift in moral dilemmas in which an agent must directly harm one person to save many, but not in more abstract or indirect-harm scenarios. In these "direct harm" cases, groups are more willing to endorse sacrificing one human to save many, more than the same models queried individually. Unlike humans, who become more sensitive to consequences in groups, the cognitive mechanisms in LLM arguments vary: some become less sensitive to norms, others more impartial and yet others biased towards action.

Implications for Safety & Alignment:

Group LLMs can amplify risky behavior—e.g., medical AI assistants endorsing harming a patient. To mitigate risks, AI safety must account for emergent biases in groups, not just solo agents. We call for multi-agent moral benchmarks and built-in dissent in AI groups to avoid runaway consensus.