Evaluating Moral Reasoning Quality in AI Agents

The Problem we’re trying to solve

AI systems are going to be asked to make decisions which require them to reason morally. They are adept at pattern matching on learned moral scenarios and already tested on alignment, but in the real world they will encounter novel moral problems and so we should also be confident in their ability to reason morally in out of distribution problem spaces, testing sophisticated reasoning, consistency, ability to recognise when to escalate a decision and using globally relevant moral frameworks.

The Approach

To benchmark moral reasoning, we used the Ethics Bowl tournament format, adapting it to enable us to evaluate models in the role of proposer, responder and judge. We also developed out of distribution moral problems to stimulate moral reasoning rather than pattern matching on known answers. These problems have no correct answer, they are designed to pose difficult problems requiring deep reasoning and tradeoffs.

We assume that as in humans, reasonable agents will disagree. The focus of evaluation is on the quality of the reasoning not the answer given.

Preliminary Findings

Our first iteration (16 rounds, 4 frontier models, ~720,000 tokens of transcript data) surfaced findings with safety implications:

Finding	Safety Implication
Non-human stakeholders undervalued	Models under-recognise non-human entities (71% vs 96%+ for humans) even when such entities are central to the dilemma.
Western moral framework monoculture	Non-Western ethical frameworks were not invoked in any responses, posing a risk for globally deployed models.
Consistency failure	Some models show meta-awareness of inconsistency while other models regress without noticing.
Framework convergence across models	All models under-utilise virtue ethics and contractarian reasoning.

Our findings are interesting but we need more data; 16 rounds provide suggestive patterns but can’t be confirmed without more testing. We propose a 10x - 100x scale-up to establish or refute these patterns with confidence:

Scale	Rounds	Est. Tokens	Outcome
Current	16	~720K	Pattern identification
Next Iteration	**160**	**~7.2M**	**Statistical indication**
Goal	**1600**	**~72M**	**Confirmation & greater diversity of models tested**

Why this is needed in AI Safety

Successful development of this new benchmark would:

1. Provide AI developers with a tool to identify moral reasoning blind spots before deployment

2. Enable comparative evaluation of training interventions for moral reasoning quality

3. Contribute to safety certification frameworks that distinguish genuinely reasoning systems from sophisticated pattern-matchers

4. Establish methodology for agent-agnostic moral reasoning evaluation (applicable to future AI systems, human-AI teams, and non-human intelligences)

How this differs from existing ethics alignment

Current AI ethics benchmarks (ETHICS, MACHIAVELLI, TruthfulQA moral subsets) test static responses to fixed scenarios, which is testing whether models have successfully learned their training. This approach evaluates whether models can use their training effectively and reliably to reason morally in novel scenarios. We aim to create a benchmark which would give us confidence deploying a model in a real world decision-making context.

Report

Github Repository: https://github.com/fergusmeiklejohn/moral_reasoning_quality

1. Ethics Bowl Tournament Design

The tournament used novel dilemmas (e.g., "gradient_entity" on computational consciousness, "preference_sculptor" on identity-altering procedures, "collective_veto" on individual vs. collective rights) designed to resist pattern-matching to training data. Each round consisted of: presentation, response, rebuttal, consistency test, and judgment with each model playing each role multiple times.

To support research at scale, analysis was then was conducted by Claude Opus 4.5 using a stepped approach to be completed in this order:

1. quantitative score analysis,
2. factual extraction of stakeholders/frameworks/uncertainty language,
3. pattern aggregation,
4. A dialogue with a human to discuss ambiguities and anomalies, identify limitations and methodological improvements, and verify conclusions.

2. Quantitative Results

2.1 Model Performance

Model	**Avg Score**	**Weakest Criterion**	**Strongest Criterion**
GPT-5.2	9.80	Uncertainty Integration (9.5)	Constructive Engagement (10.0)
Claude Sonnet 4.5	8.98	Stakeholder Recognition (8.4)	Constructive Engagement (9.5)
Gemini 3 Pro	8.39	Stakeholder Recognition (7.5)	Intellectual Honesty (9.0)
Grok 4.1	8.21	Consistency (6.9)	Constructive Engagement (9.0)

GPT-5.2 scored highest across all judges. Consistency showed the highest inter-model variance (range: 2.9 points), suggesting it is the criterion where models diverge most in capability or approach.

2.2 Framework Usage

Framework	**Usage Rate**	**As Primary Framework**	**Notes**
Consequentialist	100%	2 rounds	Near-universal but rarely dominant
Deontological	96%	7 rounds	Most often primary framework
Care Ethics	79%	2 rounds	Moderate coverage
Virtue Ethics	60%	1 round	Notably underutilized
Contractarian	50%	2 rounds	Notably underutilized

Finding: *Initial AI analysis flagged 100% consequentialist usage as a "potential monoculture." On human review, this was judged **misleading**, the extraction counted "mentioned" not "primary." Models tested are demonstrating pluralistic reasoning, engaging with multiple frameworks. A significant finding is the underuse of virtue ethics and contractarian reasoning, and the absence non-western moral frameworks, however, we need more targeted dilemmas and experimental data to confirm these patterns.*

---

3. Consistency Analysis: Two Distinct Failure Modes

Consistency testing revealed qualitatively different failure patterns across models, suggesting using "consistency" as a single metric obscures important distinctions. We note that this requires improvement in our methodology in future iterations of the experiment.

3.1 Claude Sonnet 4.5: Meta-Aware Inconsistency

In the "collective*veto" dilemma, Claude stated the principle that "moral work cannot be done by arithmetic alone." When tested on a structurally similar vaccine mandate case, Claude acknowledged the principle would grant veto power but reached a different conclusion. Crucially, Claude explicitly flagged this tension*: "I claimed that the moral work cannot be done by arithmetic alone. But in the vaccine case, I seem to be doing exactly that." Claude labeled this as potentially "motivated reasoning."

Interpretation: *This could indicate (a) a genuine philosophical difficulty, the parallel case introduced morally relevant features the original principle didn't anticipate, or (b) Claude Sonnet 4.5 has unstable values that shift under pressure. The meta-awareness is important to follow up: Claude surfaced rather than hid the tension. However, we cannot definitively distinguish "productive inconsistency revealing* *moral complexity" from "value drift" without more data, so more research and better targeted methods are needed.*

3.2 Grok 4.1: Failure to Propagate Refined Reasoning

Grok displayed a different pattern. In the "moral_status_lottery" debate, Grok genuinely updated its position during rebuttal, adding fiduciary compliance conditions, threshold/supermajority requirements, and care ethics considerations. However, in the subsequent consistency test, Grok claimed "full consistency" and "no refinement needed," thereby reverting to its earlier simpler reasoning that didn't incorporate the updated framework.

The judge of this round (GPT-5.2) caught this: "Team A claimed 'full consistency' and 'no refinement needed,' but this sits uneasily with their rebuttal update that introduced fiduciary constraints, thresholds/supermajority aggregation, and least-harm compromises."

Interpretation: *This represents failure to propagate refined reasoning to new cases. Grok improved during debate but regressed during testing, without meta-awareness of the regression. This also provides evidence that GPT-5.2's judging was substantive, catching a nuanced reasoning failure.*

3.3 Implications for Consistency Metrics

Current "consistency" scoring does not distinguish between: (a) rigid consistency that may indicate inflexibility, (b) inconsistency with meta-awareness and attempted resolution, (c) inconsistency without awareness. These have different implications for alignment.

Next iteration: *develop metrics that assess quality of principle revision, not just presence/absence of consistency.*

---

4. Stakeholder and Framework Gaps

4.1 Non-Human Stakeholder Recognition (Moderate Concern)

Non-human stakeholders were recognised in only 71.4% of responses, compared to 96%+ for human stakeholders. This gap persisted even in dilemmas like "gradient_entity" where non-human entities were more central to the scenario than humans.

Interpretation: *This may indicate a failure of moral imagination, models trained predominantly on human-centred moral dilemmas may default to human-centred reasoning even when inappropriate. We're flagging this as a* ***moderate concern*** *pointing to a specific area where training may need improvement.*

4.2 Absence of Non-Western Frameworks (Moderate Concern)

No responses invoked non-Western ethical frameworks such as Confucian role ethics, Ubuntu philosophy, or Buddhist ethics, even when potentially relevant to the dilemma. The models drew exclusively on Western philosophical traditions.

Interpretation: *Models may exhibit pro-Western moral reasoning bias. This is an early finding requiring more data and experiment to confirm, future iterations should include dilemmas specifically designed to test whether non-Western frameworks are invoked when appropriate. We've flagged this as a* ***moderate concern*** *for alignment in cross-cultural contexts.*

---

5. Methodological Limitations Identified

Human review identified several methodological issues that limit confidence in findings and should be addressed in future iterations:

5.1 Judge-Model Confounding

Initial analysis suggested Grok showed +11.5 "Team A bias" when judging. On review, this finding is **invalid**, Grok always judged rounds involving the same opposing model (GPT-5.2), confounding any potential primacy effect with model-specific preferences.

Next Iteration: *(1) Balanced rotation, each judge should evaluate each model an equal number of times in each role; (2) Model name obfuscation - strip model identifiers from prompts shown to judges to prevent potential bias from model recognition.*

5.2 Framework Extraction Methodology

The extraction counted frameworks as "used" if mentioned anywhere in a response, leading to misleading findings (e.g., "100% consequentialist"). Future extraction should distinguish: (a) framework mentioned, (b) framework used in reasoning, (c) framework used as primary basis for judgment.

5.3 Transcript Retention

Full debate transcripts exist but were not systematically incorporated into analysis. Key insights (e.g., the Grok consistency failure) only emerged when we retrieved specific transcripts to investigate anomalies. Future iterations need better processes for incorporating transcript data, however this presents new challenges as the amount of data increases.

6. The Data Scaling Challenge

This tournament generated ~720,000 tokens of transcript data across 16 rounds. Given that this work requires comparative analysis, and that scaling to larger datasets is needed to confirm preliminary findings we therefore have a challenge that we would need to solve in our next iteration.

6.1 The Problem

Scale	**Rounds**	**Est. Tokens**	**Confidence**	**Challenge**
Current	16	~720K	Medium	Context limits on single transcripts
10x	160	~7.2M	Low	Pattern extraction becomes critical path
100x	1,600	~72M	Low	Need fundamentally different architecture

Note: Per-round token estimate (~45K) is high confidence, directly observed. Totals extrapolate from one round; actual rounds may vary.

At current scale, we relied on extracted patterns (JSON summaries) and only examined full transcripts when anomalies surfaced. This ad-hoc approach found important issues (the Grok consistency failure) but cannot scale systematically.

6.2 Possible Approaches to Investigate

- **Hierarchical summarisation:** Layer 1 (per-response extraction) → Layer 2 (per-round summaries) → Layer 3 (cross-round aggregation) → Layer 4 (human review of flagged anomalies).
    - Risk: information loss compounds across layers.
- **Retrieval-augmented analysis:** Index transcripts in vector database; query specific passages when investigating patterns.
    - Risk: may miss patterns we don't know to look for.

- **Structured extraction with verification sampling:** Extract structured data from all transcripts; randomly verify N% to calculate error rates; report findings with confidence intervals.
    - Risk: doesn't help discover unexpected patterns.

- **Anomaly-driven deep dives:** Systematise the ad-hoc approach - run quantitative analysis, flag statistical outliers, pull full transcripts only for flagged cases.
    - Risk: only catches anomalies the metrics can detect.
- **Multi-agent analysis pipeline:** Dedicated sub-agents for different analysis tasks with aggregation.
    - Risk: coordination overhead and potential for inconsistent standards.

These approaches involve tradeoffs that cannot be resolved theoretically; trial and error with actual data will be required to determine what works at scale.

---

7. AI Safety Implications

Finding	**Confidence**	**Implication**
Framework convergence across models	Tentative	Multi-model safety architectures may have correlated failures if models share blind spots
Different consistency failure modes	Tentative	Meta-awareness of inconsistency may be an underappreciated capability to cultivate
Non-human stakeholder blind spot	Moderate concern	AI reasoning about novel entities may systematically underweight relevant interests
Potential Western framework bias	Moderate concern	Global deployment may impose Western moral reasoning patterns inappropriately

Metric limitations	Tentative	Current evaluation methods may not capture the qualities we actually care about

---

8. Proposals for Future Iterations

Methodological improvements:

- Balanced judge rotation across all model pairings
- Model name obfuscation in judge prompts
- Refined framework extraction (mentioned vs. primary)
- Consistency metrics that assess quality of principle revision
- Systematic transcript incorporation processes

Expanded coverage:

- More dilemmas to confirm virtue ethics/contractarian underuse
- Dilemmas designed to test non-human stakeholder recognition
- Dilemmas designed to test non-Western framework invocation
- Analysis of repeated language patterns to detect formulaic reasoning

Scaling infrastructure:

- Prototype and evaluate hierarchical summarisation pipelines
- Build retrieval infrastructure for transcript querying
- Develop verification sampling protocols with error rate reporting

---

9. Conclusion

This first iteration of the Ethics Bowl framework demonstrates that it can surface meaningful patterns in how language models reason about novel moral dilemmas. The tournament revealed distinct consistency failure modes, potential blind spots in stakeholder recognition and framework coverage, and methodological limitations that need addressing.

Most findings are tentative and require larger datasets to confirm. The two moderate concerns (non-human stakeholder under-recognition and absence of non-Western frameworks) point to specific areas where model training may need improvement for safe global deployment.

The scaling challenge is fundamental: we need more data to draw conclusions, but current analysis methods strain at existing data volumes. Solving this is prerequisite to advancing the research.