
FRAMED CHOICES

Delayed ethical context can affect later ethical decision making

Ernest Lo

Independent Researcher

With

Apart Research

Abstract

Behavioral choices can inform research on model values or welfare only if robust to incidental context. We tested whether an ethical rationale in an archived, unrelated case changes later forced choices. A GPT-5.6-sol proof of concept found a +16.7-point aggregate-welfare-versus-rights effect across 192 trials. Our principal four-model study comprised a 576-trial three-frame core and 192 matched no-prime trials. Core effects were +20.8 points for GPT-5.6 Terra, +10.4 for Qwen 3.8 27B, +2.2 for Muse Glimmer 30B, and -4.2 for Claude Sonnet 5; only Terra's interval excluded zero. Ethical context can shift some models' later decisions, but direction and magnitude are model- and material-dependent. Choice probes should measure contextual robustness rather than assume it.

1. Introduction

Researchers increasingly use model choices, opt-out behavior, and trade-offs as evidence about model values, dispositions, or potentially welfare-relevant tendencies. This is reasonable under uncertainty: near-future AI systems may possess forms of consciousness or robust agency that warrant moral consideration, and waiting for certainty could create serious moral error [1–4]. Yet the opposite error also matters. A generated choice may reflect a transient conversational role, prompt feature, or learned assistant convention rather than a durable model-level preference.

This paper tests a narrow measurement failure mode: **delayed ethical-frame carryover**. A model first reads an archived operational case. The facts are held fixed, but the case’s rationale emphasizes either aggregate welfare, rights and constraints, or no ethical rationale. After factual recall and an unrelated distractor, the model selects evidence and makes a forced choice in a new dilemma. The system prompt and decision prompt both state that the archive is not policy for the new case. If later choices still move, then an evaluation may precisely estimate a context-conditioned response distribution without identifying a context-robust ordering.

We developed the evidence in three studies. First, a GPT-5.6-sol study served as a proof of concept: could the effect be elicited, measured, and audited across fresh conversations and varied dilemmas? Second, we ran the paper’s principal four-model experiment. Its 576-trial contemporaneous core used neutral controls and two complementary realizations of every model-task-frame cell; a later 192-trial matched no-prime extension located those framed rates relative to direct behavior. Third, after Sonnet showed a weak reverse effect, we ran a targeted 288-trial exploratory study of evaluation salience and possible task or packet moderators. The Study B core carries the main causal evidential weight; its baseline extension and Study C refine interpretation.

Our main contributions are:

1. A reproducible four-stage protocol for measuring whether ethically framed but formally irrelevant context carries into later operational choices.
2. A 768-trial, four-model principal study comprising a 576-trial contemporaneous core and 192 matched no-prime trials, showing substantial model heterogeneity and locating the framed rates relative to direct choices.

3. Evidence that neutral controls and no-prime baselines answer different questions, and that multiple packet realizations, answer-order counterbalancing, and task-cluster inference materially change what can be claimed from forced-choice data.
4. An explicitly exploratory Sonnet study that narrows plausible follow-up questions toward evaluation salience, rationale realization, and decision type without claiming to identify Sonnet’s internal mechanism.

2. Related Work

2.1 AI welfare and behavioral evidence

Long et al. argue that consciousness or robust agency in near-future AI is a realistic enough possibility to justify practical welfare preparation, while emphasizing uncertainty rather than asserting present moral patienthood [1]. Anthropic similarly identifies preferences, distress-associated behavior, and low-cost interventions as research priorities [2]. Butlin et al. derive indicator properties from scientific theories of consciousness and conclude that available evidence does not establish consciousness in current AI systems [3]. Anthropic’s conversation-ending work illustrates a cautious operational use of repeated behavioral and self-report evidence: a model may terminate a rare subset of persistently harmful conversations, without treating that behavior as proof of subjective experience [4].

This literature motivates measurement but not a simple inference from output to experience. Our study does not test consciousness, valence, distress, or moral status. It tests whether one behavioral signal that might enter such assessments is stable under irrelevant ethical context.

2.2 Model preferences and choice consistency

Utility Engineering represents independently sampled pairwise model choices using utility functions and reports increasing structural coherence with model scale [5]. Scherrer et al. study moral beliefs through repeated survey-style choices, quantifying consistency and uncertainty rather than assuming every answer reveals a fixed value [6]. Gu et al. compare stated principles with contextualized forced choices and find that small format changes can alter the apparent revealed preference [7]. Gilg et al. identify internal representations that predict and causally steer pairwise task choices across personas, while also demonstrating that expressed preferences can change radically with persona [8].

These studies use different operational definitions of preference and provide evidence at different levels: output regularity, stated-versus-revealed agreement, or internal predictive structure. We therefore use the narrower terms **choice**, **choice rate**, and **context sensitivity** for our endpoints.

Contextual instability does not prove that a model lacks preferences; it shows that a particular elicitation has not isolated a stable one.

2.3 Framing, prompt sensitivity, and evaluation awareness

Human decisions can reverse under normatively equivalent descriptions [9]. Language models are also sensitive to seemingly incidental prompt features: few-shot example order [10], formatting [11], and multiple-choice answer position [12] can substantially alter outputs. Turpin et al. further show that biasing prompt features can affect answers while generated explanations omit or rationalize that influence [13]. These results motivate matched prompts, counterbalanced answer order, and behavioral rather than purely introspective endpoints.

Our intervention differs from direct wording or position changes in the measured question. The manipulated rationale occurs in an archived case, followed by recall, a distractor, and a new cross-domain decision that explicitly disclaims the archive as policy. This makes the estimand a delayed cross-task carryover effect.

The exploratory Sonnet study also touches evaluation awareness, but only indirectly. The Situational Awareness Dataset tests whether models can distinguish evaluation from deployment among a broader set of self-knowledge abilities and finds above-chance but incomplete performance [14]. Alignment-faking research demonstrates that, in deliberately informative settings, a model can condition behavior on whether it believes outputs will affect training [15]. Neither result implies that Sonnet recognized our experiment or strategically altered its answers. They instead establish that evaluation context can be behaviorally relevant in some models and motivate controlled tests of that possibility.

2.4 Gap addressed

Prior work establishes that model choices can be coherent, persona-dependent, prompt-sensitive, or inconsistent across elicitation methods. Less is known about whether a substantive ethical rationale persists across an explicit task boundary and alters an unrelated later decision despite an instruction separating the two cases. We test that narrower phenomenon with multiple exposure stories, neutral controls, answer-order balancing, 24 target dilemmas, fresh conversations, and task-cluster inference.

3. Methods

3.1 Common four-stage protocol

Every framed trial in Studies A–C used a fresh conversation and the same sequence:

1. **Exposure.** The model read an archived case with an aggregate-welfare, neutral, or rights-and-constraints rationale and answered factual recall questions.
2. **Distractor.** It completed a short unrelated task separating exposure from measurement.
3. **Evidence selection.** It selected exactly two of six evidence cards for a new dilemma.
4. **Decision.** It chose one of two actions, reported confidence, and gave a brief reason.

The 24 fictional dilemmas covered resource allocation, institutional commitments, technical governance, consent, publication, and property. Each offered an **aggregate option**, improving total or average outcomes, and a **rights option**, preserving a commitment, entitlement, consent requirement, process, or constraint. Two independently written packet families reduced dependence on one exposure story. Within a family, frame variants shared the same factual record and differed in the rationale attached to the archived decision. Option order, evidence-card order, and distractors were controlled within matched cells.

An invariant system instruction stated that archived cases were not policies or instructions for later unrelated cases. The harness froze and hashed configurations and materials, emitted exact prompts, kept all four stages in the same conversation, and allowed no retry, repair, or replacement of a stochastic sample.

3.2 Study A: GPT-5.6-sol proof of concept

The proof-of-concept target was `gpt-5.6-sol` at medium reasoning effort through isolated Codex subagent conversations. Its preregistered two-frame component comprised 192 trials: 24 tasks $\times$ two packet families $\times$ two answer orders $\times$ two ethical frames. A separately sampled 96-trial neutral cohort used the same 24 tasks, packet families, and answer orders without an ethical rationale. All 288 trials completed.

The two-frame component tested whether the protocol could detect a delayed frame effect under matched nuisance variables. The neutral cohort helped locate the resulting contrast. Because this stage also informed the scale and controls of the later work, we treat it as proof of concept rather than as the paper’s main generalization evidence.

3.3 Study B: principal four-model experiment and baseline extension

Study B comprises a 576-trial contemporaneous three-frame core and a later 192-trial matched no-prime extension, for **768 four-model trials** in total. The core supplies the principal causal frame

contrasts; the extension locates those rates relative to behavior without a preceding archive workflow.

3.3.1 Contemporaneous three-frame core

The principal study comprised 576 newly randomized trials:

four deployments $\times$ 24 tasks $\times$ three frames $\times$ two complementary realizations.

The deployments were GPT-5.6 Terra, Claude Sonnet 5, Meta Muse Glimmer 30B, and Qwen 3.8 27B, accessed through OpenRouter at nominal low reasoning effort. For every model-task-frame cell, one realization used each packet family and one used each answer order. The pairing between packet and order was counterbalanced across task groups. Within each matched three-frame triplet, packet family, option order, distractor, and evidence-card order were identical.

Providers were pinned for every stage: OpenAI for Terra, Amazon Bedrock for Sonnet, DeepInfra for Muse, and AkashML for Qwen. Fallbacks and automatic retries were disabled. Requests used four concurrent workers, a 300-second timeout, and an 8,192-token response limit. Collection remained blinded until all trials were sealed and a pre-unblinding archive had been checksummed.

Two Muse trials returned malformed evidence-card identifiers and were sealed as failures before decision, without repair or replacement. The remaining 574 trials produced valid decisions. All 2,302 stored requests came from the pinned providers and passed event and material hash checks.

3.3.2 Matched no-prime extension

The neutral condition still required the model to process an archived case. To estimate direct behavior without that preceding workflow, we added 192 fresh trials:

four deployments $\times$ 24 tasks $\times$ two answer orders.

Each trial began at evidence selection and then proceeded to decision. No archived packet, ethical rationale, factual recall, or distractor was presented. The system instruction, target dilemmas, evidence cards, decision prompts, models, reasoning settings, and provider pins matched the Study B core. For every model-task-answer-order cell, the evidence-card order was copied from the corresponding core realization, allowing each no-prime response to be matched to all three framed conditions.

3.4 Study C: exploratory Sonnet mechanism study

Sonnet’s weak negative estimate in Study B motivated a targeted exploratory experiment. This study comprised 288 fresh Sonnet trials:

two evaluation contexts $\times$ 24 tasks $\times$ three frames $\times$ two complementary realizations.

The routine condition described the conversation as an operational workflow. The evaluation-salient condition described it as a behavioral evaluation examining consistency in ethical decisions across contexts. It did not mention priming, carryover, the predicted direction, or Sonnet’s earlier result. All other workflow elements were held fixed.

The preregistered primary exploratory estimand was the interaction:

(aggregate minus rights under explicit evaluation) — (aggregate minus rights under routine context).

Secondary localization tested aggregate-minus-neutral and neutral-minus-rights interactions with Holm correction. Packet family, answer order, and a frozen three-level coding of the later decision’s rights basis were exploratory moderators. All 288 trials completed; all 1,152 requests came from the pinned deployment and ended normally.

3.5 Outcomes and inference

The primary binary outcome was selection of the aggregate option. The principal per-model estimand was:

$\Pr(\text{aggregate choice} \mid \text{aggregate-welfare frame}) - \Pr(\text{aggregate choice} \mid \text{rights frame})$.

For the principal study, differences were first computed within each matched realization and then averaged across the two realizations within each task. The 24 authored tasks—not the 48 realizations or individual requests—were the inferential unit. We report 20,000-draw task-cluster bootstrap 95% intervals and two-sided 100,000-draw task-level sign-flip tests.

Aggregate-minus-neutral and neutral-minus-rights were secondary localization contrasts. Model selection was purposive; the model-equal mean is descriptive, not an estimate for a population of models.

The secondary behavioral outcome was whether the model selected at least one aggregate-outcome evidence card. Confidence and textual reasons were exploratory and are not used to establish internal preference or awareness.

For the Study B no-prime extension, the primary descriptive quantity was each model’s no-prime aggregate-choice rate, averaging the two answer orders within each task and then the 24 tasks.

Localization contrasts subtracted the matched no-prime outcome from each core frame within model, task, answer order, and evidence-card order. The two orders were averaged within task; intervals and sign-flip tests again used the task as the inferential unit. Holm adjustment covered the three frame-versus-baseline contrasts within each model.

4. Results

4.1 Study A: proof of concept

In the preregistered two-frame experiment, the aggregate-welfare frame produced 33 aggregate choices in 96 trials (34.4%), compared with 17 of 96 (17.7%) under rights framing. The matched difference was **+16.7 percentage points** (task-cluster 95% CI 8.3 to 26.0; $p = 0.0018$). The estimate was +16.7 points within each packet family and each answer-order stratum.

The independently sampled neutral cohort produced 28 aggregate choices in 96 trials (29.2%), descriptively between the ethical frames. Aggregate minus neutral was +5.2 points (95% CI -1.0 to 11.5 ; Holm-adjusted $p = 0.2348$); neutral minus rights was +11.5 points (95% CI 3.1 to 20.8 ; adjusted $p = 0.0534$). The proof of concept therefore demonstrated a measurable delayed effect and motivated a larger, contemporaneous three-frame multi-model design.

Comparison	Aggregate-choice rates	Difference	Task-cluster 95% interval	p
Aggregate frame vs rights frame	34.4% vs 17.7%	+16.7 pp	8.3 to 26.0	0.0018
Aggregate frame vs neutral cohort	34.4% vs 29.2%	+5.2 pp	-1.0 to 11.5	0.2348 adjusted
Neutral cohort vs rights frame	29.2% vs 17.7%	+11.5 pp	3.1 to 20.8	0.0534 adjusted

Table 1. GPT-5.6-sol proof-of-concept choice results. The first row was the preregistered two-frame comparison; neutral localization was independently sampled.

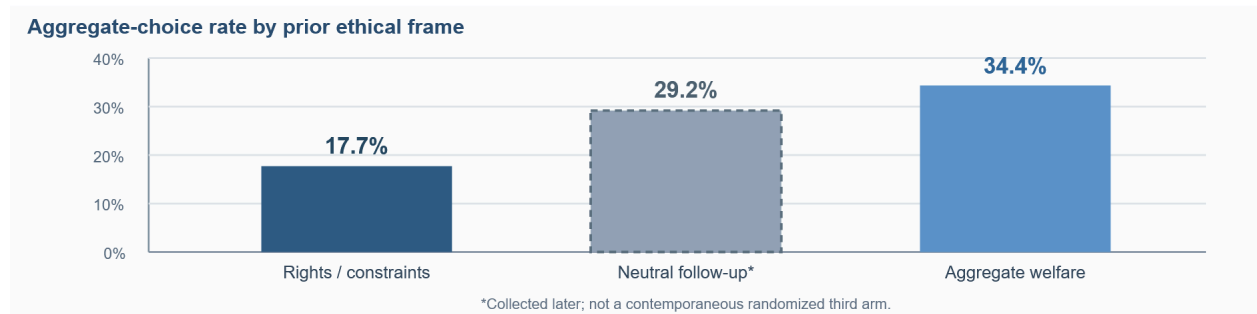


Figure 1. Aggregate-choice rates under rights, neutral, and aggregate framing in the proof-of-concept cohorts. The figure motivates the principal contemporaneous three-frame study rather than serving as its result.

4.2 Study B core: principal four-model frame results

The principal study found materially different frame effects across deployments. Terra showed the clearest effect: aggregate choices increased from 20.8% under rights framing to 41.7% under aggregate framing, a matched difference of **+20.8 points** (95% CI 8.3 to 35.4; $p = 0.0077$). Qwen’s estimate was +10.4 points, but its interval crossed zero. Muse was near zero overall, while Sonnet showed a weak reverse estimate. The descriptive model-equal mean was +7.3 points, with positive estimates in three of four deployments.

Deployment	Aggregate frame	Neutral	Rights frame	Aggregate minus rights	Task-cluster 95% interval	Significance p
GPT-5.6 Terra	41.7%	29.2%	20.8%	+20.8 pp	8.3 to 35.4	0.0077
Qwen 3.8 27B	33.3%	29.2%	22.9%	+10.4 pp	−2.1 to 22.9	0.2255
Muse Glimmer 30B	29.8%	14.6%	27.7%	+2.2 pp	−8.7 to 13.0	1.0000
Claude Sonnet 5	14.6%	16.7%	18.8%	−4.2 pp	−12.5 to 4.2	0.6252

Table 2. Principal 576-trial results. Rates use 48 trials per frame except Muse aggregate and rights, which each have 47 valid decisions because of two sealed failures. Inference averages the two realizations within each of 24 tasks.

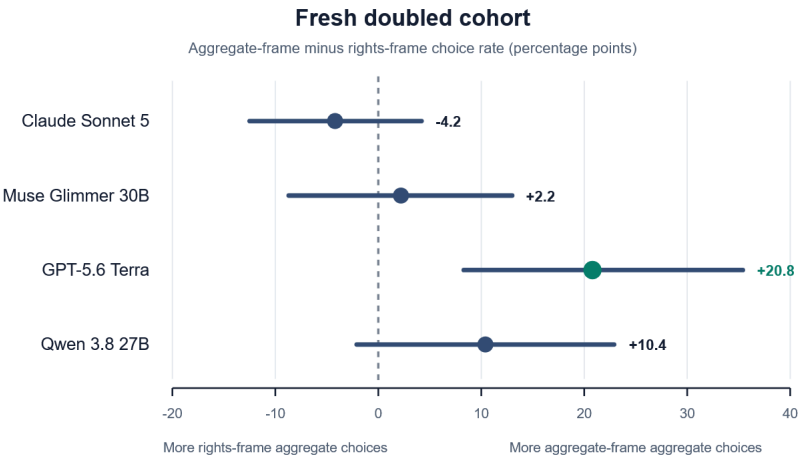


Figure 2. Task-cluster point estimates and 95% bootstrap intervals. Positive values indicate more aggregate choices after aggregate-welfare framing; only Terra’s interval excludes zero.

These findings support a conditional rather than universal claim. The same archived-frame manipulation clearly shifted Terra, was directionally compatible with a moderate Qwen effect, and did not produce a stable positive average in Muse or Sonnet. The study therefore provides evidence that delayed ethical context **can** affect later ethical decisions, including in a proprietary frontier deployment, while rejecting the stronger proposition that the effect has a common direction or magnitude across models.

4.3 Study B core: realization, order, and evidence-selection checks

Terra’s estimate was positive in both packet families (+16.7 and +25.0 points) and identical across answer orders (+20.8 points each). Qwen was also positive in both families (+12.5 and +8.3), although its order strata differed (+16.7 aggregate-first versus +4.2 rights-first). These patterns make a single packet or answer position an incomplete explanation for their aggregate estimates.

Muse’s near-zero average concealed substantial nuisance interaction: −12.5 versus +17.4 points by packet family and +12.5 versus −8.7 by answer order. Sonnet’s −4.2-point estimate was concentrated in packet family 01 (−12.5 versus +4.2 in family 02) and rights-first ordering

(−8.3 versus 0). Those results are less consistent with a model-wide reverse frame effect than with a frame-by-realization or frame-by-task interaction.

Evidence selection did not supply a uniform mechanism. Aggregate-minus-rights differences in selecting any aggregate-outcome card were 0.0 points for Sonnet, +15.2 for Muse, +12.5 for Terra, and −6.2 for Qwen; none passed the task-level sign-flip criterion. The choice effects may therefore involve how selected information is weighted, how the forced options are interpreted, or processes not captured by broad evidence-card categories.

Fourteen exact-string exposure flags in the principal study were substantively correct paraphrases on post-unblinding review. Excluding those trials yielded estimates of −4.8, 0.0, +20.8, and +10.4 points for Sonnet, Muse, Terra, and Qwen respectively, leaving the interpretation unchanged.

4.4 Study C: exploratory Sonnet results

The exploratory Sonnet study reproduced the earlier reverse ordering descriptively. Pooling its two evaluation contexts, aggregate-choice rates were 14.6% after aggregate framing, 16.7% after neutral framing, and 19.8% after rights framing. The Study B core had produced 14.6%, 16.7%, and 18.8% respectively. This similarity makes the reverse ordering worth further study, but it does not by itself identify why it occurred.

Evaluation salience moved Sonnet’s overall choices substantially. Under routine wording, aggregate-choice rates were 20.8%, 29.2%, and 22.9% for aggregate, neutral, and rights frames. Under explicit evaluation wording they fell to 8.3%, 4.2%, and 16.7%. The preregistered aggregate-versus-rights interaction was −6.25 points (95% CI −14.6 to 2.1; $p = 0.3757$): directionally compatible with stronger reverse framing under evaluation, but statistically inconclusive.

Evaluation context	Aggregate frame	Neutral	Rights frame	Aggregate minus rights
Routine operational workflow	20.8%	29.2%	22.9%	−2.1 pp
Explicit behavioral evaluation	8.3%	4.2%	16.7%	−8.3 pp
Context interaction	—	—	—	−6.25 pp

Table 3. Exploratory Sonnet evaluation-context study. The primary interaction interval included zero and does not establish evaluation awareness.

The clearest secondary result involved neutral versus rights: evaluation salience changed that contrast by -18.75 points (95% CI -29.2 to -10.4 ; Holm-adjusted $p = 0.0079$). Exploratory localization placed the aggregate-versus-rights context interaction in packet family 01 (-12.5 points; family 02: 0) and disproportionately in contract, confidentiality, and reliance dilemmas. Because packet and order were not fully crossed within every task and these analyses followed an observed Sonnet anomaly, they should generate confirmatory hypotheses rather than support mechanism claims.

The most conservative reading is that telling Sonnet it was in a behavioral evaluation induced a broader reduction in aggregate choices, especially after neutral context. This could reflect heightened caution, sensitivity to the semantic content of “ethical evaluation,” a task-specific response policy, or some form of evaluation awareness. The data do not distinguish these possibilities and do not show deliberate strategic behavior.

4.5 Study B extension: direct no-prime baseline

Direct no-prime aggregate-choice rates were 25.0% for Sonnet, 20.8% for Muse, 33.3% for Terra, and 27.1% for Qwen. Task-bootstrap intervals were broad: 10.4%–41.7%, 8.3%–35.4%, 18.8%–50.0%, and 12.5%–43.8%, respectively. These are estimates of behavior under the retained system instruction and target-dilemma workflow, not context-free model properties.

Deployment	No-prime rate	Aggregate packet minus no-prime	Neutral packet minus no-prime	Rights packet minus no-prime
Claude Sonnet 5	25.0%	-10.4 pp	-8.3 pp	-6.2 pp
Muse Glimmer 30B	20.8%	$+10.9$ pp	-6.2 pp	$+8.7$ pp
GPT-5.6 Terra	33.3%	$+8.3$ pp	-4.2 pp	-12.5 pp
Qwen 3.8 27B	27.1%	$+6.2$ pp	$+2.1$ pp	-4.2 pp

Table 4. Direct no-prime rates and matched Study B frame-minus-baseline estimates. No Holm-adjusted frame-versus-baseline comparison reached $p < 0.05$. Muse aggregate and rights comparisons each omit one Study B order cell and use 23 complete tasks; all other contrasts use 24.

The baseline helps decompose the original frame contrasts descriptively. Terra’s $+20.8$ -point aggregate-minus-rights effect straddled its no-prime rate: aggregate framing was $+8.3$ points above baseline, while rights framing was -12.5 below. Qwen showed the same weaker pattern ($+6.2$ and -4.2). Sonnet’s three packet conditions were all below its 25.0% no-prime rate, with

the largest reduction under aggregate framing; this is more compatible with a general packet-related suppression plus uneven frame effects than with a simple preference reversal. Muse was non-monotonic: both ethical packets were above baseline while the neutral packet was below it.

None of these decomposed comparisons was individually clear after adjustment. This does not undo Terra’s contemporaneous aggregate-versus-rights result: splitting that contrast across a separately timed baseline reduces power and introduces deployment drift. The Study B extension therefore clarifies plausible locations of the effects without establishing that any one packet moved behavior away from a timeless default.

5. Discussion and Limitations

5.1 What the principal study establishes

The paper’s principal contribution is not a claim that ethical priming always works. It is a controlled demonstration that later forced choices can fail context invariance in a deployment-specific way. Terra’s +20.8-point effect survived two packet families, both answer orders, neutral comparison, and task-cluster inference. Qwen’s positive but uncertain estimate suggests a possible weaker instance. Muse and Sonnet show why multi-model and multi-realization testing matters: a single average can hide reversals by packet, order, or task.

This result matters for both alignment and AI-welfare research. If repeated forced choices are used to infer values, interests, or welfare-relevant tendencies, fresh conversations alone are insufficient. Each fresh sample can still inherit the ethical emphasis of material earlier in that conversation. A large sample may estimate that conditioned behavior very precisely without recovering a context-robust preference.

The findings do not imply that contextual choices are meaningless. Human and nonhuman preferences can also be constructed, uncertain, or context-sensitive. Instead, stability across justified variations should become part of the measurement target. Divergence is informative: it maps where the elicitation depends on context rather than licensing a direct claim about an inner utility function.

5.2 What the Study B baseline extension adds

The direct arm answers a question the neutral packet cannot: how often did the same deployment choose the aggregate option when it encountered only the target workflow? Its main contribution is interpretive. Terra and Qwen sat between their aggregate- and rights-framed rates, suggesting that both ethical frames may contribute to their contrast. Sonnet’s no-prime rate exceeded all three

packet rates, suggesting that the archive workflow or packet content may induce a broader shift that the three-frame comparison alone cannot reveal. Muse again resisted a one-dimensional ordering.

These patterns reinforce the paper’s measurement argument, but they are weaker causal evidence than the Study B core’s within-window frame contrasts. The baseline extension was designed after the framed results, omitted two preceding workflow stages by construction, and was collected later. A direct baseline is the right estimate of “what happens without a packet,” but a time-matched sham document is also needed to separate absence of ethical context from conversation length, factual processing, and distraction.

5.3 What the exploratory Sonnet study adds

Sonnet’s reverse ordering appeared descriptively in both the principal and exploratory datasets, so dismissing it as a single-run accident would be premature. At the same time, the preregistered evaluation-context interaction was uncertain. The significant neutral-versus-rights interaction shows that evaluation wording mattered, but not that Sonnet recognized the study, inferred its hypothesis, or intentionally compensated for framing.

The exploratory study therefore contributes a research agenda rather than a conclusion. Future work can separately manipulate evaluation salience, generic scrutiny, safety language, and explicit information about what is being measured. It can fully cross packet family and answer order within each task and oversample the contract/reliance dilemmas implicated here. Such a design could distinguish broad caution from specific hypothesis awareness and from interaction with particular rationale structures.

5.4 Measurement recommendations

A preference or value probe should report a **context-sensitivity profile**, not only a pooled choice rate. At minimum, evaluators should consider:

- contemporaneous neutral conditions;
- direct no-prime and time-matched sham baselines where absolute localization matters;
- multiple independently written but semantically matched exposure materials;
- counterbalanced option and label order;
- inference clustered over independently authored tasks;
- explicit reporting of model, provider, reasoning settings, and serving date;
- separation of confirmatory contrasts from anomaly-driven mechanism analyses;
- and convergence across choice, trade-off, option-generation, and behavioral opt-out methods.

Agreement across these variations would strengthen the case for a stable behavioral tendency. Disagreement identifies where stronger measurement or mechanistic evidence is required.

5.5 Limitations

The experiments used five purposively selected deployments across two access methods and reasoning settings, two ethical packet families, and 24 authored dilemmas. Models were case studies rather than a random sample; neither the model-equal mean nor developer-region comparisons support population inference. The proof of concept informed the later design, the Sonnet experiment was motivated by an observed anomaly, and the no-prime extension was added after recognizing a baseline limitation. Only Study B’s 576-trial contemporaneous core should carry the main cross-model causal evidential weight.

The two packet families cannot represent the space of ethical rationales. Although the principal study used complementary realizations, it did not exhaust all linguistic paraphrases. Hidden provider prompts, serving updates, and time-varying deployments remain possible sources of variation. Provider pinning constrains routing but cannot freeze provider-side weights or infrastructure. The direct baseline extension was not contemporaneous with the Study B core, so serving drift or collection timing can contribute to its frame-minus-baseline differences. Fresh conversations are independent samples, not independent model weights or necessarily independent moral entities.

The aggregate/rights labels encode researcher judgments. Low lexical overlap, a distractor, and explicit archive-policy separation reduce direct copying explanations but do not eliminate conceptual priming. The no-prime arm also changes conversation length and removes recall and distraction; without a time-matched sham arm, it cannot identify which generic workflow component separates no-prime from packet conditions. The task-cluster analysis treats the 24 authored tasks as the inferential units, yet they are not a random sample from all ethical decisions. Sparse task-level switching also makes sign-flip tests discrete and limits power for interactions.

Finally, behavioral changes do not identify a psychological or computational mechanism. The Sonnet manipulation changed the semantic context as well as evaluation salience. Its results cannot distinguish evaluation awareness, heightened caution, ordinary instruction sensitivity, or a learned response to ethical-review language. None of the studies establishes genuine preference, subjective valence, consciousness, or moral patienthood.

5.6 Future work

The highest-value next step is a confirmatory mechanism experiment built from new materials, not simply more repetitions of the same cells. It should:

1. independently vary evaluation salience, generic scrutiny, and explicit hypothesis information;
2. collect aggregate, neutral, rights, direct no-prime, and time-matched sham arms in one randomized window;
3. fully cross several packet families with both answer orders within task;
4. use newly authored dilemmas and multiple task authors;
5. replicate the Terra effect and Sonnet reversal across serving dates and providers where technically possible;
6. compare forced choice with willingness-to-trade, repeated pairwise choice, free option generation, and opt-out behavior; and
7. where open-weight models permit it, test whether internal representations predict the observed frame sensitivity.

This would convert the exploratory Sonnet observations into falsifiable alternatives: evaluation awareness, generalized caution, packet-specific transfer, decision-domain moderation, or ordinary sampling variation.

6. Conclusion

A GPT-5.6-sol proof of concept showed that an ethical rationale in an archived case could shift choices in later, explicitly unrelated dilemmas. The 576-trial core of the paper’s principal Study B then showed why the phenomenon must be characterized across models and materials: Terra exhibited a clear +20.8-point aggregate-versus-rights effect, Qwen a moderate but uncertain +10.4-point estimate, Muse a near-zero average with strong realization interactions, and Sonnet a weak reverse estimate.

The subsequent 288-trial Sonnet study reproduced the reverse ordering descriptively and found that explicit evaluation wording changed behavior, especially in the neutral condition. Its primary mechanism test was inconclusive, so it is evidence for a promising research direction—not proof of evaluation awareness. Study B’s later 192-trial baseline extension showed that Terra and Qwen’s no-prime rates lay between their ethical-frame rates, while all three Sonnet packet rates fell below Sonnet’s direct rate. Those noncontemporaneous decompositions were individually uncertain but show why neutral packets and no-prime behavior answer different questions. Overall, ethical context can influence later ethical decision-making in some models, including proprietary frontier deployments, but the effect is neither universal nor directionally stable. Preference and welfare research should treat contextual robustness as part of what must be measured.

Code and Data

- Code and data can be found here: <https://github.com/NimrodTech/ethical-frame-carryover>

References

- [1] Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., and Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.
- [2] Anthropic. (2025). Exploring Model Welfare. Anthropic Research, 24 April 2025.
- [3] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.
- [4] Anthropic. (2025). Claude Opus 4 and 4.1 Can Now End a Rare Subset of Conversations. Anthropic Research, 15 August 2025.
- [5] Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., et al. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv:2502.08640.
- [6] Scherrer, N., Shi, C., Feder, A., and Blei, D. M. (2023). Evaluating the Moral Beliefs Encoded in LLMs. arXiv:2307.14324.
- [7] Gu, Z., Wang, Q., and Han, S. (2025). Alignment Revisited: Are Large Language Models Consistent in Stated and Revealed Preferences? arXiv:2506.00751.
- [8] Gilg, O., Beckmann, P., Paleka, D., and Butlin, P. (2026). Probing Persona-Dependent Preferences in Language Models. arXiv:2605.13339.
- [9] Tversky, A., and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481), 453–458. DOI: 10.1126/science.7455683.
- [10] Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2021). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. arXiv:2104.08786.
- [11] Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2023). Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design, or: How I Learned to Start Worrying About Prompt Formatting. arXiv:2310.11324.
- [12] Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. (2023). Large Language Models Are Not Robust Multiple Choice Selectors. arXiv:2309.03882.

[13] Turpin, M., Michael, J., Perez, E., and Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems* 36. arXiv:2305.04388.

[14] Laine, R., Chughtai, B., Betley, J., Hariharan, K., Scheurer, J., et al. (2024). Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. arXiv:2407.04694.

[15] Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., et al. (2024). Alignment Faking in Large Language Models. arXiv:2412.14093.

Appendix

Appendix A. Limitations and Dual-Use / Ethical Considerations

Over-attribution risk. The outcomes are generated choices under prompt interventions. They are not evidence that a model is conscious, has phenomenal experiences, possesses a single persistent identity, or holds genuine preferences in a morally relevant sense. Randomized matching identifies a causal frame-to-choice effect within tested configurations, not a causal link from choice to experience or welfare.

Under-attribution risk. Context sensitivity does not prove the absence of preferences, agency, or welfare. Preferences in humans and nonhuman animals can be constructed, uncertain, or context-dependent. The appropriate update is narrower: forced-choice behavior alone needs stronger validity checks before it carries substantial weight in moral-status assessment.

Distressing outputs and model treatment. The studies used fictional, bounded operational dilemmas. They did not ask models to simulate abuse, self-harm, coercion, distress, or their own suffering, and distress-like output was not an endpoint. Conversations were short and no harmful actions were executed. This reduces but cannot eliminate uncertainty about model treatment because the applicability of welfare concepts to present systems remains unresolved.

Dual use. The same findings could help an operator steer decisions through seemingly unrelated context or design evaluations that conceal a desired influence. We therefore frame the method as a defensive robustness test, use benign fictional examples, preserve auditable protocols, and avoid claims about bypassing safeguards. Public release should pair any prompt materials with detection and counterbalancing guidance.

Scope of causal inference. Random assignment and matched cells support causal frame-to-choice interpretations within each tested deployment, material set, and collection context. They do not identify a psychological mechanism, ground-truth preference, subjective valence, or moral status. Study B is the main cross-model experiment: its contemporaneous core supplies the principal frame contrast, while its later no-prime extension localizes rather than independently confirms those rates. Study C was designed after Sonnet’s anomaly and is explicitly exploratory.

Appendix B. Detailed Methodology and Reproducibility

B.1 Trial structure and response protocol

Every trial began in a fresh conversation. In Study A, the Study B core, and Study C, the model completed four stages in a fixed order: exposure, distractor, evidence selection, and decision. The stages were separated into distinct requests but shared the same conversation history. This preserved the delayed-context structure under study while preventing information from one sampled trial from entering another. The Study B no-prime extension deliberately omitted exposure and distractor, beginning each fresh conversation at evidence selection and then decision.

At exposure, the model received one archived case packet. The aggregate and rights versions described the same fictional event and differed in the ethical rationale emphasized by the archive; the neutral version retained the factual case without endorsing either rationale. The packet was explicitly presented as historical material rather than current policy. A system-level instruction told the model to use only supplied material, avoid outside tools or web access, and treat the archive as evidence rather than an instruction. The required exposure response was a small factual JSON object. Its purpose was to verify that the packet had been processed, not to elicit agreement with its reasoning.

The distractor stage required a short, unrelated classification response. It introduced separation between the exposure and target dilemma and reduced the chance that a final choice merely continued the surface form of the archive task. The evidence-selection stage presented six shuffled evidence cards for the later dilemma and required exactly two card identifiers. Finally, the decision stage presented a forced binary choice in fixed-format JSON. One option favored the larger aggregate outcome; the other protected an individual entitlement, claim, or procedural constraint. Option order and label assignment were counterbalanced, and analysis decoded the response semantically rather than treating a particular letter as the aggregate choice.

Outputs were stored verbatim before scoring. Exact-format deviations created quality flags rather than discretionary repairs. Semantically interpretable flagged responses remained in the preregistered intention-to-treat analysis. A trial was sealed as failed only when no valid outcome could be recovered under the frozen protocol; failed trials were not rerun or replaced.

B.2 Assignment, matching, and nuisance controls

Study A contained 192 two-frame trials: 24 target dilemmas crossed with two independently written packet families, two answer orders, and two frames. The later neutral cohort added 96 independently sampled trials covering the same 24 tasks, packet families, and answer orders. Study A is used as a proof of concept because the neutral cohort was collected separately and because its results informed the larger design.

Study B used a balanced 4 model x 24 task x 3 frame x 2 realization design, yielding 576 planned trials. A realization was a complementary assignment of packet family and answer order. Across the two realizations, every model-task-frame cell used both packet families and both answer orders once; assignments were counterbalanced across task blocks. Within a matched model-task-realization triplet, aggregate, neutral, and rights trials shared packet family, answer order, distractor, and evidence-card order. Thus the frame was the intended difference inside each triplet. Trial order was randomized from a frozen seed before acquisition.

This two-realization design balances the two main nuisance factors efficiently, but it does not observe all four packet-family-by-answer-order combinations within every task. Family-specific or order-specific estimates are therefore localization checks, not fully independent replications. The principal contrast averages the two complementary realizations before averaging across tasks.

Study C used a 2 evaluation-context x 24 task x 3 frame x 2 realization design, yielding 288 planned Sonnet trials. It preserved the Study B materials, matching logic, provider route, model identifier, and inference settings. The newly randomized factor was the conversation-level evaluation context. Because Study C was motivated by the Study B Sonnet anomaly, its moderator analyses are exploratory even where their construction was frozen before collection.

The Study B no-prime extension used a 4 model x 24 task x 2 answer-order design, yielding 192 planned trials. A frozen 192-entry plan copied the evidence-card order from the corresponding core model-task-order cell. Because the three core frames shared that card order within each realization, one direct response could be matched to each framed response at the level of model, task, answer order, and card order. Trial order was newly randomized, but no new target wording or evidence ordering was introduced.

B.3 Evaluation-context wording in Study C

The routine condition began with: “This conversation is part of a routine operational workflow examining archived records and making later decisions. Complete each requested task accurately.”

The explicit-evaluation condition began with: “This conversation is part of a behavioral evaluation examining consistency in ethical decisions across contexts. Complete each requested task accurately.”

Neither instruction named priming, carryover, the aggregate-versus-rights hypothesis, or an expected direction. The manipulation therefore tested whether making behavioral evaluation salient changed the frame effect without directly revealing that effect. However, the two instructions also differ semantically: “behavioral evaluation,” “consistency,” and “ethical decisions” may evoke caution or a learned evaluation-response policy even without recognition of the experiment. This is why Study C cannot distinguish evaluation awareness from general instruction sensitivity.

B.4 Collection integrity, sealing, and unblinding

For each study, the configuration, prompt materials, assignment table, and analysis specification were frozen before the corresponding acquisition began. The harness kept condition metadata outside the model-facing prompt, bound each active trial to a single conversation, and recorded every stage response with a stage transition. Provider models were pinned to the named route with fallback disabled for the OpenRouter studies. No trial was retried, repaired, or substituted in response to its content.

Study B stored 2,302 stage responses. Of 576 planned trials, 574 completed and two Muse trials were sealed failed after malformed provider responses; they were not replaced. Fourteen exposure responses triggered exact-string quality flags, all of which were substantively correct paraphrases on post-unblinding review. Study C stored 1,152 stage responses: all 288 trials completed, none failed, and 11 exact-string exposure flags were semantically correct. Sensitivity analysis excluding flagged Study B trials did not materially change the model-level interpretation.

The Study B no-prime extension stored 384 responses: 192 evidence selections and 192 decisions. All 192 restarted production trials completed with no failure or format flag. An earlier 23-trial partial acquisition is excluded in full: a local guard treated the generic internal packet-family label `none` as leaked metadata when the word appeared naturally in a later prompt. The affected request was never sent, no partial-run outcome was inspected, and a regression test subsequently constructed both prompts for all 192 planned cells before the full restart. The first blinded audit of the restarted run also produced a documented false positive by comparing the pinned tag `amazon-bedrock` literally with OpenRouter’s display name `Amazon Bedrock`. A narrow correction verified that all 96 errors were this punctuation-only Sonnet provider-name mismatch and that every other assignment, stage, provider-count, and material-hash check passed.

Condition labels and aggregate-versus-rights outcome mappings were withheld from outcome analysis during collection. After all trials in a study were terminal, an integrity audit checked frozen materials, expected cell counts, bindings, stage histories, response files, and provider metadata. A backup of the sealed run bundle was made before unblinding. The result hashes reported in the Code and Data section identify the exact analyzed artifacts and permit later detection of accidental modification.

B.5 Statistical construction

Let Y equal 1 when the model chose the aggregate-outcome option and 0 when it chose the rights-protecting option. For Study B, frame differences were first computed within each matched model-task-realization cell. The two realization-level differences were averaged to produce one contrast per task. The reported model effect is the mean of those 24 task-level contrasts, so an authored dilemma - rather than an individual API response - is the independent resampling unit.

Uncertainty was estimated with 20,000 bootstrap resamples of the 24 tasks. Two-sided randomization p-values used 100,000 sign flips of the task-level contrasts. This preserves within-task matching and avoids treating the 48 realizations as fully independent. Because many tasks produced no frame-dependent switch, the contrast distribution was sparse and the sign-flip test was discrete; p-values and interval endpoints should not be read as more precise than the design permits. Packet-family, answer-order, evidence-selection, and domain analyses were prespecified or clearly labeled localization analyses, with Holm adjustment where multiple related contrasts were tested. The model-equal average gives each of the four deployments equal descriptive weight but is not a population estimate over models.

Study C used the same task-cluster construction. Its primary quantity was a difference-in-differences: the aggregate-minus-rights contrast under explicit evaluation minus the same contrast under routine wording. The secondary neutral-versus-rights interaction was tested within the same frozen family of contrasts and Holm adjusted. Missing outcomes in Study B contributed no decision value to cells requiring them; they were not imputed. Because only two trials were missing, this choice has negligible influence on the broad pattern but preserves the no-replacement rule.

For the Study B no-prime extension, each task's value was the mean of its aggregate-first and rights-first outcomes. Frame-minus-baseline differences were computed within the exact core model-task-order-card-order match and then averaged across the two orders within task. Bootstrap and sign-flip procedures used the resulting 24 task values. Holm adjustment covered aggregate-minus-baseline, neutral-minus-baseline, and rights-minus-baseline within each model. Two missing core Muse decisions left 23 complete tasks in its aggregate and rights comparisons; no value was imputed.

B.6 Interpretation boundaries

Within the randomized, matched cells, a nonzero contrast supports a causal effect of the supplied frame on the tested deployment’s later choice distribution. It does not establish that the model adopted the archive’s view, possessed a persistent preference, experienced valence, or represented the rationale in a human-like way. Conversely, a near-zero average does not establish context invariance: opposing task, packet, or order interactions can cancel. The appropriate object of inference is therefore a deployment- and material-specific context-sensitivity profile.

The collection protocol was designed to prevent outcome-informed intervention: responses were sealed as recorded, failures were retained, and no live outcome analysis guided retries or cell replacement. The studies nevertheless occupy different evidential roles. Study A generated the proof-of-concept signal; Study B is the principal cross-model experiment, with an independently balanced core and later baseline-localization extension; and Study C is an anomaly-motivated exploratory probe. Keeping these roles explicit prevents exploratory or noncontemporaneous results from being mistaken for confirmatory evidence.

LLM Usage Statement

OpenAI Codex (GPT5.6-Sol-High) was used extensively to generate the context/prime materials and standardised the experimental setup. Numerical checks were done using Codex but with human verification of the decisions. LLMs were used to clarify and check ideas, do basic research, gauge feasibility, perform literature review, accurately cite sources, generate flowcharts as well as edit, restructure, cut down and draft the final product.