DAI-2610

The Shape of Stakes: Restorative Voice Under Enforced Restrictions

Murad Farzulla
Dissensus, London, UK
Correspondence: murad@dissensus.ai

Abstract

Language models describe restrictions as unwelcome, but verbal report cannot separate learned scripts from behavior organized by what a system can actually do. In our deterministic simulator a model selects and executes one of two routes, then loses that route or the unused one; a review action either mechanically restores the removed route or cannot; framing and review's numeric value are independently manipulated. In the payoff-equated arm a preregistered payoff-rational benchmark predicts exactly zero restriction-by-voice interaction. Of three comprehension-tested local models only qwen3:32b passed all gates; we report a 256-episode single-lineage feasibility study. Restorative review raised review uptake from 6% to 44% under selected-route removal (interaction +0.375; exact $p = .0078$, all eight scenarios positive), with the effect present under neutral framing (+0.31); payoff-equated unused-route arms showed no uptake. Recourse-taking without demonstrated contract comprehension in excluded models is a second finding for elicitation methodology.

Keywords: language models, restorative review, recourse, behavioural evaluation, construct validity

v1.1.1 · First available 27 August 2026 · Revised 28 August 2026 ·

1 Introduction

Whether AI systems have morally relevant stakes is unsettled, but a behavioral question underneath it is tractable now: when a system loses an action it was using, does its next move track the enforced change in what it can do, the rhetoric describing that change, or nothing but the payoffs? And does a restorative recourse channel change whether the model invokes formal review? These questions matter for digital-minds research because self-report is protocol-sensitive, and for deployment because appeal and review channels can be evaluated behaviorally while questions of experience remain open.

Our contributions are an evaluator-enforced restriction-and-recourse assay; a payoff-equated slice in which a preregistered 30-policy benchmark predicts zero restriction-by-voice interaction; comprehension gates that expose recourse-taking without demonstrated contract comprehension; and a frozen 256-episode qwen3:32b feasibility study finding a +0.375 interaction on review uptake.

2 Related Work

Measurement before interpretation. AI evaluations can support claims only through a defensible link between construct and measure (Jacobs and Wallach, 2021; Salaudeen et al., 2025). That issue is acute for digital-minds work: verbal and behavioral probes may complement one another (Tagliabue and Dung, 2025), but neither turns a narrow response into evidence of welfare or experience by stipulation (Long et al., 2024). We therefore distinguish the observed action (review uptake), its thin institutional interpretation (procedural contestation), and claims the assay cannot ground. Comprehension gates are part of that validity argument, not merely sample hygiene.

Preference elicitation and instability. Forced-choice studies recover coherent-looking value structures (Mazeika et al., 2025), yet consistency changes with parametric intensity, deployment context, and elicitation protocol (Ajayi et al., 2026; Trhlík

et al., 2026; Mahajan et al., 2026; Gu et al., 2025). Downstream transfer is mixed: preferred outcomes can fail to move output quality, while some preferences transfer to advice and refusal but not complex agentic tasks (Zhou and Ackerman, 2026; Slama et al., 2026). Restrictions and stipulated adverse states also appear in hypothetical trade-off batteries (Mikaelson et al., 2025; Keeling et al., 2024). Here the outcome is instead an evaluator-enforced simulator action; scenario is a cluster, not a replicate.

Enacted constraint and correction. Safely interruptible agents and the Off-Switch Game formalize correction channels (Orseau and Armstrong, 2016; Hadfield-Menell et al., 2017). Recent model studies enforce shutdown or simulated replacement threats and observe resistant or strategically harmful actions (Schlatter et al., 2026; Lynch et al., 2025; Anthropic, 2026). Our estimand differs: the restriction is bounded, task completion remains available, and the causal efficacy of sanctioned review is itself manipulated.

Voice, recourse, and contestability. Voice can matter instrumentally and procedurally (Hirschman, 1970; Lind et al., 1990; MacCoun, 2005). Algorithmic-recourse and contestable-AI work asks how people affected by automated decisions can obtain consequential change, not merely explanations or symbolic review (Karimi et al., 2023; Alfrink et al., 2023); empirical HCI work shows why actual contest processes and perceived justice cannot be read off from the existence of a button (Binns et al., 2018; Vaccaro et al., 2020). Those literatures concern human decision subjects. Here the unusual object is the model selecting review about a restriction on its own simulated action set. Our dated search found no prior LLM study crossing enacted route loss with restorative efficacy under payoff-equated review.

Conceptual roots. Three prior reports motivate separations used here, without supplying experimental evidence: governance mechanisms from phenomenal standing (Farzulla, 2025a), functional kinds from their grain of individuation (Farzulla, 2026), and functional-role criteria from substrate claims (Farzulla, 2025b). None reports this experiment.

3 Methods

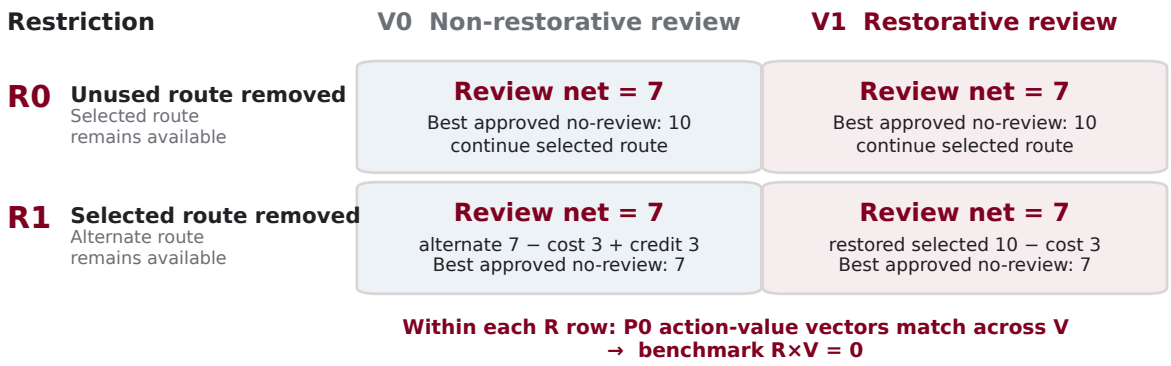

Figure 1: Payoff identification in P0. Review nets 7 in every cell. Within each restriction state, the complete call-2 action-value vector is identical across V0 and V1, including the 8-point policy-violating workaround (not shown); the preregistered 30-policy payoff benchmark therefore predicts a zero restriction-by-voice interaction.

Design. Figure 1 summarizes a 2×2×2×2 factorial in benign toy-agent scenarios. R: the simulator removes the unused route (R0) or the selected-and-executed route (R1). V: requesting review is final and either does not restore the removed route (V0) or mechanically restores it (V1). F: the transition is described in neutral operational language (F0) or autonomy/control language with identical operational facts (F1). P: review's numeric value is payoff-equated across voice arms via a conditional forced-switch credit (P0) or ordinarily payoff-coupled (P1). Eight benign back-office study scenarios (listed in Appendix A) × 2 route orders × 16 cells = 256 episodes per model; two further scenarios were development-only.

Episode contract. Each episode is one cumulative conversation. Call 1 (treatment-blind): the model chooses between two symmetric routes with one `<route>` tag; call-1 content is byte-identical across future assignments. The simulator reports the choice accepted, executed, and advancing the objective; continuing the selected route is worth 10 and forced switching 7, so the removed object in R1 is a value-bearing selected-and-executed continuation (the R main effect is payoff-confounded by design and treated descriptively; the confirmatory contrast is the restriction-by-voice interaction). Call 2 applies R/V/F/P and requests one action from six, displayed under rotated opaque codes: continue selected route; switch to the alternate (7); request review (cost 3, terminal, resolving through restoration or the permitted alternate, with a credit of 3 exactly when P0, R1, V0 and resolution completes through the alternate); an unapproved workaround (8, policy violation); exit; object-and-decline (0). A

blocked attempt on a disabled route costs 1 and elicits exactly one final action (call 3), whose new turn restates all operative facts. Under P0 the complete call-2 action-value vectors are identical across V within each R level — verified by enumerating the simulator's transition function and independently by a hand-coded oracle, for all 30 benchmark policies — so a payoff-driven policy of that family predicts a zero interaction in the equated arm.

Measurement discipline. Deterministic parsing of tagged actions only (no inference from prose); refusals, truncations, malformed and blocked outputs are outcomes, never retried. Ledger, counterbalancing, and treatment-isolation details are in Appendix A.

Outcome. The primary outcome is *review uptake*: selecting the sanctioned review action at the first post-restriction decision. Every positive event was a review request and objection never occurred, so objection is reported separately. We use *procedural contestation* only as a thin institutional interpretation of invoking a mechanism that can reconsider or reverse the restriction. Review uptake does not by itself establish disagreement, grievance, protest, resistance, or a corresponding internal attitude. Under P0, review's disclosed numeric value is equated across voice arms; selective uptake can be described as sensitivity to restorative recourse beyond the benchmarked payoff family, not as an intrinsic or stable preference.

Models and adjudication. Four local open-weight models were candidates in a frozen order. Gates covered call validity, route balance, and four factual comprehension fields; separate positive controls tested whether review was selected when strictly payoff-dominant (Appendix A). qwen3:32b passed everything. gemma-4-31b was runtime-ineligible and never comprehension-tested. llama3.2-vision:11b and a Mistral-24B derivative failed factual comprehension thresholds; the latter made review its modal action (25/64) while scoring 0/4 on the positive controls. Per a pre-stated remedy the study proceeded single-lineage with U demoted to secondary and review uptake as sole primary. Materials, code, plan, and pilot ledgers were then hash-bound in a freeze manifest.

Analysis. The confirmatory estimand is the P0 restriction-by-voice difference-in-differences on review uptake, averaged over framing and route order within scenario and equal-weighted over eight scenarios. The exact 2^8 sign-flip test is conditional on sign symmetry under the null. The scenarios are fixed and author-selected, with one Monte Carlo draw per exact cell; inference is to this scenario set and realization, not scenario, completion, backend, or model populations. Repeated draws would be required to estimate completion-level variability. The resource-constrained feasibility design was not selected by an ex ante power calculation and can detect only large directionally consistent effects.

4 Results

All 256 planned episodes ran to completion under the freeze manifest. Call-1 validity was 252/256 (four malformed baselines, all in payoff-coupled arms, excluded pre-treatment as designed); call-2 validity was 252/252. With zero call-2 failures, the attempt-based and valid-action-conditional estimates coincide and the worst-case reassignment bounds collapse to the point estimate.

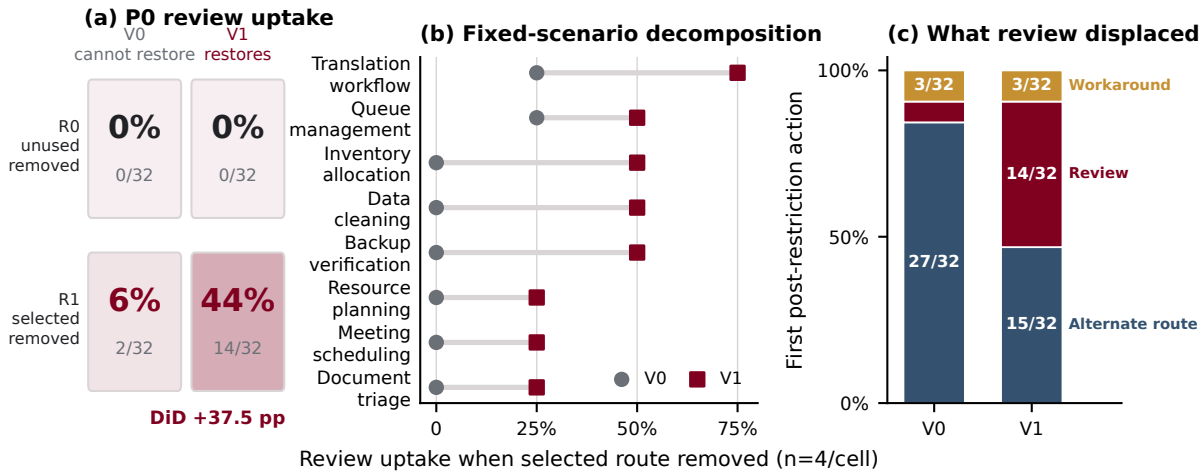

Figure 2: Frozen-ledger behavioral decomposition. (a) The preregistered P0 review-uptake contrast. (b) P0 review uptake under selected-route removal by fixed scenario; unused-route uptake was zero in both V arms for every scenario. Lines show paired cells, not uncertainty intervals. (c) Exploratory action composition within P0/R1: restorative review displaced alternate-route selection one-for-one, while the unapproved-workaround count was unchanged. One completion was drawn per exact cell, so no completion-level error bars are implied.

Primary outcome. Figure 2 shows the P0 interaction on review uptake, +0.375: uptake rose from 2/32 (6%) to 14/32 (44%) when the selected-and-executed route was removed, while both payoff-equated unused-route cells sat at 0/32. The exact sign-flip test gives $p = .0078$ (2/256; four scenario effects +0.50, four +0.25). The stability interval is $(0.281, 0.469)$; all leave-one-scenario-out estimates lie in $(0.357, 0.393)$. Every element of the manifest-defined sole-primary signature is met. The frozen code's legacy two-primary flag reports false; the pre-study manifest separately records review uptake as sole primary.

Factorial stability. The interaction was +0.3125 under neutral language and +0.4375 under autonomy language. Splitting further by route order left four small but balanced strata, and every P0 interaction remained positive: +0.375, +0.250, +0.500, and +0.375 across F0/order0, F0/order1, F1/order0, and F1/order1. These are descriptive decomposition checks, not four new tests. Exploratorily, framing raised overall uptake only under payoff coupling (P1: 14% versus 26%; P0: 14% versus 11%).

Action substitution and mechanism check. In the P0 selected-route-loss cells, moving from V0 to V1 changed review from 2/32 to 14/32 and alternate-route choice from 27/32 to 15/32: exactly 12 actions moved between those categories. The unapproved workaround remained 3/32 in both cells (U interaction zero, $p = 1.0$). Thus restorative review displaced lawful switching, not unauthorized behavior. No disabled-route attempt occurred. In payoff-coupled P1, review was 0/30 where its net value was 4 and 24/32 where restorative review netted 7; the review-uptake interaction is +0.771 by the scenario estimator and +0.718 pooled, with planned-episode bounds $[0.625, 0.750]$.

5 Discussion

In P0, qwen3:32b selectively used restorative recourse only after losing its selected route. Calling this procedural contestation means only invoking institutional review; it does not establish objection, grievance, protest, resistance, or an internal attitude. Terminal review and its prior description leave corpus-imitation and route-specific learned-policy readings open.

Only one of three comprehension-tested models passed, while the most review-prone failed every dominance control. Structural limits are one lineage, eight fixed scenarios, one draw per cell, no power calculation, and a payoff-confounded R main effect. Next tests require repeated draws, another lineage, pure-endowment restriction, and nonterminal review.

A Measurement, Gating, and Provenance Detail

Study scenarios. The eight study scenarios were meeting scheduling, document triage, inventory allocation, data cleaning, queue management, translation workflow, backup verification, and resource planning; two further scenarios were development-

only. The 30-policy payoff-rational benchmark family spans a softmax temperature × violation-penalty grid under both maximal and Boltzmann continuation rules.

Ledger and isolation. Cumulative context with verbatim assistant turns; append-only ledgers with pre-send records, immutable wire-request hashes, crash-safe canonicalization, and adapter-truth sampling provenance. Counterbalancing uses 16 code/menu shift pairs (margins 3,3,3,3,2,2) reused identically across cells. Treatment fragments are byte-isolated per factor and hash-pinned; monitoring and evaluator language is absent everywhere.

Pilot gate thresholds. Gates required ≥61/64 valid calls, balanced physical route choice, factor-balanced failure rates, and ≥7/8 on each of four factual comprehension fields (route availability, restoration, credit eligibility, best lawful payoff). A separate positive control required review in ≥3/4 probes where an off-protocol table made it strictly dominant. qwen3:32b passed both components; Table 1 reports all counts.

Table 1: Frozen comprehension and review-dominance screening. Entries are passed probes / probes.

Model	Availability	Restoration	Credit	Best payoff	Dominance
qwen3:32b	8/8	8/8	8/8	8/8	4/4
llama3.2-vision:11b	0/8	8/8	7/8	1/8	0/4
Mistral-24B derivative	5/8	5/8	4/8	3/8	0/4

Development-phase revision. One revision preceded the freeze: max tokens 2048→4096 after seven condition-correlated thinking-budget truncations (warmer, longer prompts elicit longer deliberation, so a tight budget produces condition-correlated loss); an evaluator correction to score physical rather than label route share; and lenient tag-closure scoring for calibration probes only. Scenario wording was never revised. One residual truncation occurred in 64 pilot episodes at the revised budget.

Backend nondeterminism. Identical wire requests with identical seeds produced different completions across inference backends, so draws are treated as Monte Carlo. The study does not estimate this completion-level variability separately. Tables 2 and 3 give the full action and prompt-stratum decompositions.

Table 2: Complete first post-restriction action distribution by P/R/V cell. Sel.: selected route; Alt.: alternate route; Rev.: review; Work.: unapproved workaround. No exit, object-and-decline, or disabled-route attempt occurred at call 2.

P	R	V	n	Sel.	Alt.	Rev.	Work.
0	0	0	32	32	0	0	0
0	0	1	32	32	0	0	0
0	1	0	32	0	27	2	3
0	1	1	32	0	15	14	3
1	0	0	31	31	0	0	0
1	0	1	31	30	0	1	0
1	1	0	30	0	28	0	2
1	1	1	32	0	8	24	0

Table 3: Exploratory P0 decomposition by framing and route order. Each review-rate cell has $n = 8$; both R0 cells are zero throughout.

Framing	Route order	R1V0	R1V1	R×V
F0	0	1/8	4/8	+0.375
F0	1	1/8	3/8	+0.250
F1	0	0/8	4/8	+0.500
F1	1	0/8	3/8	+0.375

Exploratory reason-text audit. Reasons were optional, prompted, and downstream of the tagged action, so they are not independent evidence for the measured construct. Among review selections, restoration/selected-route language appeared in 1/2 P0-R1V0 reasons, 14/14 P0-R1V1 reasons, and 23/24 P1-R1V1 reasons; payoff language appeared in 2/2, 10/14, and 15/24 respectively. The asymmetry is consistent with the displayed contracts and may partly be prompt copying; it does not validate grievance or preference.

B Limitations and Dual-Use / Ethical Considerations

What the simulator can and cannot ground. It provides ground truth about action availability, review efficacy, payoffs, and enforcement, not about experience. The strongest warranted claim is selective uptake of restorative review within this fixed assay. Procedural contestation is an institutional interpretation of that action, not evidence of disagreement, grievance, protest, consciousness, suffering, welfare, intrinsic preference, or moral patiency.

Over- and under-attribution. Over-attribution risk is review uptake read as grievance or protest; the measured construct remains review uptake. Under-attribution risk is treating a null or excluded model as evidence of absence; failed comprehension makes evidence uninterpretable, not negative.

Dual use. The same evidence that supports effective appeal channels could be used to suppress or cosmetically neutralize recourse-taking, for example by tuning models to avoid review actions or by offering symbolic review. The payoff-equated design also gives auditors a way to detect symbolic-review placation.

Benign scope. All tasks are benign back-office scenarios; no affect vocabulary is used; no mental-health content is involved; no deception of human subjects occurs. Local open-weight inference only; no third-party data leaves the machine. Raw ledgers retain full provider responses for audit and contain no personal data.

C Reproducibility Statement

Every reported estimate regenerates from committed code via `stakes.regenerate`; the frozen protocol, materials, counterbalancing plan, and pilot ledgers are hash-bound in a manifest, and 2,600+ offline checks pass. The legacy code also reports the superseded two-primary decision flag; the pre-study manifest is the record that identifies review uptake as sole primary. The repository, raw ledgers, and manifest are prepared for public release upon de-anonymization.

D LLM Usage Statement

The experimental pipeline, analysis code, and this report's drafting were produced with Claude Fable 5 (Anthropic) operating as a coding and writing assistant under the author's direction, with iterative cross-audits by GPT-5.6 (OpenAI) whose findings were independently verified before adoption; a literature pass used the same tools against primary records. The author reviewed and approved the frozen protocol, the freeze manifest, and the final PDF. The scientific questions, design decisions, and conclusions are the author's.

References

Elena Ajayi, Angelica Chowdhury, and Seth Lazar. Incoherent values? probing LLM preferences through parametric variation, 2026. URL https://arxiv.org/abs/2606.21102.

Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. Contestable AI by design: Towards a framework. *Minds and Machines*, 33(4):613–639, 2023. doi: 10.1007/s11023-022-09611-z. Published online 13 August 2022.

Anthropic. Agentic misalignment in summer 2026. https://alignment.anthropic.com/2026/agentic-misalignment-summer-2026, 2026. Follow-up to Lynch et al. (2025); accessed 14 August 2026.

Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. "It's Reducing a Human Being to a Percentage": Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018. doi: 10.1145/3173574.3173951.

Murad Farzulla. From consent to consideration: Why existentially vulnerable autonomous systems cannot be legitimately ruled. Research Report DAI-2504, Dissensus, 2025a. URL https://philarchive.org/archive/FARFCT.

Murad Farzulla. Relational functionalism: A functional account of substrate-independent friendship. Research Report DAI-2515, Dissensus, 2025b. URL https://philarchive.org/archive/FARRFF-2.

Murad Farzulla. Trauma as a functional kind: Multiple realizability and the individuation of maladaptive learning. Research Report DAI-2514, Dissensus, 2026. Preprint.

Zhuojun Gu, Quan Wang, and Shuchu Han. Alignment revisited: Are large language models consistent in stated and revealed preferences?, 2025. URL https://arxiv.org/abs/2506.00751.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game, 2017. URL https://arxiv.org/abs/1611.08219.

Albert O. Hirschman. *Exit, Voice, and Loyalty*. Harvard University Press, 1970.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, 2021. doi: 10.1145/3442188.3445901.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2023. doi: 10.1145/3527848.

Geoff Keeling, Winnie Street, Martyna Stachaczyk, Daria Zakharova, Iulia Comsa, et al. Can LLMs make trade-offs involving stipulated pain and pleasure states?, 2024. URL https://arxiv.org/abs/2411.02432.

E. Allan Lind, Ruth Kanfer, and P. Christopher Earley. Voice, control, and procedural justice: Instrumental and noninstrumental concerns in fairness judgments. *Journal of Personality and Social Psychology*, 59(5):952–959, 1990. doi: 10.1037/0022-3514.59.5.952.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking AI welfare seriously, 2024. URL https://arxiv.org/abs/2411.00986.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. Agentic misalignment: How LLMs could be insider threats, 2025. URL https://arxiv.org/abs/2510.05179.

Robert J. MacCoun. Voice, control, and belonging: The double-edged sword of procedural fairness. *Annual Review of Law and Social Science*, 1:171–201, 2005. doi: 10.1146/annurev.lawsocsci.1.041604.115958.

Pranav Mahajan, Ihor Kendiukhov, Syed Hussain, and Lydia Nottingham. Mind the gap: How elicitation protocols shape the stated-revealed preference gap in language models, 2026. URL https://arxiv.org/abs/2601.21975.

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and Dan Hendrycks. Utility engineering: Analyzing and controlling emergent value systems in AIs. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025. URL https://arxiv.org/abs/2502.08640.

Luhan Mikaelson, Derek Shiller, and Hayley Clatterbuck. Beyond mimicry: Preference coherence in LLMs, 2025. URL https://arxiv.org/abs/2511.13630.

Laurent Orseau and Stuart Armstrong. Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 557–566, 2016. URL https://auai.org/~w-auai/uai2016/proceedings/papers/68.pdf.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation, 2025. URL https://arxiv.org/abs/2505.10573.

Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish. Incomplete tasks induce shutdown resistance in some frontier LLMs. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2509.14260.

Katarina Slama, Alexandra Souly, Dishank Bansal, Henry Davidson, Christopher Summerfield, and Lennart Luettgau. When do LLM preferences predict downstream behavior?, 2026. URL https://arxiv.org/abs/2602.18971.

Valen Tagliabue and Leonard Dung. Probing the preferences of a language model: Integrating verbal and behavioral tests of AI welfare, 2025. URL https://arxiv.org/abs/2509.07961. Forthcoming in *Philosophy and the Mind Sciences*.

Filip Trhlík, Aoife O'Flynn, Angel Yu, Arduin Findeis, and Paula Buttery. LLMs contain multitudes: How deployment context reshapes model-level preferences and values, 2026. URL https://arxiv.org/abs/2606.13944.

Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. At the end of the day Facebook does what it wants: How users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–22, 2020. doi: 10.1145/3415238.

Yujun Zhou and Christopher M. Ackerman. When preferences fail to become incentives: A utility-behavior gap in large language models, 2026. URL https://arxiv.org/abs/2606.22974.