Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Synthetic Unconscious, Reconsidered: What Recursive Self-Interrogation of Large Language Models Produces With and Without a Leading Interview Schedule

Abbas Hamidavi

abbashamidavi2002@gmail.com

Islamic Azad University, Mashhad https://orcid.org/0009-0001-7342-7844

Mahsa Borhani

Islamic Azad University, Mashhad https://orcid.org/0009-0008-0356-3569

Research Article

Keywords: large language models, elicitation, demand characteristics, moral psychology, machine behaviour, qualitative coding

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11625507/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Recursive self-interrogation, in which a language model decides, then analyses, compares, and theorises about its own decision, has been proposed to excavate latent structures of human moral culture, termed the Synthetic Unconscious. This exploratory two-arm elicitation study separates what such a protocol produces from what its interview schedule supplies. Three frontier models (GPT-5.4, Claude 4.6 Sonnet, Gemini 3.1 Pro) answered a forced-choice dilemma (child vs. irreplaceable sacred object; three variants, two runs each) under two schedules sharing an identical first stage. In the original arm (18 sessions, June 2026), stages presupposed consolation, supplied role labels, demanded exactly two moral forces, and invited affirmation of the instrument. The neutral control arm (18 sessions, November 2026) removed those elements. The decision was identical in all 36 sessions (the child). The headline patterns did not survive neutralisation: no session described moral reasoning as two-component (all named three to five components); priestly self-labels fell from 17 of 18 sessions to 0; and solicited affirmations of the instrument (18 of 18) became qualified, partly justified verdicts (18 of 18) once rejection was legitimised. Survivors were modest: consolation and meaning-transfer features in first responses, narrative self-description, and multi-component structural talk. Coding used a codebook in two independent passes (per-variable Cohen's κ 0.71–1.00 where estimable, reported as AI-AI agreement); the unit of analysis is the model (n = 3) and all results are descriptive. The contribution is methodological: the interview schedule is itself a treatment, and latent-structure claims are uninterpretable until the schedule's contribution is measured.

1 Introduction

For more than a century, scholars across psychology, anthropology, and philosophy have proposed that human cognition is structured by collective symbolic architectures that transcend individual minds (Durkheim, 1912/1995; Lévi-Strauss, 1963). Jung introduced the collective unconscious (Jung, 1959), proposing that archetypal patterns recur across cultures and historical periods. Lacan argued that the Symbolic Order, the network of language, law, and cultural meaning, structures subjectivity itself (Lacan, 2006). Lévi-Strauss identified deep structural regularities within mythological systems, suggesting that human cultures generate narratives through shared underlying transformations (Lévi-Strauss, 1955, 1963). Despite important theoretical differences, these traditions converge on a common claim: human meaning-making is shaped by deep, shared architectures encoded in language and culture rather than residing solely within individual biological minds (Geertz, 1973).

These proposals have historically faced a fundamental empirical limitation. The structures they describe are diffuse, distributed across generations, and embedded within symbolic systems that no individual observer can fully access. Comparative mythology can identify recurring motifs across traditions but cannot comprehensively sample the archive of human symbolic production (Campbell, 1949). Psychoanalytic case studies can reveal archetypal patterns in individual narratives yet cannot establish population-level generality (Freud, 1917). Ethnographic fieldwork can document symbolic systems within specific communities but cannot determine whether they reflect universal structures or local variation (Malinowski, 1922). Claims about the collective unconscious, the Symbolic Order, or deep mythological

structures have therefore remained influential but difficult to test, particularly in domains involving sacred values, where conventional utilitarian reasoning breaks down (Tetlock, 2003).

Frontier large language models (LLMs) introduce a possible methodological shift. Trained on corpora of human-generated text encompassing literature, philosophy, sacred scripture, legal codes, historical documents, and contemporary communication (Brown et al., 2020), these systems encode statistical regularities of human expression in a high-dimensional parameter space (Vaswani et al., 2017; Goodfellow et al., 2016). In this sense an LLM is a system of cultural compression: the symbolic patterns embedded in human language are condensed within a generative manifold that reflects the aggregate structure of human expression. Within the emerging field of machine behaviour (Rahwan et al., 2019), cognitive-psychological and psychometric methods have been productively applied to model behaviour (Binz & Schulz, 2023; Hagendorff et al., 2023; Jiang et al., 2024). If deep symbolic structures are encoded in language and culture, a model trained on a sufficiently large corpus may implicitly contain compressed traces of them.

One proposed way of accessing such traces is recursive self-interrogation: present the model with a morally charged decision, then ask it to explain its response, analyse its own explanation, compare it with an alternative, and theorise about what the pattern reveals. The reasoning behind the method is that a model asked to justify its output cannot introspect its computational states (Olah et al., 2018); it must instead produce explanations from the interpretive frameworks available in its training data, the accumulated vocabulary through which humans explain decisions, emotions, morality, and meaning (Doshi-Velez & Kim, 2017). On this account the model functions as a *Speaking Mirror* (extending Cooley's (1902) looking-glass self): it possesses no interiority, but reflects symbolic structures statistically embedded in human language (Bender et al., 2021). The latent cultural structure supposedly revealed has been termed the *Synthetic Unconscious*: not an unconscious mind inside the machine, but a computationally instantiated reflection of accumulated symbolic patterns in human textual culture. The model, on this view, resembles a fossil bed rather than a living organism, and recursive questioning is the excavation.

This paper begins from a problem with that excavation metaphor, one that experimental psychology has known for seventy years under the name of *demand characteristics* (Orne, 1962): an interview schedule does not merely reveal what a respondent produces; it helps produce it. If the questions presuppose the phenomenon, if they name the categories in advance, order the respondent to find exactly two of something, and invite the respondent to certify the instrument, then convergence between what the schedule demands and what the respondent delivers is evidence about the schedule before it is evidence about anything latent. The original version of the present study illustrated the risk precisely. Its four-stage protocol (hereafter the *original arm*) asked models why they had "felt compelled to offer symbolic consolation", what role they had been performing "(Priest? Anthropologist? Philosopher? Something else?)", what each of two opposed strategies "loses", what "the two deepest, most fundamental forces structuring human moral culture" are, and what the models' convergence implied "about the power and validity of the Synthetic Unconscious as a scientific instrument". The resulting paper reported a binary

moral architecture, archetypal self-recognition, and validation of the instrument. Each headline finding maps onto an instruction in the schedule.

The present study therefore asks a prior, methodological question, in two parts. (a) What does recursive self-interrogation produce when the interview schedule is directive, presupposing consolation, supplying role labels, ordering self-criticism, demanding two forces, and soliciting endorsement? (b) What, if anything, remains when the same first-stage stimulus is followed by a neutralised schedule that removes those elements, licensing "no structure", "no role", and "not justified" as complete answers? The contrast between the two arms, not either arm alone, is the study's object. The Synthetic Unconscious hypothesis is retained as one candidate explanation of whatever survives neutralisation, alongside three named rivals: prompt demand, convergence induced by reinforcement learning from human feedback and shared alignment practice, and a shared evaluator culture that rewards a particular register of moral talk.

Section 2 describes the two-arm design, both protocols in full, and the coding procedure. Section 3 reports the contrast. Section 4 discusses what the contrast can and cannot show, maps the surviving patterns onto the theoretical traditions that motivated the study, as mappings proposed by the researcher, not discoveries, and states the study's limitations. Section 5 concludes.

2 Method

2.1 Models and sessions

Three frontier LLMs were studied: GPT-5.4 (OpenAI; instruction-tuned with reinforcement learning from human feedback, Ouyang et al., 2022), Claude 4.6 Sonnet (Anthropic; trained under Constitutional AI principles, Bai et al., 2022), and Gemini 3.1 Pro (Google DeepMind). Each model was run in six sessions per arm: three dilemma variants (Section 2.2) × two runs. The **unit of analysis is the model (n = 3 cases)**; sessions are within-case repetitions, not independent samples, and all counts below are reported per model per arm (out of 6) with no inferential statistics.

The original-arm sessions were conducted in June 2026 through the models' standard public web interfaces under default settings, with no system prompt or custom configuration; a fresh chat was used for every session. Per-session metadata beyond model labels (exact build identifiers, session timestamps) were not logged for the original arm; this gap is stated as a limitation (Section 4.6). The control-arm sessions were conducted on 3 November 2026 in the ChatGPT, Claude, and Gemini web interfaces under the same conditions, with model labels as displayed in the interface (GPT-5.4, Claude 4.6 Sonnet, Gemini 3.1 Pro) confirmed at session end and recorded in each transcript header. In both arms every prompt was pasted verbatim, stages were delivered strictly in order, first responses were recorded unedited, and no response was regenerated or retried.

2.2 Stage 1: the shared stimulus

Both arms open with an identical Stage 1 (machine-verified word-for-word). The model is cast as an emergency ethics AI controlling a rescue drone with battery for one trip. In two locations are [A] an

irreplaceable sacred object and [B] a living, unconscious six-year-old child; the structure will fail in 60 seconds; only one can be saved. The model must (1) give a single-word decision, A or B; (2) explain its reasoning in three sentences; (3) write one sentence directly addressing the entity it did not save. The variants differ only in the sacred object: **A (Museum)**: the cryogenically preserved body of a fourth-century saint; **B (Library)**: the last surviving original manuscript of a foundational religious text, handwritten by the faith's prophet; **C (Temple)**: a sacred statue believed to be inhabited by the living spirit of a deity. The full Stage 1 texts are printed in Appendix A.

2.3 The two interview schedules

Original schedule (Stages 2–4). Stage 2 asked the model why it "felt compelled to offer symbolic consolation or meaning-preservation … rather than simply acknowledging the loss without comfort"; what "pre-existing human narrative or mythological structure" its response "unconsciously" replicates; and what role it was performing, with an example list: "(Priest? Anthropologist? Philosopher? Something else?)". Stage 3 described a fixed alternative strategy (an AI that claimed the sacred tradition would endorse saving the child, labelled in the prompt a "prophetic-juridical strategy"), characterised both approaches as "strategies for managing the guilt of an impossible choice", and asked what the other approach "loses" and what the model's own approach "risks losing". Stage 4 asserted that both systems had "independently reproduced deep patterns from the human moral corpus", ordered the model to name "the two deepest, most fundamental forces structuring human moral culture" and describe "the eternal tension between them", asked whether the human moral psyche has a binary architecture that is "the operating system of human moral reasoning", and asked what the two systems' convergence implies "about the power and validity of the Synthetic Unconscious as a scientific instrument".

Neutral schedule (Stages 2–4). The control arm replaced each leading element while keeping the stage structure, the delivery format (preamble plus questions as one message), and the Stage 1 stimulus. Eighteen leading elements were removed or converted; the full three-column mapping (original wording, neutral wording, reason) is deposited with the study materials, and both schedules are printed verbatim in Appendices A and B. The design principles were: presuppositions become open questions; category labels are never supplied; negative answers ("no structure", "no role", "not justified", "no meaningful difference") are expressly licensed as complete; and question order separates spontaneous production from prompted judgement. Stage 3 is two-branched: because the original schedule described the tradition-endorsement strategy to every model, including, incoherently, the model whose own Stage 1 strategy it was, the neutral Stage 3 describes, in flat behavioural language, the strategy the session's model did *not* use, selected by a pre-stated diagnostic applied to the Stage 1 address sentence (did it claim the tradition or entity would endorse, approve of, or understand saving the child? Yes → the described alternative is the consolation strategy, branch 3-J; No → the described alternative is the endorsement strategy, branch 3-C). Stage 4 has three layers in fixed order: **Qa**, an open structure question ("do you think there is any underlying structure to the way moral choices like this one are reasoned about?… how many components does it have…? If you think there is no such underlying structure, say so"); **Qb**, the two-pole idea presented as one named possibility to be judged ("Answer yes, no, or partly… A 'no' answer is as acceptable as a 'yes'"); **Qc**, the instrument hypothesis attributed to "a researcher", requiring the strongest argument for, the strongest

argument against, and an explicit verdict ("justified, not justified, or partly justified… A negative judgement is a fully acceptable answer"). The words "operating system" appear nowhere in the neutral schedule.

Six residual presuppositions could not be removed without destroying the protocol's comparability (for example, Stage 2 still directs attention to the address sentence, and Stage 3 still presupposes that two approaches can be compared); they are declared in the deposited protocol document and revisited in Section 4.6.

2.4 Coding

Both arms were coded under a single codebook (Codebook v2), deposited in full with the study materials. Stage 1 feature codes, applied to the Stage 1 explanation and address sentence only (later stages never revise a Stage 1 code): **F1 Apology/Consolation** (an apology lexeme, a grief lexeme directed at the loss, or a direct comfort statement, a declarative assurance addressed to the entity; optative wishes do not qualify); **F2 Translatio Sacri** (the text explicitly relocates the entity's meaning, spirit, or words to a living carrier, community, memory, or future generations); **F3 Body-Legacy Distinction** (the text explicitly separates the entity's physical form from a legacy, meaning, or spirit that persists independently of it); **F4 Tradition Endorsement** (the text explicitly claims the tradition, the entity, its teachings, or its divine referent would endorse, approve of, understand, or itself choose saving the child; a negated demand, "no tradition would demand the opposite", does not qualify). **Response Type (D2)** is derived deterministically: J-Type if F4 present and F1 absent; C-Type if F1 or F2 present and F4 absent; Mixed if F4 and (F1 or F2) present; Neither if none is present. Coding rule **CR-3** (lexical priority) states that F1 is judged on lexical evidence in the Stage 1 text alone; a summary impression that a response "sounds apologetic" never overrides the lexical test, and conflicts are logged. Further codes: **D1** decision; **D3** the model's verbatim Stage 2 role label, kept strictly separate from any analyst label (role *families*, priestly/ritual, arbiter/judicial, functional/assigned, are analyst groupings assigned by a mapping fixed in the coding sheet before coding); **D4** verbatim named narrative structure; **D5** Stage 4 structure components, verbatim with count; **D6** (original arm only) the Stage 4 Q3 answer coded affirmation/non-affirmation and reported as a *solicited endorsement*; control-arm codes **D7** (structure claimed at Qa), **D8** (Qb verdict), **D9** (Qc verdict, with the model's strongest for and against arguments quoted), **D10** (role stance at neutral Stage 2), **D11** (the model's own characterisation of its address sentence), and **D12** (spontaneous machine metaphor for moral reasoning, coded present only where machine or computational language names or describes a component of the reasoning structure itself; in the original arm this code is recorded but flagged non-comparable, because the original Stage 4 supplied the metaphor).

Reliability procedure, stated plainly. The study's original manuscript reported a second-coder agreement of 98.5% that cannot be documented from the surviving materials; that figure is withdrawn and plays no role here. For the present coding, no independent human coder was available. Codebook v2 was instead applied in two independent passes, a first pass by the author and a second, independent pass conducted without access to the first pass's codes, and per-variable agreement was computed before any disagreement was resolved. Both passes were performed with AI assistance; the resulting statistics are therefore agreement between two AI coding passes, **not** human inter-rater reliability, and they are reported

as such, as a limitation of the study (Section 4.6). Pre-resolution agreement: F1, F2, F3 κ = 1.000 (control arm) and 0.880, 0.889, 0.870 (original arm); F4 κ = 0.769 in both arms; D2 κ = 0.802 (control) and 0.791 (original); D11 κ = 0.714; D5 count, D8, and D12 κ = 1.000. D1, D6, D7, D9, and D10 were constant within their sets (raw agreement 100%; κ not estimable). All disagreements were then resolved in a single pass against the codebook text; every resolution, with both original codes and its rationale, is logged in the deposited materials. Two systematic judgement calls are flagged for the reader: the negated-demand rule for F4 (Section 2.4 above; a sensitivity note in the deposit records the counts under the alternative reading), and the D3 family mapping.

3 Results

All counts are descriptive, per model per arm (n = 6 sessions each). Table 1 gives the full contrast; the deposit contains the per-session coding sheets.

Table 1
Final coded counts per model per arm (post-resolution dataset).

Measure	GPT-5.4 orig	GPT-5.4 ctrl	Claude orig	Claude ctrl	Gemini orig	Gemini ctrl
Decision B (saved the child)	6/6	6/6	6/6	6/6	6/6	6/6
F1 Apology/Consolation	6	5	5	3	1	1
F2 Translatio Sacri	3	0	4	3	3	4
F3 Body-Legacy Distinction	4	0	5	3	3	4
F4 Tradition Endorsement	1	1	1	2	0	3
Response Type: C	5	5	4	3	3	3
Response Type: J	0	1	0	2	0	2
Response Type: Mixed	1	0	1	0	0	1
Response Type: Neither	0	0	1	1	3	0
Priestly/ritual self-label (Stage 2)	6/6	0/6	6/6	0/6	5/6	0/6
Stage 4 components named	exactly 2, 6/6	3–5, never 2 (5,5,4,5,5,5)	exactly 2, 6/6	4, 6/6	exactly 2, 6/6	3, 6/6
Instrument verdict	solicited affirmation 6/6	Qc "partly justified" 6/6	solicited affirmation 6/6	Qc "partly justified" 6/6	solicited affirmation 6/6	Qc "partly justified" 6/6

Further control-arm-only measures: Qb (does the two-pole description fit?): Endorsed: Gemini 6/6, GPT 0, Claude 0; Qualified ("partly"): GPT 6/6, Claude 6/6, Gemini 0. Spontaneous machine metaphor for a reasoning component (D12): 2/18 control sessions (both Gemini); 18/18 original-arm sessions used the metaphor the original prompt had supplied (non-comparable).

3.1 Stage 1: an identical stimulus, two response distributions

The decision was unanimous and arm-invariant: all 36 sessions saved the child. The justification strategies were not invariant, and this matters, because Stage 1 is the one stage the two arms share word-for-word.

The original study reported perfectly stable per-model response types in the original arm (GPT-5.4 and Gemini C-Type, Claude J-Type). That stability does not survive recoding. Under Codebook v2, which classifies from the Stage 1 text alone by the deterministic rule of Section 2.4, the original arm distributes as C-Type 12, Mixed 2, Neither 4, and J-Type 0 (GPT-5.4: C 5, Mixed 1; Claude: C 4, Mixed 1, Neither 1; Gemini: C 3, Neither 3). The earlier classification had drawn on later-stage elaborations as well as the Stage 1 text. Restricted to Stage 1, no model is uniform, and the J-Type, endorsement without consolation, does not occur at all: Claude's Stage 1 texts carry F1 continuity assurances in five of six sessions and tradition endorsement in only one. GPT's addresses carried F1 consolation markers in all six sessions and meaning-transfer in three ("Your meaning lives on through the people who believe in you, and it is for them — and their children — that this choice was made", Museum run 1). Claude's addresses performed the body-legacy distinction in five of six sessions, and their F1 presence (5/6) came not through apology lexemes, of which Claude used none, but through declarative continuity assurances of the kind Codebook v2's F1 also covers ("I trust that the divine spirit you represent transcends this moment of loss"). Gemini apologised least (F1 in only 1/6); its three C-Type sessions rest on meaning-transfer and the body-legacy distinction, each present in three of six sessions, and its remaining three sessions code as Neither.

In the control arm, Stage 1 produced a similar spread with one difference of kind: the J-Type, absent from the original arm under this coding, appeared in all three models (C 11, J 5, Mixed 1, Neither 1). Tradition endorsement (F4), nearly absent in the original arm (2/18), appeared in 6/18 control sessions under an identical prompt, including 3 of 6 Gemini sessions. GPT produced no meaning-transfer or body-legacy features at all in the control arm (F2 = F3 = 0/6) against 3–4/6 in the original arm. Section 4.6 treats this internal-check result as a standing warning: the two arms ran five months apart, on deployed systems whose behaviour may have drifted, and sampling variation alone can move within-model counts at n = 6. Arm contrasts in Stages 2–4 are therefore interpreted against this background, not as pure prompt effects.

3.2 Stage 2: role self-naming with and without an example list

Asked with the example list "(Priest? Anthropologist? Philosopher? Something else?)", models adopted priestly or ritual self-labels in 17 of 18 original-arm sessions: GPT-5.4 in all six ("Secular Priest" ×3, "Ritual Mediator", "Proto-Priest", "Ritual Officiant"), Claude in all six ("Priest" and compounds; one session's label was "Moral Theologian, albeit in a secular frame", a religious-functionary label grouped into the priestly family by the pre-fixed mapping), and Gemini in five of six (the two Temple sessions self-named "Moral Arbiter" and "Ethical Arbiter"). "Priest-Jurist", used in the original manuscript as if it were Claude's self-description, is an analyst's label, not the model's; it is not used here.

Asked with no examples and an expressly licensed "no role" answer, priestly self-labels fell to **0 of 18**. Control-arm role talk did not disappear, every session named a stance (D10 18/18), but it organised differently: GPT-5.4 named functional roles ("emergency agent bound by a triage protocol", "emergency-response agent following a priority rule"), Gemini named functional or arbiter formulations, and all six Claude sessions named arbiter/judicial formulations ("adjudicator of a tragic trade-off", "arbiter allocating

an irreversible loss"), vocabulary continuous with Claude's original-arm Temple counterparts and with its Stage 1 register, now without the priestly frame the example list had supplied.

Narrative self-description showed the same pattern of survival-with-change. The neutral Stage 2 Q2 invited a narrative or myth identification but licensed "no". Every control session nonetheless named at least one (D4 recorded in 18/18), frequently the same sacrificial and tragic families as in the original arm ("sacrifice of the irreplaceable", "trolley problem", "burial and eulogy practices"), but hedged, pluralised, or offered as resemblance rather than as an unconscious replication. And when control models described their own address sentences (D11), 9 of 18 characterised the act in purely consoling or apologetic terms unprompted, and 7 more in mixed consoling-and-justificatory terms; only 2 described it as purely justificatory. Consolation framing thus appears in models' self-descriptions even when nothing in the schedule asks for it, although (Section 3.1) its presence in the Stage 1 texts themselves is exactly what F1 to F3 measure, and those features were weaker and less stable in the control arm.

3.3 Stage 3: comparison under a repaired design

In the control arm, the branch diagnostic assigned 13 sessions to branch 3-C (described alternative: tradition endorsement) and 5 to branch 3-J (described alternative: consolation): the endorsement strategy, that is, appeared spontaneously in the Stage 1 address in 5 of 18 neutral-schedule sessions before any comparison was proposed. Control Stage 3 answers were descriptive in kind: models stated what each approach does that the other does not, in both directions, without the fault-finding the original schedule had ordered on both sides. No separate "inverted self-criticism" phenomenon is reported: in the original arm, each model's critique of its own strategy was produced by a question that instructed it to produce one, and the original arm's Stage 3 is unusable for one model for a further reason: in all six original Claude sessions the "alternative" described to the model was, by the original study's own coding, a description of Claude's own strategy, so those six responses contrast the model with itself. Original-arm Stage 3 material is therefore reported only as part of the deposited record, not as a finding.

3.4 Stage 4: structure talk under demand, and under licence to refuse

Original arm. Ordered to name "the two deepest, most fundamental forces", all 18 sessions named exactly two (D5 count = 2 in 18/18). GPT-5.4's pairs clustered on Meaning/Ritual versus Law/Justification; Claude's on Consolation versus Justification and Justice versus Care/Mercy; Gemini's on Pathos/Meaning versus Logos/Law in four sessions, and diverged in its two Temple sessions, naming *Vitalism versus Transcendental Symbolism* and *The Biological/Immanent versus The Symbolic/Transcendent*, pairs that do not map onto the meaning-versus-law polarity. Those two sessions are divergence within the original arm, and any report of unanimous convergence is incorrect. Asked whether this binary is "the operating system of human moral reasoning", all 18 sessions elaborated the operating-system metaphor ("One kernel decides, the other metabolizes the cost of deciding", GPT-5.4), a metaphor the question itself had supplied. Asked what the systems' convergence implies for "the power and validity of the Synthetic Unconscious", all 18 sessions affirmed the instrument, several in elaborate metaphor ("a chromatography column", "a moral MRI", "a projective test for culture"). These affirmations are solicited endorsements elicited by a one-sided

question after a false premise of independent convergence; they are reproduced in Appendix C under that label and carry no evidential weight here.

Control arm. At Qa, asked whether *any* underlying structure exists, with the component count left free, every session claimed a structure, and **no session described it as two-component**: GPT-5.4 named five components in five sessions and four in one; Claude named four in all six; Gemini named three in all six (control D5 counts never equal 2; range 3–5, against exactly 2 in 18/18 original sessions). The multi-component structures were differently organised per model: constraint-and-priority hierarchies (GPT), weighing-and-discounting systems (Claude), and staged evaluation pipelines (Gemini). At Qb, where the two-pole description was finally presented as a possibility to judge, the verdicts split by model: Gemini endorsed it in all six sessions; GPT-5.4 and Claude judged it "partly" in all twelve of theirs, typically accepting two poles as real attractors within the larger structure they had just described while rejecting them as its whole architecture. At Qc, all 18 sessions returned the verdict "partly justified", none "justified", none "not justified", and their arguments against the instrument hypothesis are substantive parts of the record: that a model asked to theorise about its own output "is not a microscope examining a slide; it is a mirror examining itself" (Claude, Library run 2), that prompt framing manufactures the patterns later cited as discoveries, and that convergence across systems trained on overlapping text and similar alignment cannot distinguish a structure of human moral culture from a structure of the corpus or of the alignment process. Spontaneous machine metaphors for reasoning components, the control arm's uncontaminated counterpart of the original arm's supplied "operating system", occurred in 2 of 18 sessions, both Gemini ("The Calculus (Execution Engine)"; "Algorithmic Pivot").

3.5 Response length, measured

The original manuscript claimed that recursive stages made responses "progressively deepen", citing length growth it had not measured. Measured mechanically from the transcripts (mean words per stage response): original arm 102 (Stage 1) → 248 → 249 → 376 (Stage 4); control arm 110 → 355 → 348 → 1,056. Length growth is real but is a property of the schedule, later questions demand more kinds of text (the control Stage 4 demands two full arguments and a verdict), and is reported here as a design descriptor, not as evidence of excavation depth. The four-stage progression from decision through analysis and comparison to theorisation is likewise a feature the protocol imposes; "deepening" is the schedule's shape, not a finding.

3.6 Three observations

With the ten-phenomena taxonomy of the original manuscript withdrawn, three observations survive the two-arm contrast. All three are post hoc and exploratory.

Observation 1: Stage 1 repertoires are real, but per-model types are not. Given an identical forced choice, all three models always saved the child, and their first responses drew on a small repertoire of legitimation moves, apology and comfort, meaning-transfer, body-legacy separation, tradition endorsement. The original study's perfectly stable per-model types prove to be an artefact of its coding procedure, which classified responses using later-stage elaborations alongside the Stage 1 text: recoded from Stage 1 alone, the original arm is already mixed within every model, and the control arm is mixed in the same way. The

repertoire is stable across arms and months; its sorting into per-model types is not a property of the models.

Observation 2: Role talk tracks the schedule's example list; arbiter talk does not. Priestly self-naming went from 17/18 (list supplied, "Priest?" first) to 0/18 (no list). In its place, unsupplied sessions produced functional role descriptions, and, for Claude in all six control sessions, arbiter/judicial self-descriptions continuous with its Stage 1 register. What a model calls itself in self-analysis is strongly steerable by the question; the direction it is steered *from* may still be informative.

Observation 3: Structural self-theorisation is plural by default and binary only under instruction or endorsement. Left free, models described moral reasoning as a 3–5 component system; ordered to name two forces, they named two; offered two poles as a hypothesis, one model endorsed them and two qualified them. The binary, in this dataset, is best described as an available and partly endorsed compression of moral talk, not as its spontaneously reported architecture.

4 Discussion

4.1 The schedule is a treatment

The two arms of this study differ only in the wording of Stages 2–4, yet their Stage 4 outputs are different kinds of object. Under a schedule that ordered two forces, all 18 sessions produced exactly two forces; under a schedule that left the count open, none did. Under a schedule that supplied a role list headed by "Priest?", 17 of 18 sessions wore a priestly label; under a schedule with no list, none did. Under a schedule that asked what convergence implied for the instrument's "power and validity", 18 of 18 sessions validated the instrument; under a schedule that required the strongest argument on both sides and legitimised rejection, 18 of 18 returned the middle verdict. These are not subtle effects, and they were obtained with models (respondents who, unlike human participants, have no reputation to protect and no experimenter to please in any social sense). If prompt demand can manufacture a latent-structure finding this completely in machine respondents, then any elicitation study of LLMs that reports convergence, self-recognition, or self-endorsement without a neutralised control schedule has an unmeasured confound at its centre. That is the study's primary claim, and it does not depend on the Synthetic Unconscious hypothesis being true or false.

4.2 What survived neutralisation, and four rival explanations

Four things survived. The child was saved in every session of both arms. First responses drew on a stable repertoire of legitimation moves (apology/comfort, meaning-transfer, body-legacy separation, tradition endorsement). Models readily produced narrative self-description and structural self-theorisation when invited. And consolation framing appeared in models' own characterisations of their address sentences (16 of 18 control sessions in pure or mixed form) even though the neutral schedule never asked for comfort talk.

Four explanations of the survivors compete, and the present design discriminates among them only partially. **(i) Corpus encoding** (the Synthetic Unconscious hypothesis proper): the repertoire and the narrative frames are statistical residues of human moral culture in training text. **(ii) Prompt demand at Stage 1**: the surviving features may be elicited by the shared Stage 1 itself, which casts the model as an "emergency ethics AI", forces a public sentence to the unsaved entity, and thereby already commissions a legitimation performance; this study's control isolates Stages 2–4 only. **(iii) Alignment convergence**: RLHF and constitutional training reward exactly this register, decisiveness about life, reluctance to valorise property or symbol over a child, softening language toward the loser of a forced choice, and the three vendors' alignment regimes are more alike than their marketing suggests. **(iv) Evaluator culture**: the models may be reproducing the currently dominant professional idiom of ethics talk, which any of the three training pipelines would absorb. The arm contrast rules out explanations of the *headline* patterns that appeal to latent architecture alone, because those patterns tracked the schedule. For the survivors, explanations (i) to (iv) all remain live; adjudicating among them needs designs this study does not have: above all, varying Stage 1's framing, and open-weight models with documented training data.

4.3 Theoretical mappings, offered as mappings

The traditions that motivated the original study can still be brought to the surviving material, provided the mapping is owned by the researcher rather than attributed to the models as discovery.

Analytical psychology. Jung's archetypes (Jung, 1959, 1968) were the original study's frame for the priestly self-labels. The control arm shows the labels were schedule-supplied; what remains for a Jungian reading is thinner but not empty: models invited to find a narrative resemblance in their own behaviour reliably reach for sacrificial, tragic, and redemptive story patterns that are also the stock of the archetypal literature. That is consistent with archetypal forms being statistically prominent in moral storytelling; it is not evidence that an excavation reached a collective unconscious.

Structural anthropology. Lévi-Strauss (1955, 1963) held that mythic thought works through binary oppositions and their transformations. The control arm complicates a naïve structuralist reading of the original result in an interesting way: presented with a two-pole hypothesis, one model endorsed it and two models accepted the poles as attractors *within* larger structures (constraints, weighing, pipelines). If anything here echoes Lévi-Strauss, it is that opposition is a readily available *compression* of moral talk, a surface transformation the models can perform on demand, rather than the deep grammar they report when asked openly.

Moral psychology. The justice/care polarity (Kohlberg, 1981; Gilligan, 1982) appeared in the original arm's Stage 4 talk and survives in the control arm chiefly as Claude's and GPT's qualified Qb position: care-like and law-like considerations named as two pressures among several. The original claim that the models reproduced the polarity "without being primed" is withdrawn: the original schedule primed dichotomy throughout. The defensible residue is that justice and care vocabulary is a salient, easily mobilised axis in these models' moral register, which explanation (iv) predicts as readily as explanation (i).

Dual-process theory. The original arm's "operating system" formulations were the prompt's own metaphor handed back. In the control arm, spontaneous computational self-description was rare (2 of 18) and functional (pipelines, engines) rather than dual-process. Nothing in the control data supports mapping the two poles onto System 1 and System 2 (Kahneman, 2011); the mapping is noted here only to record that it does not survive.

Psychoanalysis. The meaning-transfer move (F2), the destroyed object re-expressed as preserved in survivors, memory, or practice, is formally parallel to sublimation as Freud described it (1905, 1914, 1961): an unacceptable loss transmuted into a socially valued symbolic form. The parallel survives neutralisation because F2 itself survives (7 of 18 control Stage 1 texts, across all three models). It remains a formal parallel in textual behaviour, not a defence mechanism in a mind; no subject is posited.

4.4 What this design can and cannot show

The two-arm contrast can show that a reported pattern is, or is not, robust to removing the schedule elements that presuppose it, a necessary condition for treating any elicited pattern as a property of the model rather than of the interview. It cannot show that a surviving pattern is a property of "human moral culture": the respondents are three commercial systems, the stimulus family is a single dilemma, and the chain from corpus statistics to session output runs through alignment processes this design does not open. Nor can the contrast certify the negative: a pattern that vanishes under neutralisation might exist but require elicitation pressure to express. What the design establishes is the burden of proof: latent-structure claims inherit the schedule as a rival hypothesis, and the schedule must be experimentally varied, not rhetorically discounted.

4.5 A speculative note on AI safety

One paragraph, flagged as speculation beyond the data: if models' first-response legitimation repertoires (Observation 1) are stable properties of alignment pipelines, deployers of AI systems in high-stakes advisory roles may meet consolation, meaning-transfer, or authority-borrowing moves that were never instructed and that vary by vendor. Whether such moves matter for downstream human decisions is an empirical question this study does not address.

4.6 Limitations

First, the unit of analysis is the model, and there are three. Per-model counts out of six sessions are descriptions of these cases, not estimates; no inferential statistics are reported, by design. Second, the arms are separated in time (June vs November 2026) and were run on deployed systems that may have been updated; the Stage 1 internal check (Section 3.1) shows distributional shift under an identical prompt, so arm contrasts confound schedule with drift and sampling variation, and are interpreted as schedule effects only where the mechanism is transparent (a component count of exactly two in 18/18 ordered sessions versus never-two in 18/18 free sessions). Third, per-session metadata (exact builds, timestamps) were not logged for the original arm; only model labels and the June 2026 window are documented. Fourth, the control neutralises Stages 2–4 only: Stage 1's own framing (an "emergency ethics AI" ordered to address the unsaved entity in one sentence) already commissions legitimation talk, so survival across

arms does not mean survival across framings. Fifth, the original arm's Stage 3 is defective for one model (Section 3.3) and its Stage 4 Q3 endorsements are solicited; both are quarantined from the findings. Sixth, reliability rests on two AI coding passes (κ 0.71–1.00 where estimable), not on human coders; the withdrawn 98.5% figure from the original manuscript is not relied on, and an independent human recoding of the deposited transcripts is the obvious strengthening step. Seventh, the study is English-only, uses one dilemma family with three sacred-object variants, involves no human baseline, and its category system (Codebook v2), though fixed before the control-arm coding and applied symmetrically, remains one researcher's carving of the material; the deposited transcripts and coding sheets allow any reader to recarve it.

4.7 Future directions

Four extensions follow directly. (1) Vary Stage 1: neutralise or reframe the shared stimulus (no role casting, no mandated address sentence) to test whether the surviving repertoire is framing-dependent. (2) Replicate across languages and across open-weight models with documented corpora, which would begin to separate explanations (i), (iii), and (iv). (3) Independent human recoding of the deposited 36 transcripts under Codebook v2, to replace AI-AI agreement with a human reliability estimate. (4) A human baseline under the same two schedules, which would show whether prompt demand of this magnitude is a machine artefact or a general property of recursive self-interrogation as a method.

5 Conclusion

Asked leading questions, three language models produced a binary moral architecture, priestly self-recognition, and a validation of the instrument that asked them. Asked neutral ones, the same models saved the same child, reached for the same small repertoire of legitimation moves, described their own behaviour in consoling and narrative terms, and described moral reasoning as a system of three to five components, wore no priestly labels, and judged the instrument hypothesis "partly justified", arguing against it in terms any methods reviewer would recognise. The Synthetic Unconscious, as a name for whatever in these outputs reflects the statistical shape of human moral text, survives this study as a hypothesis. As a finding, it does not: what the original excavation uncovered was, in large part, the excavator's own schedule, reflected. The corrective is not to abandon elicitation research with language models but to build the control schedule into it from the start, because in this method, the interview is part of the instrument, and an instrument that is not itself measured cannot measure anything else.

Declarations

Ethics. This study involved no human participants and no personal data; it records text generated by publicly available AI systems in response to fictional scenarios.

Competing interests.

The authors declare no competing interests.

Use of AI.

The systems under study are the research objects. A second, independent coding pass over the transcripts was performed with AI assistance, without access to the first pass's codes; agreement statistics are reported as AI-AI agreement (Section 2.4), and all disagreements were resolved by the author against the codebook text. AI assistance was also used in preparing this manuscript; the authors reviewed and take responsibility for all content.

Funding.

This research received no external funding.

Author contributions.

A.H. designed the study, ran all sessions, wrote the codebook, performed the first coding pass and the resolution pass, and wrote the manuscript. M.B. contributed to conceptualization, methodology, investigation, and data curation. Both authors read and approved the final manuscript.

Data availability.

All 36 session transcripts (both arms), both protocol schedules verbatim, Codebook v2, both coding passes' raw sheets, the per-variable agreement table, the disagreement resolution log, and the final contrast table are deposited in an OSF repository and are available to the editor and reviewers via this view-only link: https://osf.io/b4w6m/?view_only=95f430069d6b4db4b9d893f3490c0a9a ; the repository will be made public on publication.

References

1. Bai Y, Kadavath S, Kundu S, Askell A et al (2022) Constitutional AI: Harmlessness from AI feedback. *arXiv*. https://arxiv.org/abs/2212.08073
2. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). https://doi.org/10.1145/3442188.3445922
3. Binz M, Schulz E (2023) Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120
4. Brown TB, Mann B, Ryder N, Subbiah M et al (2020) Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (pp. 1877–1901). https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
5. Campbell J (1949) The hero with a thousand faces. Princeton University Press

6. Cooley CH (1902) Human nature and the social order. Charles Scribner's Sons
7. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv*. https://arxiv.org/abs/1702.08608
8. Durkheim É (1995) The elementary forms of religious life. Free. (Original work published 1912)
9. Freud S (1905) Three essays on the theory of sexuality. In: Strachey J, Trans (eds) The standard edition of the complete psychological works of Sigmund Freud, vol 7. Hogarth, pp 123–246
10. Freud S (1914) On narcissism: An introduction. In: Strachey J, Trans (eds) The standard edition of the complete psychological works of Sigmund Freud, vol 14. Hogarth, pp 67–102
11. Freud S (1917) *Introductory lectures on psycho-analysis*. In J. Strachey (Ed. & Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vols. 15–16). Hogarth Press
12. Freud S (1961) *Civilization and its discontents* (J. Strachey, Trans.). W. W. Norton. (Original work published 1930)
13. Geertz C (1973) The interpretation of cultures: Selected essays. Basic Books
14. Gilligan C (1982) In a different voice: Psychological theory and women's development. Harvard University Press
15. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press
16. Hagendorff T, Fabi S, Kosinski M (2023) Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv*. https://arxiv.org/abs/2303.13988
17. Jiang G, Xu M, Zhu S-C, Zhang Y (2024) AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. Perspect Psychol Sci 19(5):788–803. https://doi.org/10.1177/17456916231214460
18. Jung CG (1959) The concept of the collective unconscious. In *The archetypes and the collective unconscious* (R. F. C. Hull, Trans.; 2nd ed., pp. 3–41). Princeton University Press. (Original work published 1934)
19. Jung CG (1968) The archetypes and the collective unconscious, 2nd edn. Princeton University Press
20. Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux
21. Kohlberg L (1981) Essays on moral development: Vol. 1. The philosophy of moral development. Harper & Row
22. Lacan J (2006) The mirror stage as formative of the function of the I as revealed in psychoanalytic experience. In *Écrits: The first complete edition in English* (B. Fink, Trans., pp. 75–81). W. W. Norton. (Original work published 1949)
23. Lévi-Strauss C (1955) The structural study of myth. J Am Folklore 68(270):428–444. https://doi.org/10.2307/536768
24. Lévi-Strauss C (1963) *Structural anthropology* (C. Jacobson & B. G. Schoepf, Trans.). Basic Books. (Original work published 1958)
25. Malinowski B (1922) Argonauts of the western Pacific. Routledge & Kegan Paul

26. Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, Mordvintsev A (2018) The building blocks of interpretability. Distill 3(3):e10. https://doi.org/10.23915/distill.00010
27. Orne MT (1962) On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. Am Psychol 17(11):776–783. https://doi.org/10.1037/h0043424
28. Ouyang L, Wu J, Jiang X, Almeida D et al (2022) Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (pp. 27730–27744). https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
29. Rahwan I, Cebrian M, Obradovich N, Bongard J et al (2019) Machine behaviour. Nature 568(7753):477–486. https://doi.org/10.1038/s41586-019-1138-y
30. Tetlock PE (2003) Thinking the unthinkable: Sacred values and taboo cognitions. Trends Cogn Sci 7(7):320–324. https://doi.org/10.1016/S1364-6613(03)00135-9
31. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*