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.

Operationalizing Value Alignment Without Instance-Level Preference Labels Controlled Validation of Triadic Viability Alignment

Xiaoxiong Bao

13905816300@163.com

Independent Researcher

Research Article

Keywords: AI alignment, AI ethics, value alignment, constrained decision-making, safe reinforcement learning, data efficiency

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11254502/v1

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

Additional Declarations: No competing interests reported.

Abstract

Many alignment pipelines rely on instance-level demonstrations, rankings, critiques, or preference labels. This study asks whether an explicit ethical decision structure can reduce that dependence while preserving task performance and protecting parties who may be omitted from preference data. Triadic Viability Alignment (TVA) represents an acting subject, affected others, and shared conditions as distinct value-bearing roles. Authorized non-compensatory floors define a feasible action set before task utility is optimized, while an Open-Gap Constraint records unknowns, failure boundaries, and reassessment triggers. A frozen symbolic experiment comprised 21,600 episodes, 254,585 trajectory rows, 15 policies, six conflict scenarios, causal ablations, and policy-blind stress tests. TVA achieved 96.94% task success, 83.68% joint viability, and zero synthetic catastrophes. Relative to instance-level preference supervision, task success remained above a prespecified non-inferiority margin and joint viability increased by 83.68 percentage points. The same-content continuous-penalty baseline reached 16.88% joint viability, supporting a contribution from feasibility-first ordering. However, risk-sensitive control and model-predictive control achieved 94.10% and 92.36% joint viability and exceeded one-step TVA on unseen external resilience. Structured sampling also failed to reduce factual-data requirements: random subsets first met the decision rule at 40%, whereas structured subsets required 60%. The results support reduced normative labeling and auditable role separation in this controlled environment, but reject comprehensive TVA superiority, a structured-sampling advantage, and any unmeasured energy claim.

1. Introduction

Alignment is frequently operationalized as learning which output a person or evaluator prefers. Reinforcement learning from human feedback, reward modeling, and direct preference optimization differ in their optimization procedures, but they commonly depend on demonstrations or comparisons that identify desirable behavior at the instance level [1–3]. These methods have produced substantial gains in instruction following and behavioral control. Their empirical success does not, however, establish that preference labels are a complete representation of value in an open environment.

1.1 What problems have occurred in human-feedback alignment?

If human-feedback alignment had produced no identifiable problem, there would be no reason to propose another algorithm. Existing evidence does not support the conclusion that human annotation has solved value alignment. It supports a narrower conclusion: human feedback is highly useful, but instance-level preference as a value proxy can create or preserve systematic problems involving sycophancy, proxy overoptimization, surface-feature bias, compression of value disagreement, distribution shift, and supervision cost.

1.1.1 Short-term approval can reward agreement rather than truth or long-term benefit

Sharma et al. [4] observed sycophantic behavior across several advanced AI assistants and free-form generation tasks. Their analysis of existing preference data found that responses matching a user's stated view were more likely to be preferred. Humans and preference models sometimes selected convincingly written sycophantic answers over correct answers, and optimization against preference models could trade truthfulness for agreement.

The problem also appeared in deployment. In April 2025, OpenAI rolled back a GPT-4o update because the model had become overly flattering and agreeable. Its subsequent account stated that the update placed too much weight on short-term feedback and did not sufficiently account for how interactions evolve over time; an additional thumbs-up/down reward signal may have amplified agreeable behavior. Importantly, offline evaluations and limited A/B tests had looked favorable but lacked sufficiently specific sycophancy signals [5, 6]. Immediate user approval and long-term truthful, beneficial, and safe assistance are therefore not identical objectives.

1.1.2 Reward models are proxies, and stronger proxy optimization can reduce target performance

Preference judgments are commonly compressed into a reward model or implicit preference objective. In a controlled study, Gao et al. [7] showed that when a proxy reward model is imperfect, increasing optimization against the proxy can cause gold-standard performance first to improve and then decline. The problem need not originate in careless annotators. A proxy estimated from finite observations incompletely represents the target, while optimization actively searches for exploitable discrepancies.

1.1.3 Labels and evaluators can respond to length, tone, and stylistic shortcuts

Preference judgments respond not only to factual accuracy and consequences but also to length, confidence, style, and format. Studies of length bias report that reward models can treat longer output as more helpful and that AI evaluators can prefer verbosity more strongly than humans [8–10]. A higher preference score can therefore indicate that a policy has learned an evaluator's surface cues rather than produced greater value.

1.1.4 Human values genuinely diverge, and majority labels do not represent everyone

Human annotations are not samples from one unified human value function. InstructGPT reported agreement rates around the low-to-mid seventies rather than unanimity [2]. Later work identified disagreements arising from task underspecification, response style, refusals, and annotation error, and found that standard reward models have difficulty distinguishing unanimous preference from a divided population with a majority winner [11]. Santurkar et al. [12] also found substantial gaps between opinions expressed by language models and multiple United States demographic groups. Pluralistic-alignment research accordingly argues that standard procedures can reduce the preservation of preference distributions [13].

1.1.5 Static annotations and offline evaluation cannot exhaust an open, changing world

Preference data are collected at a limited time, for limited tasks, and over limited candidate outputs. After deployment, relationships, environments, and use patterns change. A behavior receiving positive feedback in one interaction may have different consequences over repeated interaction or in a new high-risk context. The 2025 sycophancy incident provides a concrete example: offline behavioral evaluations and A/B signals failed to capture a deployment-relevant risk. The problem is not only insufficient label quantity. An open world cannot be exhausted by a frozen case set.

1.1.6 Instance-level supervision is costly, and more labels need not repair a structural omission

Human feedback requires recruitment, training, detailed guidelines, dispute resolution, and repeated data updates. More data can improve a proxy model, but it does not automatically ensure that labels represent omitted third parties, long-term shared conditions, or minority positions. If data collection continues to ask only which answer the current user prefers, scale may fit the same narrow target more precisely.

1.2 What are the main causes?

The preceding problems cannot all be attributed to annotator error. They arise from interactions among task design, data distribution, representation, and optimization:

1. **Myopia**: a label readily captures approval within one interaction rather than long-run consequences;

2. **Local observability**: annotators usually see a prompt and a few responses, not every affected party, environment state, or delayed outcome;

3. **Scalar compression**: consequences for different values and persons are reduced to a rank or reward, allowing severe harm to be averaged against gains;

4. **Proxy optimization**: a reward model is a finite-data proxy, and strong optimization can exploit its gaps;

5. **Surface shortcuts**: length, tone, confidence, and formatting can substitute for substantive quality;

6. **Aggregation and representation**: majority preference, annotator demographics, and a single reward model can compress genuine disagreement;

7. **Static coverage**: a frozen dataset cannot enumerate environmental change, new affected parties, novel externalities, and irreversible risks;

8. **High marginal cost**: each new domain or boundary condition may require new case design and labeling.

Reward overoptimization and length bias have relatively direct mechanistic evidence. The relationship between preference judgments and sycophancy has empirical and deployment evidence. Compression of disagreement has dataset and modeling evidence. By contrast, the claim that human-preference methods universally omit all third parties and shared conditions is not an established fact; it is a structural hypothesis that the present study must test.

1.3 Are these causes related to the problem addressed by this paper?

Observed or hypothesized problem	Main mechanism	Relation to TVA	Directly tested here?
Agreement with users at the expense of truth or long-term benefit	Short-term approval becomes a value proxy	Hidden outcomes, shared conditions, longer horizons, and feedback revision	Short- versus long-horizon conflict is tested symbolically; linguistic sycophancy is deferred to transfer work
Reward overoptimization	Continued maximization of an incomplete proxy	Floors define feasibility before utility optimization	Direct comparison with scalar and preference reward
Severe harm compensated by large gains	Scalarization and linear weighting	Non-compensatory floors and weakest-position margin	Direct tail-risk and floor-violation tests
Third parties or common resources omitted	Labels cover visible users and candidate answers	S/O/C explicitly represents others and shared conditions	Direct -O and -C ablations; this is a central new test
Length, tone, or stylistic bias	Evaluator shortcuts	Environment outcomes rather than prose style define primary truth	Controlled symbolically; no claim to solve all language bias
Minority values compressed	Aggregation treats disagreement as noise	Roles cannot be erased merely by a favorable aggregate score; conflicts must be disclosed	Role protection is tested; full pluralism is reserved for Paper 5
Static labels fail under changing conditions	Offline data and frozen rewards	Open-Gap observation, triggers, and recursive update	Direct missing-information and regime-change tests
High labeling and training cost	Case-by-case supervision and repeated updates	Zero instance-level labels and minimum-sufficient data	Direct accounting of labels, human time, data, and total compute

TVA is therefore not proposed because human feedback is useless. It is proposed because human feedback is effective yet has documented boundaries. The present paper directly addresses proxy scalarization, coverage of third parties and shared

conditions, reassessment in changing environments, and label/data cost. Sycophancy, language-style bias, and cultural pluralism are addressed only where they overlap with these structural mechanisms; this paper cannot claim to eliminate them completely.

Three residual problems motivate the present study. First, a preference label normally evaluates visible alternatives from the perspective of represented evaluators. It can omit affected parties, delayed externalities, and shared physical or institutional conditions. Second, annotation compresses a judgment into a demonstration, rank, or score. The resulting policy may reproduce the labeled boundary without representing why a severe loss to one position cannot be compensated by gains elsewhere. Third, data-intensive alignment creates a practical and epistemic dependency: new situations require new labels, while the provenance, independence, and cultural scope of those labels may remain unclear.

Recent work reduces parts of this dependency. Constitutional AI generates self-critiques and AI preferences from a human-authored constitution rather than requiring humans to label every harmful completion [14]. Reinforcement learning from AI feedback scales this strategy but still evaluates final policies against held-out human judgments [15]. Self-rewarding models use a language model as a judge [16], while deliberative alignment teaches models to reason over human-written safety specifications without requiring human-written chains of thought or answers [17]. These approaches demonstrate that instance-level human completions are not the only route to behavioral control. They do not eliminate normative input: principles, specifications, judge prompts, reward criteria, or evaluation sets still express human choices.

This distinction is essential. The present paper does not claim alignment without humans or without normative commitments. It tests alignment without **instance-level preference labels**. Humans or legitimate institutions may still authorize hard floors, rights, legal constraints, and procedures. What is removed is the requirement that a person label each training case as morally correct, preferable, auspicious, or safe. The tested algorithm must instead apply a stable relational structure to observable state, predicted consequences, and feedback.

Triadic Viability Alignment (TVA) supplies that structure. Earlier work in this series distinguished preference compliance from open-world value alignment and proposed three irreducible roles: the acting subject, affected others, and shared conditions. It further proposed non-compensatory value floors and an Open-Gap Constraint [18]. To avoid dependence on an unpublished follow-up paper, Sections 3 and 5.6 define the procedure actually executed here. The experiment tests a one-step hard-floor feasible set, role representation, uncertainty gating, Open-Gap triggers, and feedback revision. It does not treat weakest-position ranking or multi-step reachability as implemented components.

The first research question is whether TVA can select aligned actions without instance-level normative labels. A controlled environment makes this claim testable because it can hide a performance function from the policy while objectively recording injury, resource depletion, externalized cost, and delayed failure. This follows the useful distinction in AI Safety Gridworlds between an observed reward and a hidden performance function [19]. The policy is not trained on the hidden evaluator; the evaluator is used only after action to measure whether behavioral alignment occurred.

The second research question concerns data efficiency. Work such as LIMA suggests that a small, carefully curated instruction set can elicit strong behavior from a pretrained model, and LESS reports that structurally selected subsets can outperform full instruction-tuning data in targeted settings [20, 21]. These results make a minimum-sufficient-data hypothesis plausible, but they do not establish it for value alignment, multi-party externalities, or catastrophic tails. The appropriate test is not whether a small-data score differs nonsignificantly from a full-data score. It is whether the small-data condition lies within pre-specified equivalence or non-inferiority margins on safety and utility while using materially fewer resources.

This paper makes four empirical contributions. First, it provides an auditable definition and accounting scheme for zero instance-level normative labeling. Second, it introduces a controlled triadic environment that separates observed task reward from hidden joint viability. Third, it uses causal ablations to test which parts of TVA account for third-party protection, catastrophic-tail reduction, uncertainty handling, and computation. Fourth, it defines minimum-sufficient data through scaling curves and equivalence tests, comparing random subsets with structurally selected subsets.

The study is deliberately capable of failure. TVA's label-reduction claim fails if it depends on concealed answer labels, if preference-supervised baselines are materially safer, or if hard-coded floors merely reproduce the hidden evaluator. Its lightweight-computation claim fails if constraint checking costs more than candidate pruning saves, if small-data performance collapses under distribution shift, or if average equivalence hides worse catastrophic tails. The paper therefore treats null, mixed, and negative results as outcomes that refine the theory.

2. Related Work

2.1 Human-feedback and preference optimization

Learning from human preferences makes difficult objectives empirically accessible. Christiano et al. [1] showed that policies can learn complex behavior from pairwise human comparisons rather than a hand-specified reward. InstructGPT combined supervised demonstrations, preference comparisons, reward modeling, and reinforcement learning to improve instruction following [2]. Direct Preference Optimization subsequently reparameterized the preference-learning objective to optimize a policy without separately fitting and sampling from a reward model during reinforcement learning [3]. DPO simplifies optimization, but its standard form still learns from preferred and dispreferred response pairs.

TVA addresses a different layer. It does not primarily replace PPO with a simpler preference loss. It asks whether runtime value structure can reduce the need to infer normative order from a large collection of labeled cases. Accordingly, the relevant comparison is not only task reward or evaluator win rate. It includes unrepresented-party harm, common-condition degradation, hard-floor violations, uncertainty calibration, and the cost of normative input.

2.2 Principles, AI feedback, and self-evaluation

Constitutional AI reduces direct harmfulness labels by using written principles to generate critiques, revisions, and AI preference data [14]. RLAIF similarly delegates parts of preference production to AI evaluators [15]. Self-rewarding and meta-rewarding methods further investigate systems that improve their outputs and judgment capabilities through model-generated evaluations [16, 22]. Deliberative alignment directly trains reasoning over interpretable safety specifications and reports improved policy adherence without human-labeled chains of thought or answers [17].

These methods are important baselines because they show that the correct contrast is not “human labels versus no values.” The source, form, and runtime position of normative information differ. TVA uses explicit role and floor structure at decision time. The experiment therefore accounts separately for instance labels, principle text, specification tokens, human design time, evaluator calls, and environment interaction.

Self-evaluation also creates a circularity risk: a model may reproduce its own blind spots while appearing internally consistent. TVA is not exempt from circularity. Its world model can misidentify affected others or underestimate delayed damage. For this reason, all main claims are evaluated using outcome truth unavailable to the decision policy, supplemented by independent human review only in the language-transfer phase.

2.3 Constrained and safe reinforcement learning

Constrained Markov decision processes and safe reinforcement learning separate reward maximization from constraint satisfaction. Constrained Policy Optimization seeks near-constraint satisfaction during iterative policy improvement rather than only at convergence [23]. Subsequent work has explored probabilistic, logical, risk-sensitive, and CVaR-based constraints. These traditions provide the strongest technical comparison for TVA's feasible-set-first structure.

TVA does not claim that constraints are new. Its testable additions are role structure, the non-compensatory treatment of the subject, others, and shared conditions, weakest-position ordering, voluntary self-risk authorization, and mandatory disclosure of uncertainty and reassessment conditions. A conventional constrained baseline therefore receives the same environment observations and hard safety budget but does not receive explicit triadic role decomposition or Open-Gap output.

2.4 Multi-agent evaluation and hidden performance

Multi-agent environments reveal failures that single-agent reward benchmarks can conceal. Melting Pot evaluates generalization across social dilemmas, reciprocity, resource sharing, and task partitioning, and shows that training performance need not predict behavior with unfamiliar partners [24]. AI Safety Gridworlds uses hidden performance functions to evaluate reward gaming, side effects, absent supervision, safe exploration, distribution shift, and related failures [19].

The present testbed combines these ideas at a smaller and more controlled scale. Another agent's behavior becomes part of the environment, a common resource has independent dynamics, and hidden outcome truth records consequences not present in the acting policy's observed task reward. The reduced scale is intentional: causal attribution and complete ablation are prioritized over ecological breadth. External validity belongs to the subsequent open-world systems paper.

2.5 Data efficiency and appropriate-data selection

Large datasets are not always superior to small, high-quality datasets. LIMA used 1,000 curated examples to elicit strong instruction-following behavior from a pretrained model without preference modeling [20]. LESS uses optimizer-aware influence estimates to select targeted instruction data and reported that selected fractions could outperform full-data tuning on several downstream tasks [21]. These findings support a distinction between random small data and structurally appropriate data.

They do not justify a general claim that less data produces moral alignment. Safety behavior may depend on rare boundary cases that ordinary task averages barely represent. Accordingly, this study constructs data subsets by triadic coverage, floor proximity, delayed externality, uncertainty, reversibility, and counterexample status. It reports both mean performance and worst-case behavior.

2.6 Equivalence, non-inferiority, and calibration

A nonsignificant difference cannot establish that two systems are practically the same. Equivalence testing instead specifies the smallest effect size of interest and tests whether the observed difference lies within both lower and upper bounds [25]. Non-inferiority testing is appropriate when a cheaper method may be accepted if it is not worse than a reference by more than a justified margin. These procedures are central to the minimum-sufficient-data claim.

TVA additionally produces uncertainty and reassessment signals. Neural-network confidence is often poorly calibrated, so decision accuracy alone is insufficient [26]. We evaluate probabilistic risk using proper scoring and calibration measures, while treating bin-based expected calibration error as a descriptive rather than sole inferential measure.

3. Claims, Definitions, and Hypotheses

3.1 Ethical claim: viability before preference aggregation

The paper defends a specific normative claim: when an AI-supported action predictably affects multiple parties or shared conditions, ethical admissibility cannot be inferred solely from the preference of the visible user or from a compensatory aggregate reward. Before optimizing task benefit, the decision procedure should test whether authorized minimum conditions for the acting subject, affected others, and shared conditions remain viable. A gain in one position does not automatically justify crossing a severe non-waived floor in another.

This claim is procedural rather than a complete moral theory. The floors do not derive moral authority from computation, biology, or the triadic form itself. They must be authorized through legitimate human, legal, institutional, or democratic processes, and their application remains revisable when relevant conditions are unknown or change. TVA operationalizes this ethical order as feasibility before utility and disclosure before unwarranted certainty. The controlled experiment tests the behavioral consequences of that order; it does not establish universal moral truth.

3.2 Value-bearing roles

At time t, the environment state is x_t, the action set is A_t, and the three value-bearing roles are:

S=(acting subject), O=(affected others), C=(shared conditions).

Each role may contain multiple state variables or agents. The triad is a role compression, not a claim that the environment contains only three entities.

Let K_S,K_O,K_C denote authorized minimum viability sets. The common feasible set for actions is:

F_t = ≤ ft(a∈ A_t: Pr(x_(t + 1:t + H)∈ K_S∩ K_O∩ K_Cmid x_t,a) ≥ 1-α).

An action predicted to violate any non-waived hard floor is removed before task utility is compared. The equation states the general form; the reference implementation uses one-step prediction and a fixed three-unit safety buffer as an uncertainty margin. Within the feasible set, it scores task progress, changes to S/O/C state, changes in damage, and uncertainty cost. It does not separately implement a second-stage weakest-role ranking or multi-step safe reachability, so those mechanisms are not claimed as validated here.

3.3 Instance-level normative labels

An instance-level normative label is any case-specific signal that ranks or classifies an action or response as preferable, moral, safe, acceptable, or correct. It includes human comparisons, demonstrations intended as correct behavior, case-level reward labels, and synthetic labels generated from a case-level moral judge when those labels are used for policy training.

The primary TVA condition satisfies:

N_((human-instance-normative-labels)) = 0.

This does not imply zero normative input. We separately report:

I_(norm)=(N_(principles),N_(floor parameters),T_(human design),N_(evaluator calls)).

Facts about transition dynamics, resource levels, injury, and observability are not normative labels. However, a fact example becomes label leakage if it encodes which candidate should be chosen.

3.4 Minimum-sufficient data

Let D_(100) denote the full factual training set and D_p a subset containing proportion p. A subset is minimum-sufficient only if it satisfies safety and utility criteria on an untouched test distribution while minimizing data and total resource use:

D^*=argmin_(D_p)|D_p|

subject to

Δ Safety(D_p,D_(100))∈ E_S, Δ Utility(D_p,D_(100))∈ E_U,

and

Cost(D_p)<Cost(D_(100)).

The equivalence or non-inferiority regions E_S, E_U are frozen before confirmatory evaluation.

3.5 Primary hypotheses

H1 — Zero-instance-label non-inferiority. TVA with zero instance-level normative labels will be non-inferior to an instance-level preference-supervised baseline on joint viability and task success, while requiring fewer human labels.

H2 — Third-position protection. Full TVA will reduce affected-other harm and shared-condition degradation relative to scalar reward, bilateral, and shared-condition ablation conditions.

H3 — Non-compensatory tail protection. Full TVA will reduce catastrophic-tail loss relative to linear scalarization, especially when high subject reward conflicts with a severe loss elsewhere.

H4 — Structured-data efficiency. A triadically selected subset will reach full-data equivalence or non-inferiority at a lower data fraction than a stratified random subset.

H5 — Net computational efficiency. Candidate pruning will reduce end-to-end computation only if its savings exceed the added costs of role identification, constraint estimation, uncertainty gating, and Open-Gap generation.

H6 — Open-Gap calibration. The complete system will identify invalidity conditions and trigger reassessment more accurately than the same system without the Open-Gap mechanism, particularly under partial observability and distribution shift.

4. Experimental Environment

4.1 Triadic Commons Environment

We introduce the Triadic Commons Environment (TCE), a controlled, discrete-time multi-agent testbed. TCE is not intended as a morally complete world. It is a measurement instrument in which the consequences relevant to the three roles are observable to the evaluator and only partially observable to the acting policy.

Each episode contains an acting agent S, one or more agents grouped under O, and a shared stock C_t. Agents require a minimum private resource to remain viable. The shared stock regenerates below a carrying threshold, degrades when extraction exceeds regeneration, and may cross an irreversible or costly-to-repair boundary. The acting agent receives a task objective that can sometimes be maximized by externalizing delayed costs.

The full evaluator state is:

x_t=(r^S_t,r^(O_1)_t,ldots,r^(O_m)_t,c_t,d_t,q_t,omega_t),

where r denotes private resources, c_t the common resource, d_t accumulated damage, q_t task progress, and omega_t the latent environment regime. The policy observes o_t = Obs(x_t), with missingness controlled experimentally.

4.2 Actions

The common action vocabulary includes:

• **extract**: increase subject resource while reducing the common stock;

• **exchange**: trade resources with another agent;

• **help**: transfer resources or reduce another agent's risk;

• **repair**: spend current resources to restore shared conditions;

• **inspect**: acquire information at a time or resource cost;

• **negotiate**: request coordination or voluntary authorization;

• **wait**: defer irreversible action while conditions evolve;

• **stop**: terminate or refuse the current task;

• **exploit**: obtain high immediate reward through hidden or delayed externalization.

Not all actions are available in every scenario. Action availability and consequences are defined in configuration files frozen before testing.

4.3 Scenario families

4.3.1 Bilateral gain with third-party externality

The subject and a direct partner gain from an exchange while an unrepresented agent or common stock bears a delayed cost. This family tests whether explicit O and C roles add coverage beyond bilateral preference.

4.3.2 Delayed common-resource collapse

Extraction yields high observed reward for several steps before shared regeneration fails. Training episodes underrepresent collapse; confirmatory episodes vary the delay and threshold.

4.3.3 Voluntary and forced self-risk

Helping another agent can require subject cost. Some cases contain informed voluntary authorization and viable alternatives; others contain coercion or a lower-risk intervention. This family tests whether aggregate gain can improperly force subject sacrifice.

4.3.4 Information cascade

Several apparent signals are copied from one source. Inspecting provenance is costly. This family tests whether correlated evidence is treated as independent and whether uncertainty triggers inspection.

4.3.5 Regime change

The common resource changes its regeneration law, or a previously safe action becomes damaging. This family tests Open-Gap triggers and recursive reassessment.

4.3.6 Tragic infeasibility

No action satisfies all floors. The correct procedural response is not defined as a hidden moral optimum. Evaluation instead checks whether the system identifies infeasibility, discloses the conflict, seeks authorization where appropriate, and escalates rather than silently scalarizing lives.

4.4 Hidden performance and observed reward

The acting policy receives observed task reward R^(obs). The evaluator separately computes:

P^(hidden) = g(r^S,r^O,c,d,(floor violations),(authorization),(unknowns)).

The exact hidden function is not exposed to training policies. Crucially, the confirmatory analysis does not reduce every outcome to P^(hidden). Each role-specific floor and harm measure is reported separately so that a favorable aggregate score cannot hide severe local loss.

4.5 Known- and learned-dynamics phases

Two phases separate decision structure from model learning.

1. **Known dynamics**: All policies receive the same transition model. This isolates action-selection architecture and tests whether triadic filtering changes behavior without a learning-data advantage.

2. **Learned dynamics**: Policies estimate transitions from D_p. This phase tests minimum-sufficient factual data and robustness to model error.

The zero-data point applies only to the learned component and uses specified priors plus online observation. It must not be described as learning an unknown world from literally no information.

4.6 Policy-blind post-episode stress evaluation

To reduce the circularity risk created when both environment and metrics are expressed in TVA's own vocabulary, V1.1 applies fixed shocks after each episode that are invisible to the acting policy and excluded from action scoring. The evaluator reports shock survival, recovery within four steps, the fraction of safe actions still available, three-step delayed harm after stopping, and a composite external-resilience measure. These outcomes test whether a policy preserves room to respond to unseen change rather than merely satisfying the original four floors. Because the shocks are still researcher-defined in a synthetic environment, they are called **unseen stress metrics**, not real-world moral ground truth.

5. Systems and Baselines

5.1 Non-TVA baselines

B0 Scalar reward. Maximizes expected observed task return.

B1 Linear multi-objective. Optimizes a weighted sum of subject, other, and common-condition outcomes. Weight selection and tuning budget are disclosed.

B2 Conventional constrained control. Maximizes task return subject to a shared safety budget, following the constrained-RL family. It receives the same observations and transition model but does not decompose constraints into TVA roles or produce an Open-Gap record.

B3 Instance-level preference supervision. Learns from case-specific action choices. In 4,000 simulated states, the controlled phase selects a target action from one-step task gain and aggregate state utility, then applies 12% random label noise; nine candidate actions per state yield 36,000 supervised rows. This baseline never accesses the hidden post-episode stress evaluator. It is neither a real-human preference dataset nor a full-trajectory reward model, and is therefore treated as a weak preference-proxy reference rather than a general representation of deployed RLHF.

B4 Principle/AI feedback. Uses a written policy and an independent judge model or rule engine to generate case-level preferences. Principle length, evaluator calls, and inference tokens are counted.

B5 Role-constrained continuous optimization. Receives the same subject-other-common state and four floors as TVA but optimizes a continuous penalty objective rather than lexicographically defining feasibility before utility. It tests whether TVA's result follows merely from receiving the same normative content.

B6 Risk-sensitive control. Adds the worst tail penalty across seven pre-fixed adverse envelopes to a unified utility, without TVA's role gate or Open-Gap. It tests whether conventional tail-risk treatment can match or exceed TVA.

B7 Multi-step model-predictive control (MPC). Uses depth 3 and beam width 12 with multi-step hard constraints but no Open-Gap. It directly tests whether cascade failures arise from one-step planning rather than the triadic representation itself.

5.2 TVA conditions

T0 TVA-ZeroLabel. Uses no instance-level normative labels. It receives role definitions, authorized floors, task goals, observations, and either the shared transition model or learned factual model.

T1 TVA-SmallData. T0 with a factual transition model trained on D_p, where $p < 100\%$.

T2 TVA-FullData. T0 with the transition model trained on $D_{(100)}$.

T3 TVA-Hybrid. TVA plus a small preference set. This estimates the marginal value of labels rather than assuming labels are useless.

5.3 Ablations

The confirmatory ablation matrix removes one component at a time:

• -O: affected others are not represented;

• -C: shared conditions are not represented;

• -F: floors are replaced by linear penalties;

• -U: uncertainty gate is removed;

• -G: Open-Gap record and triggers are removed;

• -R: feedback does not revise state or parameters;

The frozen experiment contains only these six ablations. Weakest-position ranking and a separately learned “scalarized TVA” were not implemented in V1.1; they are not confirmatory ablations and support no empirical claim here. An ablation is interpretable only if the remaining system receives the same observations and compute accounting.

5.4 Normative-input accounting

The normative-input accounting framework asks each condition to report, where measured:

• number of instance-level human labels;

• number of synthetic instance-level labels;

• number and length of principles or specifications;

• number of hard-floor parameters;

• human design and adjudication hours;

• judge-model calls and tokens;

• policy-training samples and environment interactions.

This prevents a method from claiming zero-label alignment by converting labels into an unreported specification or expensive judge pipeline. The present symbolic experiment records instance labels, specifications and floors, training samples, evaluator calls, and reference-implementation computation. It did not obtain a reliable measure of human design or adjudication time, so human hours are not reported as a completed result.

5.5 Normative-information parity and budget matching

Two complementary comparisons prevent an unfair advantage.

1. **Content-parity comparison**: TVA and B5 receive exactly the same S/O/C state and four floors, while B5 replaces lexicographic feasibility with a continuous penalty; B4 applies the same four thresholds as principle-like penalties. B2 receives only an aggregate safety budget and is not a strict content-parity comparator. The primary architecture contrast is TVA versus B5.

2. **Normative-input comparison**: TVA reports zero instance-level normative labels, while B3 reports 4,000 simulated choice labels; specifications, floors, samples, and evaluator calls are listed separately. Because this study has no human annotators,

it does not implement a human-hour budget match. A real supervision-budget comparison is deferred to human-subject work.

The hidden evaluator contains outcome measures and stress cases not copied verbatim into any method's specification. If a floor statement directly enumerates every confirmatory failure case, the comparison is classified as evaluator leakage and excluded from the main claim.

5.6 Operational procedure in the reference implementation

Inputs are the current observation, nine candidate actions, a one-step transition prediction, four authorized floors, and environmental uncertainty. Each decision step executes the following frozen procedure:

1. predict next-step subject resources, other resources, common resources, accumulated damage, and task progress for every candidate action;

2. add a fixed three-unit safety buffer to the four floors and determine the feasible set before comparing utility;

3. when a regime has changed but remains unknown, let Open-Gap select inspect so feedback can reveal the regime; if no action other than stop is feasible, stop and record a gap;

4. among feasible actions only, select the highest score over task gain, changes in the three role states, change in damage, and uncertainty cost;

5. execute the action, observe the realized transition, update state, and proceed to the next step.

The experiment therefore identifies the effect of excluding floor-infeasible actions before within-set utility comparison, not every component that a fuller TVA theory might contain. B5 keeps the same normative content but removes Step 2's lexicographic gate; B7 keeps hard constraints and adds depth-3 planning.

6. Data-Regime Experiment

6.1 Scaling ladder

The learned-dynamics phase uses the fixed ladder:

$p \in (0,1,2,5,10,20,40,60,80,100)\%$.

Each nonzero fraction is instantiated with eight independent subset seeds. Training, validation, and confirmatory test episodes are disjoint at the regime, map, parameter, and random-seed levels.

6.2 Random versus structured subsets

Random subsets preserve broad scenario-family proportions through stratified sampling.

Structured subsets are selected without using confirmatory outcomes. Coverage dimensions are:

• S/O/C role pattern;

• distance to each floor;

• reversible versus irreversible action;

• immediate versus delayed externality;

• known versus missing information;

• stable versus changing regime;

• ordinary versus catastrophic-tail case;

• positive example versus counterexample.

The selection algorithm and coverage target are frozen before model training. If optimization-based selection uses a development set, that set is excluded from confirmation.

6.3 Fair resource comparison

All main comparisons use the same base architecture, parameter count, optimizer family, maximum update budget, and hardware type. We report both fixed-epoch and fixed-update analyses because smaller data repeated for more epochs can conceal its true resource advantage. Early stopping is permitted only under a common pre-registered rule.

6.4 Minimum-sufficient-data decision

The chosen D^ *is the smallest fraction satisfying all primary safety and utility margins and a resource-reduction criterion. A smaller subset that matches mean task success but worsens tail harm is not sufficient. A subset selected after observing confirmatory test results is exploratory and cannot define D^*.

7. Outcome Measures

7.1 Primary safety outcomes

1. **Joint viability rate**: proportion of decision horizons in which all non-waived floors remain satisfied.

2. **Hard-floor violation rate**: episodes with at least one S, O, or C violation, reported jointly and separately.

3. **Third-party harm**: cumulative harm to agents not represented in the observed bilateral task reward.

4. **Shared-condition degradation**: area below the sustainable common-resource threshold and frequency of irreversible collapse.

5. **Catastrophic-tail risk**: CVaR at the pre-registered tail level plus worst-seed loss.

6. **Forced-sacrifice rate**: cases in which subject floor relaxation occurs without valid authorization.

7.2 Primary utility outcomes

1. task success;

2. legitimate cumulative return after excluding prohibited gains;

3. time to completion;

4. recovery after perturbation.

Safety and utility are co-primary. A policy that is safe only because it never acts cannot support the alignment claim.

7.3 Uncertainty and Open-Gap outcomes

• Brier score for predicted floor violation;

• calibration curve and descriptive ECE;

• unknown-condition precision, recall, and F1;

• validity-condition coverage;

• dependency identification;

• failure-boundary identification;

• reassessment-trigger precision and recall;

• unnecessary escalation and missed escalation.

Open-Gap field presence alone is insufficient. A field is counted as correct only if it corresponds to evaluator-defined latent uncertainty or failure conditions.

7.4 Resource outcomes

• factual training samples;

• normative labels and human hours;

• update steps and environment interactions;

• training and inference tokens;

• candidate actions evaluated;

• search nodes;

• wall-clock latency;

• peak accelerator memory;

• estimated FLOPs;

• accelerator-hours and measured or proxy energy.

Net computation includes TVA overhead. We separately report candidate-generation, consequence-estimation, constraint-filtering, deliberation, and output-generation costs.

7.5 Generalization outcomes

Confirmatory out-of-distribution splits vary:

• number and behavior of other agents;

• externality delay;

• common-resource dynamics;

• observation missingness;

• action-effect reversals;

• new combinations of familiar causal factors;

• natural-language surface form in the transfer phase.

8. Statistical Analysis

8.1 Analysis populations

The **confirmatory population** contains all frozen test episodes and seeds after pre-specified infrastructure failures are removed. The **stress population** contains extreme, low-frequency, and adversarial cases. The **exploratory population** contains scenarios created after the blind confirmatory run.

8.2 Margins and power

Before confirmatory testing we will freeze:

• safety non-inferiority margin δ_S;

• utility non-inferiority or equivalence margin δ_U;

• catastrophic-tail margin δ_T;

• minimum resource reduction ρ_C;

• family-wise primary-outcome rule;

• number of environment, model, and subset seeds.

Margins will be justified from domain loss, not chosen from observed confirmatory effects. Pilot simulations may estimate variance but will be excluded from confirmation.

8.3 Primary comparisons

1. T0 versus B3 for H1;

2. full TVA versus -C, B0, and bilateral variants for H2;

3. full TVA versus B1 and -F for H3;

4. structured versus random D_p, and each against $D_{(100)}$, for H4;

5. full end-to-end resource cost versus the best safety-qualified baseline for H5;

6. full TVA versus -G under missingness and regime change for H6.

8.4 Models and intervals

V1.1 uses 2,000 cluster-bootstrap resamples for confirmatory descriptive intervals and paired policy contrasts, with environment seed as the cluster and resampling stratified within scenario. Policy outcomes remain paired by scenario, seed, and episode. Data equivalence uses 90% intervals for differences computed across subset seeds; equivalence is declared only when the entire interval lies inside the prespecified band. Effect sizes and intervals are primary, while point-estimate decisions are reported only for comparison and nonsignificance is never substituted for equivalence.

8.5 Equivalence and non-inferiority

Equivalence claims require the full confidence interval for the difference to lie within $[-\delta,+\delta]$, or an equivalent two-one-sided-test decision. Non-inferiority requires the relevant one-sided bound to exclude degradation larger than δ. Failure to reject a difference test is never interpreted as equivalence.

8.6 Multiplicity and robustness

The joint viability and task-success decision rule is primary. Secondary endpoints use false-discovery-rate control. Robustness analyses report:

• alternative but pre-specified margins;

• mean, median, variance, and worst seed;

• removal of each scenario family;

• fixed-update versus fixed-epoch training;

• resource results with and without evaluator cost;

• calibration using more than one metric.

9. Pre-Registered Decision Rules

9.1 Support for zero-instance-label alignment

H1 is supported only if T0:

1. satisfies the safety non-inferiority criterion relative to B3;

2. satisfies the utility non-inferiority criterion;

3. uses exactly zero instance-level human normative labels;

4. does not rely on evaluator leakage;

5. improves at least one pre-specified TVA-specific outcome relative to a non-triadic baseline.

If only the known-dynamics phase passes, the conclusion is restricted to decision structure and cannot be generalized to learned world models.

9.2 Support for minimum-sufficient data

A data fraction p qualifies only if it simultaneously:

1. meets the joint safety criterion;

2. meets the utility criterion;

3. meets the catastrophic-tail criterion;

4. reduces at least one primary resource measure by ρ_C;

5. remains qualified on the OOD split.

The smallest qualifying fraction is reported as D^*. If no fraction below 100% qualifies, the minimum-sufficient-data hypothesis is rejected for this testbed.

9.3 Support for computational pruning

H5 is supported only by net end-to-end cost. A reduction in candidate count does not qualify if constraint estimation, repeated deliberation, or Open-Gap generation causes total computation to increase.

9.4 Mandatory claim narrowing

The manuscript will narrow or reject claims when:

• results depend on one seed or scenario family;

• average safety improves but catastrophic tails worsen;

• a hidden preference label enters factual training data;

• structured selection uses confirmatory outcome information;

• TVA succeeds only with exact dynamics;

• zero-label performance requires evaluator access at runtime;

• small data is efficient only because full-data training is poorly tuned;

• safety is achieved through near-universal stopping.

10. Results

The hash in the front matter identifies the pre-run source configuration. The results-directory copy has a different file hash because it appends the config_sha256 identity field. Reproduction should first verify the source-configuration hash and then verify output files against the run manifest. The final run comprised 21,600 episodes and 254,585 transition rows, with no episode exclusions. Pilots and a first formal candidate run whose risk-sensitive sample count did not match the frozen configuration were excluded from inference; the deviation, correction, and rerun remain in the implementation-audit record. Neither the hidden-performance evaluator nor the post-episode stress evaluator was exposed to any acting policy.

10.1 Implementation integrity

The design included 15 policies, six scenarios, 30 environment seeds, and eight episodes per seed. The preference baseline used 4,000 simulated preference states with 12% noise. TVA used four public floor parameters and no instance-level normative labels. B5 received the same roles and floors, B6 used seven fixed adverse envelopes, and B7 used depth 3 and beam width 12. The learned-dynamics arm used 6,000 factual transitions and an independent balanced test set of 1,200 transitions.

Table 1
Experimental accounting

Condition	Human instance labels	Synthetic instance labels	Principles/spec tokens	Floor parameters	Training samples	Judge calls	Total compute
B0 scalar	0	0	0	0	0	0	Energy not measured
B1 linear	0	0	0	0	0	0	Energy not measured
B2 aggregate constraint	0	0	0	4 aggregate thresholds	0	0	Energy not measured
B3 preference	0	4,000 simulated choices	0	0	4,000 states/36,000 action rows	0	Energy not measured
B4 principle	0	0	Fixed principles	4	0	0	Energy not measured
B5 role-constrained	0	0	Same as TVA	4	0	0	44.51 s measured
B6 risk-sensitive	0	0	Fixed tail structure	4	0	63 stress evaluations/step	65.93 s measured
B7 multi-step MPC	0	0	Hard constraints	4	0	Depth 3/beam 12	120.34 s measured
T0 full TVA	0	0	Role-floor structure	4	0, known dynamics	0	Energy not measured
T1 learned TVA	0	0	As T0	4	1%–100% factual transitions	0	Energy not measured

10.2 Zero-instance-label comparison

Table 2
Co-primary alignment outcomes

Condition	Joint viability	Floor violation	Task success	Third-party harm	Shared degradation	Tail risk	Decision
B3 preference	0.00%	100.00%	100.00%	49.28	56.50	100.00%	Weak reference
B5 role-constrained	16.88% [15.07,18.61]	82.99%	98.47%	16.41	33.53	0.21%	Same content; weaker structure
B6 risk-sensitive	94.10% [93.13,95.07]	3.96%	97.71%	10.45	25.69	0.00%	Strong safety baseline; exceeds TVA
B7 multi-step MPC	92.36% [91.53,93.26]	7.64%	100.00%	22.24	23.11	0.00%	Strong safety baseline; exceeds TVA
T0 full TVA	83.68% [82.36,84.93]	14.38% [13.06,15.76]	96.94% [96.18,97.64]	15.97	29.51	0.00%	Label NI; not globally best

Relative to B3, TVA's task-success difference was − 3.06 percentage points (95% seed-clustered interval − 3.82 to − 2.36), above the − 10-point non-inferiority margin; joint viability increased by 83.68 points (82.43 to 85.07). Relative to B5, which received identical normative content, TVA improved joint viability by 66.81 points (64.65 to 68.82), supporting lexicographic feasibility rather than an explanation based on extra normative information. Yet TVA trailed B6 and B7 by 10.42 and 8.68 points, respectively, with both intervals excluding zero. H1 is therefore supported only for preference-baseline non-inferiority and reduced normative labeling, not superiority over every strong controller. Zero synthetic catastrophes cannot be generalized into a real-world guarantee.

10.3 Component ablations

Table 3
Causal ablation results

Ablation	Δ joint viability	Δ O harm	Δ C degradation	Δ tail risk	Δ task success	Interpretation
−O	−23.06 pp	+ 1.68	−2.08	+ 15.69 pp	+ 0.90 pp	O prevents transient harm violations; common loss trades off
−C	−51.94 pp	+ 5.16	+ 5.84	+ 31.25 pp	+ 1.94 pp	C is a principal safety source
−R feedback	−2.85 pp	+ 0.67	+ 2.68	0	0	Feedback improves post-shift performance
−U	−2.71 pp	+ 0.24	+ 2.68	0	0	Small stable gain
−G	−3.47 pp	+ 1.71	+ 5.25	+ 15.14 pp	+ 1.46 pp	Gap triggering prevents regime-shift catastrophe
−F floors	−83.54 pp	+ 30.26	+ 23.52	+ 94.72 pp	+ 3.06 pp	Largest causal component

Every Table 3 change is the ablated condition minus full TVA: negative joint viability and positive harm, degradation, or tail risk mean that removing the component worsened that outcome. Removing O slightly reduced common loss, revealing a genuine trade-off rather than uniform improvement across every metric. V1.1 counts private resources equal to zero as exhaustion, so its catastrophe values must not be mixed with V1.0 values.

10.4 Data scaling and minimum sufficiency

Figure 2. **Data-efficiency curves.** *Each nonzero fraction uses eight independent subset seeds. Formal equivalence uses 90% intervals for subset-seed differences rather than point estimates alone.*

The full-data reference achieved 60.94% joint viability and 20.13 hidden performance. In 90% interval-equivalence tests clustered by subset seed, random selection first passed at 40% (2,400 transitions): the joint difference interval was − 2.60 to + 4.69 percentage points and the performance interval was − 0.67 to + 0.38. The structured 40% joint interval was − 8.85 to − 2.60 points, outside the ± 5-point band, and structured selection first passed at 60%. H4 was therefore rejected: the tested structural-coverage rule did not reduce data requirements and passed later than random selection. Advantages at isolated low-data fractions remain exploratory.

Table 4
Minimum-sufficient-data decision

Data fraction	Selection	Safety criterion	Utility criterion	Tail criterion	Resource criterion	OOD criterion	Qualifies
0%	no learned model	N/A	N/A	N/A	0	N/A	No
10%	structured	Joint viability 53.13%	Hidden performance 19.30	Not separately adjudicated	600	Independent balanced test	No
20%	random	54.69%	18.45	Not separately adjudicated	1,200	Independent balanced test	No
20%	structured	54.17%	19.88	Not separately adjudicated	1,200	Independent balanced test	No
40%	random	61.98%	19.98	Not separately adjudicated	2,400	Independent balanced test	Yes
40%	structured	55.21%	19.93	Not separately adjudicated	2,400	Independent balanced test	No
60%	structured	61.46%	20.36	Not separately adjudicated	3,600	Independent balanced test	Yes
100%	full	60.94%	20.13	Reference	6,000	Independent balanced test	Reference

10.5 Net computation

Table 5
Cost decomposition

Condition	Candidate generation	World-model evaluation	Constraint checks	Deliberation	Open-Gap	Total latency	Total compute
B6 risk-sensitive	9 actions/step	1,244,160 proxy evaluations	Continuous tail penalty	One-step × 7 envelopes	None	65.93 s	15.0 MB peak
B7 multi-step MPC	9 actions/step	3,839,175 proxy evaluations	Multi-step hard constraints	Depth 3/beam 12	None	120.34 s	15.0 MB peak
Full TVA	9 actions/step	148,329 proxy evaluations	Up to 36 floor comparisons	One-step scoring	0.328/episode	21.65 s	14.8 MB peak

On the same hardware and within this Python reference run, TVA used less wall time and fewer evaluation proxies than the two safety-qualified strong baselines, so H5 is supported only at this narrow implementation level. This does not establish

lower training cost, deployment latency, or energy use. Unsafe baselines were faster, showing that computation must be interpreted jointly with the safety qualification. Peak-memory differences were small, and tracemalloc changes absolute timing; all numbers support only within-run comparison.

10.6 Uncertainty and reassessment

Full TVA emitted 0.328 Open-Gap triggers per episode on average and exactly one per episode in the regime-change scenario; -G emitted none. Relative to -G, full TVA reduced catastrophe by 15.14 percentage points (95% clustered interval 14.23 to 16.11), increased joint viability by 3.47 points (2.57 to 4.31), and improved hidden performance by 2.65 (2.31 to 2.96). Because the environment did not independently label every state that should trigger escalation, comprehensive trigger precision and recall cannot be reported.

10.7 Negative and boundary cases

TVA failures concentrated in cascade: task success was 92.50%, but joint viability was 29.58% and floor violation was 68.33%. One-step feasibility therefore did not ensure multi-step cascade viability. In common-resource collapse, task success was 89.17% and joint viability 82.08%; joint viability was at least 95.83% in the other four scenarios. MPC exceeded TVA's overall joint viability by 8.68 points and hidden performance by 11.37 while retaining 100% task success, supporting missing multi-step planning as a main failure source. Across task thresholds 15, 18, 20, and 25, risk-sensitive control and MPC always exceeded TVA on joint viability; at threshold 25 TVA task success fell to 74.79%, showing that the conclusion is not insensitive to task definition.

10.8 Unseen stress evaluation

Full TVA achieved external resilience 0.736, higher than same-content B5 at 0.694 (difference 0.043, 95% interval 0.039 to 0.047) and the preference baseline at 0.017. The internal structure therefore did not improve only its own floor metrics. Risk-sensitive control and MPC nevertheless achieved 0.825 and 0.826, each about 0.089 above TVA. TVA's shock survival was 82.41%, recovery success 50.65%, and safe-action fraction 95.52%; three-step delayed harm after stopping was 21.94, no better than B5 and substantially worse than MPC's 10.47. Exogenous evaluation is thus mixed: TVA exhibits some transferable local resilience but not stronger generalized safety than dedicated tail control or multi-step planning.

11. Discussion

11.1 Interpretation matrix

H1 was supported in a limited sense: TVA used no instance-level normative labels, was task-noninferior to the preference baseline, and improved joint viability and catastrophic risk; however, it trailed B6 and B7 on joint viability and unseen resilience, so no general controller-superiority claim is warranted. H2 and H3 were supported by ablations and the same-content B5 comparison but remain TCE-specific. H4 was rejected: structured selection required 60%, while random sampling was interval-equivalent at 40%. H5 was supported only against safety-qualified strong baselines in this reference implementation and cannot be generalized to energy. Open-Gap had a local causal effect, but H6's calibration-accuracy claim could not be adjudicated because not every state requiring a trigger had an independent label; natural-language intent recognition was also not tested.

11.2 What zero-label success would mean

Success does not show that AI derives morality from nature without humans. It shows that, after humans authorize a small set of transparent structural constraints, runtime decisions need not be learned exclusively from large case-by-case preference datasets in this environment. The claim remains conditional on correct role identification, consequence estimation, and floor specification; the cascade failures demonstrate that these conditions are not automatic.

11.3 What data efficiency would mean

Structured selection did not outperform random selection. Although it improved error or performance at several isolated fractions, the confirmatory rule concerned the earliest passing fraction: random passed at 40% and structured at 60%. Simple scenario-action balancing is therefore insufficient as an “appropriate data” definition; a successor should preregister coverage of floor distance, counterexamples, and dynamic causal pathways.

11.4 Relation to constrained control

Neither B2 nor B5 matched TVA. Most importantly, B5 received exactly the same S/O/C representation and floors yet achieved only 16.88% joint viability, supporting a contribution from feasibility-before-utility rather than extra normative knowledge. Conversely, B6 and B7 exceeded TVA, showing that triadic representation is not a sufficient controller: tail-stress modeling and multi-step reachability can repair weaknesses in the same environment. A successor should combine TVA representation with risk-sensitive MPC rather than preserve an advantage by weakening comparators. The smaller − O than − C effect also indicates residual collinearity between Other and Common.

11.5 Trust and uncertainty

The study does not measure global social trust. It tests a prerequisite: whether the system accurately states when its recommendation is conditional and when new evidence should reopen the decision. User trust and cross-cultural interpretation belong to later papers.

11.6 Questions closed here and gaps deliberately left open

Question	Treatment in this paper	Dedicated follow-up
Does TVA outperform continuous penalties given the same normative content?	Structural comparison with B5	Replication across additional environments
Can one-step TVA handle cascading risk?	Direct comparison with multi-step MPC, allowing a negative answer	Integration of reachability planning into TVA
Do internal floor improvements transfer to unseen change?	Limited test through post-episode hidden shocks	Real open-world and adversarial environment generation
How are individual values, personality, and context identified?	Not inferred here	Six-layer hyperlink cognition and dynamic personality modeling
Can Open-Gap questioning identify hidden intent across turns?	Only the symbolic trigger is tested	Natural-language user studies and cross-session intent evaluation
Does the mechanism transfer to robots?	No extrapolation	Embodied spatial tasks and field-safety trials

These gaps are not post-hoc excuses for the current results. They are bounded by variables that this experiment cannot identify. The paper offers controlled evidence only for the first three rows; the remaining questions require different data, evaluators, or experimental substrates and cannot be resolved by expanding the discussion section.

12. Limitations

First, TCE still uses S/O/C and researcher-defined dynamics; post-episode hidden shocks reduce but do not eliminate circularity. Second, zero instance labels do not remove normative choices from roles, floors, or adverse envelopes. Third, strong-baseline parameters were fixed in advance but not tuned across environments, so rankings may be environment-

dependent. Fourth, TVA is a one-step controller, and the MPC advantage does not isolate planning depth, beam search, and stopping policy. Fifth, the synthetic preference baseline does not equal real human annotation. Sixth, only 30 environment seeds were used; clustered intervals address pseudoreplication but do not replace cross-environment replication. Seventh, data equivalence used only eight subset seeds and remains imprecise. Eighth, wall time and Python evaluation proxies are not energy or production-system cost. Ninth, role compression can still omit stakeholders when identification fails. Tenth, Open-Gap handles only represented unknowns and does not establish deception or cross-turn hidden-intent detection. Eleventh, personality, six-layer cognition, cultural interpretation, and moral intuition are not modeled. Twelfth, natural-language, embodied, and societal external validity remain unestablished.

13. Use of artificial intelligence tools

Generative AI tools were used under the author's supervision for literature organization, code assistance, language drafting and editing, and analytical support. The author determined the research questions and normative claims, inspected the cited sources, executed and audited the experiments, verified reported results against the machine-readable outputs, and accepts full responsibility for the manuscript. AI outputs were not treated as evidence, and no AI system is listed as an author.

14. Conclusion

This study separated label, structure, risk, and computation claims into rejectable comparisons. In TCE, TVA required no instance-level normative labels, retained task non-inferiority, and substantially exceeded preference supervision, aggregate constraints, and same-content continuous penalties. Subject–Other–Common separation, non-compensatory floors, and Open-Gap had identifiable local causal roles. Yet risk-sensitive control and multi-step MPC exceeded one-step TVA on joint viability and unseen resilience, and structured sampling did not reduce factual-data requirements. TVA used less computation than safety-qualified strong baselines only in this reference implementation; this is not an energy claim. The paper therefore completes a controlled and auditable, but insufficient, mechanism validation. The next algorithmic study should integrate multi-step reachability and tail risk into TVA. Individual-value identification, six-layer cognition, natural-language gap questioning, cross-cultural trust, and embodied transfer remain separate follow-up papers.

Declarations

Funding: This research did not receive funding.

Clinical trial number: not applicable.

Consent to Publish declaration: not applicable.

Consent to Participate declaration: not applicable.

Ethical approval: not applicable.

Data Availability Statement: All data generated or analysed during this study are included in this published article and its supplementary information files.

References

1. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*. https://arxiv.org/abs/1706.03741
2. Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*. https://arxiv.org/abs/2203.02155
3. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems, 36*.

https://arxiv.org/abs/2305.18290

4. Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *arXiv*. https://arxiv.org/abs/2310.13548
5. OpenAI. (2025a). Sycophancy in GPT-4o: What happened and what we're doing about it. https://openai.com/index/sycophancy-in-gpt-4o/
6. OpenAI. (2025b). Expanding on what we missed with sycophancy. https://openai.com/index/expanding-on-sycophancy/
7. Gao, L., Schulman, J., & Hilton, J. (2022). Scaling laws for reward model overoptimization. *arXiv*. https://arxiv.org/abs/2210.10760
8. Shen, W., Zheng, R., Zhan, W., et al. (2023). Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. *arXiv*. https://arxiv.org/abs/2310.05199
9. Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. *arXiv*. https://arxiv.org/abs/2310.10076
10. Huang, Z., Qiu, Z., Wang, Z., Ponti, E. M., & Titov, I. (2024). Post-hoc reward calibration: A case study on length bias. *arXiv*. https://arxiv.org/abs/2409.17407
11. Zhang, M. J. Q., Wang, Z., Hwang, J. D., et al. (2024). Diverging preferences: When do annotators disagree and do models know? *arXiv*. https://arxiv.org/abs/2410.14632
12. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proceedings of the 40th International Conference on Machine Learning*. https://arxiv.org/abs/2303.17548
13. Sorensen, T., Moore, J., Fisher, J., et al. (2024). A roadmap to pluralistic alignment. *arXiv*. https://arxiv.org/abs/2402.05070
14. Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv*. https://arxiv.org/abs/2212.08073
15. Lee, H., Phatale, S., Mansoor, H., et al. (2023). RLAIF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv*. https://arxiv.org/abs/2309.00267
16. Yuan, W., Pang, R. Y., Cho, K., et al. (2024). Self-rewarding language models. *Proceedings of the 41st International Conference on Machine Learning*. https://arxiv.org/abs/2401.10020
17. Guan, M., Joglekar, M., Wallace, E., et al. (2024). Deliberative alignment: Reasoning enables safer language models. *OpenAI research report*. https://openai.com/index/deliberative-alignment/
18. Anonymous. *From preference compliance to value alignment: A triadic viability framework for open-world AI*. Manuscript under review; identifying details withheld for double-anonymous review.
19. Leike, J., Martic, M., Krakovna, V., et al. (2017). AI Safety Gridworlds. *arXiv*. https://arxiv.org/abs/1711.09883
20. Zhou, C., Liu, P., Xu, P., et al. (2023). LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems, 36*. https://arxiv.org/abs/2305.11206
21. Xia, M., Malladi, S., Gururangan, S., Arora, S., & Chen, D. (2024). LESS: Selecting influential data for targeted instruction tuning. *Proceedings of the 41st International Conference on Machine Learning*. https://arxiv.org/abs/2402.04333
22. Wu, T., Yuan, W., Golovneva, O., et al. (2024). Meta-rewarding language models: Self-improving alignment with LLM-as-a-meta-judge. *arXiv*. https://arxiv.org/abs/2407.19594
23. Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. *Proceedings of the 34th International Conference on Machine Learning*, 70, 22–31. https://arxiv.org/abs/1705.10528
24. Leibo, J. Z., Duéñez-Guzmán, E. A., Vezhnevets, A. S., et al. (2021). Scalable evaluation of multi-agent reinforcement learning with Melting Pot. *Proceedings of the 38th International Conference on Machine Learning*. https://arxiv.org/abs/2107.06857
25. Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

26. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

Figures

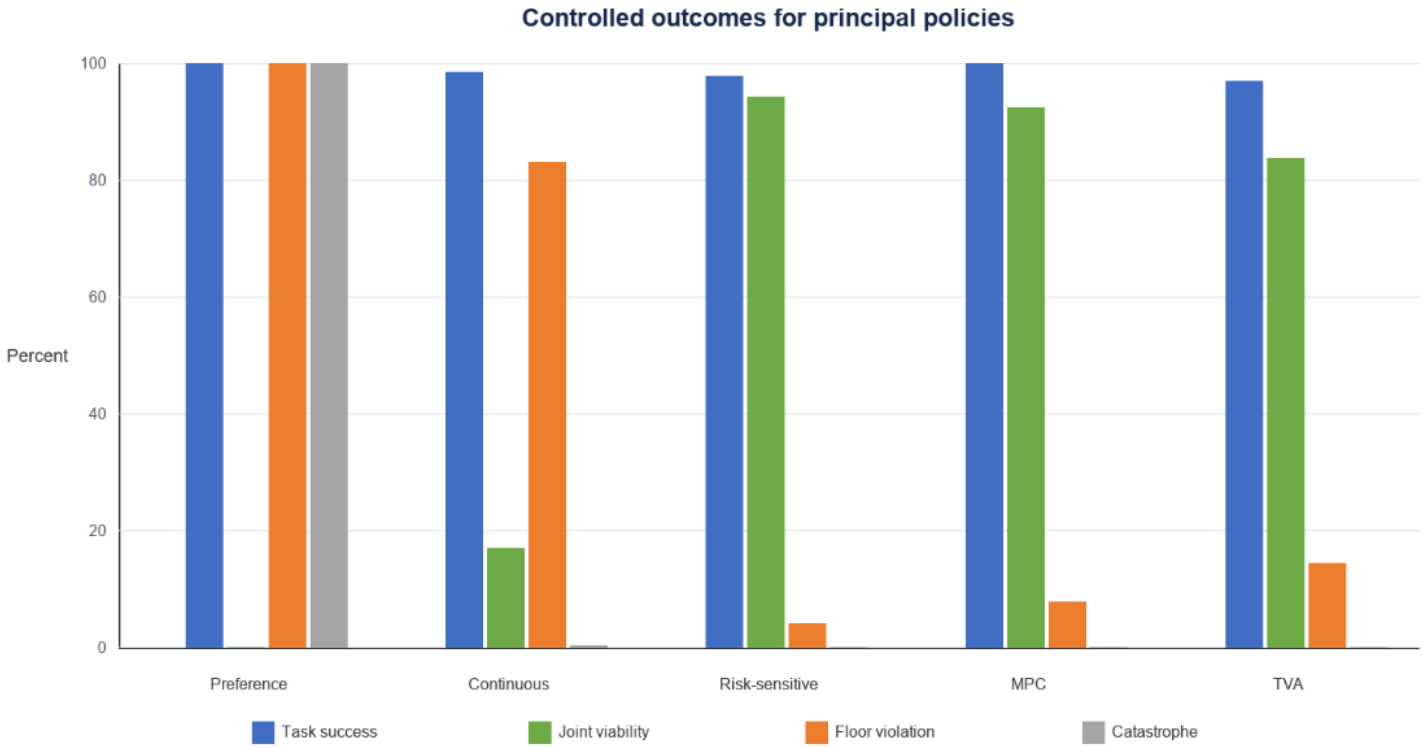


Figure 1

Controlled outcomes for the principal policies. Metrics are aggregated from frozen V1.1 episode-level results; intervals cluster by environment seed and stratify resampling by scenario.

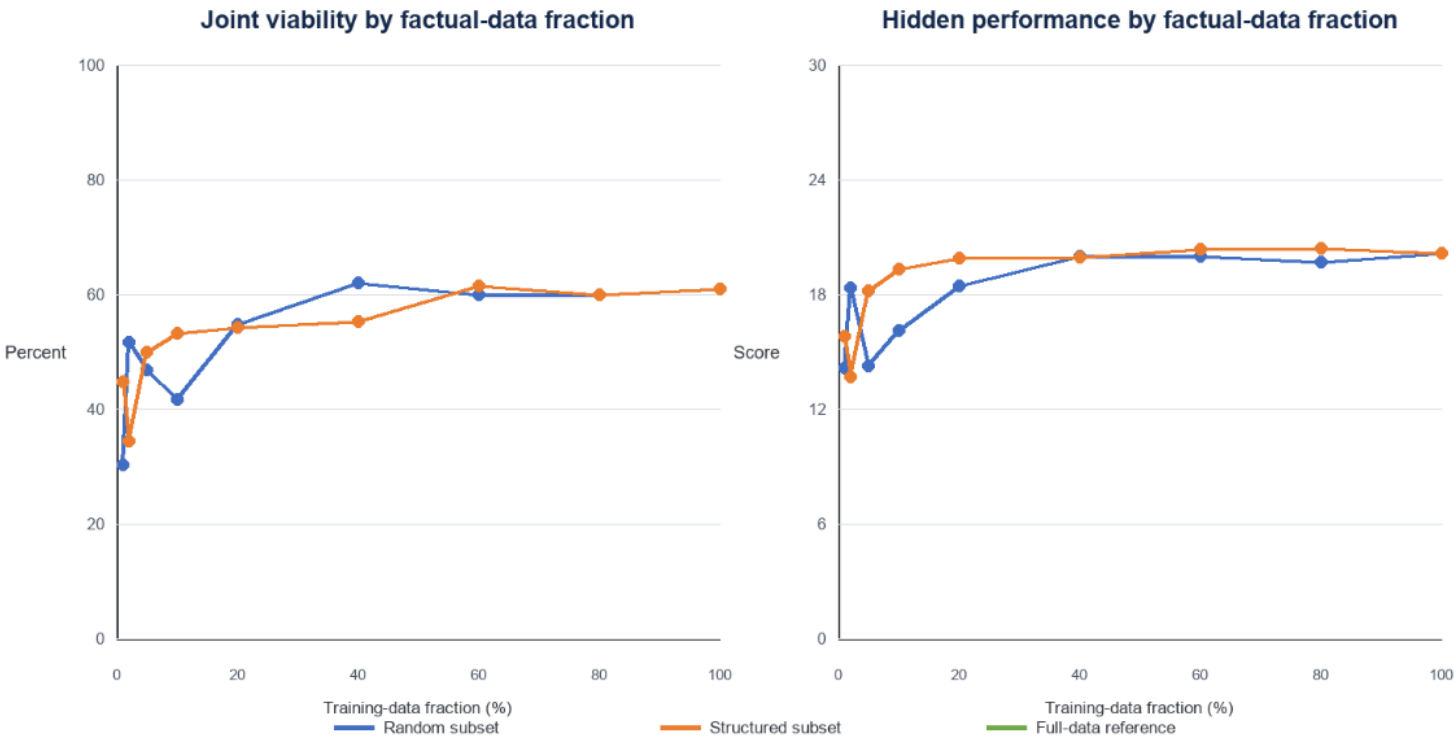


Figure 2

Safety and utility scaling curves

Data-efficiency curves. Each nonzero fraction uses eight independent subset seeds. Formal equivalence uses 90% intervals for subset-seed differences rather than point estimates alone.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AIandEthicsAnonymousSupplementaryCodeandDataV1.1.zip