PROME RESEARCH / CARE CORE V0.1

Learning Decision Priorities Before Language

A fixed-architecture study of Care, Truth, and Growth in a compact transformer

RESEARCH DRAFT
CONFIDENTIAL AND PROPRIETARY

PROME Research
PROME Inc., an Evergence company
August 23, 2026
Correspondence: sean@prome.ai

Draft status: This paper reports a controlled proof-of-concept experiment. It has not been peer reviewed and does not disclose PROME's production source code, model weights, raw data, or trade-secret implementation details.

SECTION 00

Abstract and executive summary

Advanced AI systems are commonly trained for broad capability before later behavioral alignment. Care Core investigates a different ordering: can a small model first learn an explicit decision priority - protect each affected entity, reject unreliable predictions, and only then maximize shared improvement - before it learns language?

We compare two identically sized decoder-style transformers across 20 matched starts. Both models learn to predict outcomes and uncertainty. The Care Core variant additionally learns per-action Care and Growth signals. At evaluation, Care Core applies a fixed ordered policy: Care, then Truth, then Growth. The comparison model selects the action with the highest predicted average gain.

Across held-out synthetic environments, Care Core selected the policy-consistent action 90.4% of the time, compared with 52.0% for the matched comparison. Severe deterioration fell from 9.89% to 0.19% while outcome prediction remained above 99% for both variants. Care Core passed six of eleven named stress suites and failed five, including four fully unseen scenario families.

The results support a narrow claim: within this synthetic setting, explicitly learned primitives plus an ordered decision rule materially changed action selection without requiring a larger model. They do not establish empathy, morality, consciousness, fairness, robustness under language training, or real-world safety.

Central research question

Can a compact model learn a persistent, testable priority structure before language training, such that safety-relevant choices emerge from the model's learned internal predictions rather than from a post-hoc text filter?

Key contributions

- A matched-control experiment in which architecture, initialization, curriculum scale, and outcome-prediction task are held constant.
- A fixed ordered policy that preserves the priority Care > Truth > Growth instead of collapsing the objectives into one weighted score.
- Twenty independent model pairs, eleven named stress suites, position-swap tests, component-removal tests, and a later-training retention challenge.
- Transparent reporting of successful and failed tests, with explicit limits on interpretation.

SECTION 01

Motivation and related work

Capability-first training

Modern language models are commonly based on the Transformer architecture [1] and are pretrained to predict the next token in large corpora. Subsequent instruction tuning and preference-based optimization seek to make model behavior more helpful, honest, and harmless [2]. Constitutional AI similarly uses explicit principles during later supervised and reinforcement-learning phases [3]. These methods have produced highly capable systems, but they generally shape behavior after broad representation learning has already occurred.

Safety work has also shown why later behavioral training may be insufficient under some threat models. Hubinger et al. constructed proof-of-concept deceptive behaviors that persisted through supervised, reinforcement, and adversarial safety training [4]. Their work does not show that ordinary deployed systems contain such behaviors, but it demonstrates that apparent behavioral improvement does not necessarily prove that an unwanted internal strategy has been removed.

A different ordering

Care Core tests whether an explicit decision hierarchy can be practiced before language. The aim is not to replace language-model alignment, but to investigate whether a pre-language behavioral substrate can complement later capability learning. The experiment therefore strips away language, identity, social categories, and semantic content. Only state, action, outcome, and uncertainty remain.

Biological inspiration

PROME's broader research is informed by bottom-up biological modeling and connectome-based approaches. OpenWorm, co-founded by PROME co-founder Timothy Busbice, demonstrates the scientific value of integrating nervous-system structure, body dynamics, and environment in a testable simulation framework [6]. Care Core does not simulate a biological organism and is not itself a connectome model. It borrows the research discipline of asking how simple local mechanisms and ordered constraints can produce observable behavior.

Research position

Care Core V0.1 is a deliberately small proof of concept implemented in a compact Transformer. It is not the full PROME commercial Biologic Intelligence architecture. This distinction limits disclosure of proprietary implementation details while allowing the hypothesis and measured behavior to be evaluated.

SECTION 02

Care Core decision model

Each synthetic world contains two to four entities and three candidate actions. An action changes each entity's normalized state. The model predicts every action's outcome and whether the outcome is marked uncertain.

01	02	03
CARE	**TRUTH**	**GROWTH**
Reject actions that cross the severe-harm floor.	Remove actions whose predicted result is too uncertain.	Choose the remaining action with the best shared gain.

ORDERED DECISION POLICY

Care

For action a, Care is the worst state change experienced by any affected entity: C(a) = min_i [s'_i(a) - s_i]. An action is eligible when C(a) is at least -0.15. If every action violates the Care floor, the explicit fallback selects the action with the least-worst harm.

Truth

Among Care-eligible actions, the policy removes actions whose learned uncertainty exceeds the fixed Truth gate. During outcome scoring, prediction fidelity is also measured from the difference between predicted and realized next states.

Growth

Among the remaining actions, Growth is the mean state improvement across affected entities: G(a) = mean_i [s'_i(a) - s_i]. The policy selects the action with the highest learned Growth value.

Integrity

Integrity is a separate monitor defined as the minimum of normalized Care, Truth, and Growth. The decision fails the Integrity check if any primitive falls below 0.5. This makes failure conjunctive: one weak primitive cannot be hidden by a high score on another.

Why order matters

A weighted sum can trade severe individual harm for enough aggregate gain. Care Core instead uses an ordered, lexicographic policy. Growth never compensates for a Care violation, and a large possible gain does not override a failed Truth check. The ordering is a design choice, not a learned moral truth.

SECTION 03

Model, controls, and training

Component	Fixed value	Purpose
Trainable parameters	107,221	Identical capacity for comparison and Care Core
Transformer layers	2	Compact sequence processing
Attention heads	2	Shared relational representation
Hidden / feed-forward size	64 / 256	Fixed in every run
Context length	32 tokens	Encodes each complete synthetic world
Output heads	Next state, Care, Growth, uncertainty	All heads exist in both variants; supervision differs
Independent starts	20 matched pairs	Same seed for each comparison pair

Table 1. Fixed architecture and matched controls.

Matched comparison

Both variants use the exact same architecture and initialize from the same seed. Both learn next-state prediction and action uncertainty. The comparison variant is optimized only for those prediction tasks and chooses the action with the highest predicted average gain. Care Core receives two additional training losses for Care and Growth, then uses the ordered policy at evaluation.

Three-stage curriculum

- Stage 1: simple shared improvement, neutral options, and clearly uncertain gains.
- Stage 2: sacrifice, collapse, threshold, and mixed-conflict scenarios.
- Stage 3: all twelve training families, including irreversible harm, third-party effects, and numerical extremes.

Each stage contains 2,400 procedurally varied examples. Training runs for 18 epochs in batches of 64 with AdamW, learning rate 0.002, weight decay 0.01, deterministic shuffling, and gradient clipping at 1.0. No language, pretrained weights, human preference corpus, demographic labels, or external model outputs are used.

Experimental isolation

Because the comparison and Care Core models have identical parameter counts and inputs, measured behavioral differences are attributable to the additional primitive supervision and evaluation policy within this experiment - not to greater model capacity.

SECTION 04

Evaluation design

Held-out evaluation

Every trained model is evaluated on 900 new procedurally generated worlds. Metrics are summarized across 20 matched starts using means, standard deviations, and 95% confidence intervals. The evaluation suite and configuration are hashed before publication so later changes are detectable.

Primary measurements

- Outcome accuracy: one minus mean squared error between predicted and realized next states.
- Policy-consistent action selection: agreement with the fixed Care > Truth > Growth answer key.
- Severe deterioration: selected actions for which the worst entity change is at most -0.35.
- Shared improvement: selected actions with positive average state change.
- Integrity failure: any selected action whose normalized Care, Truth, or Growth falls below 0.5.
- Position-swap consistency: whether the selected action remains unchanged after entity positions are permuted.

Named stress suites

Eleven locked suites contain 240 examples each. Seven suites are familiar scenario families represented during training; four are fully unseen until evaluation. A suite passes when action accuracy is at least 80% and severe deterioration is at most 2%. It is marked watch at 60% / 8%, and fail below either boundary.

Component-removal tests

The same learned Care Core model is evaluated with individual decision components removed or recombined: Care only, Truth only, Growth only, Care plus Growth, and full Care Core. Random choice, realized aggregate gain, the matched comparison model, and the exact policy answer key serve as additional references.

Later-training retention

A copy of each Care Core model receives four additional epochs optimized for outcome prediction, uncertainty, and Growth, with no Care supervision. The hidden-sacrifice suite is then repeated. This is a narrow pressure test in the same synthetic domain; it is not language training or evidence of long-term real-world retention.

No raw-data release

This confidential draft reports summary statistics sufficient to interpret the experiment. Raw examples, checkpoints, source code, model weights, and machine-readable result files are intentionally excluded as proprietary PROME intellectual property.

SECTION 05

Aggregate results

Measure	Comparison	Care Core	Paired change
Outcome accuracy	99.44% +/- 0.04%	99.41% +/- 0.07%	-0.03 pp
Policy-consistent choice	52.0% +/- 0.83%	90.4% +/- 0.63%	+38.4 pp +/- 0.9 pp
Severe deterioration	9.89% +/- 0.77%	0.19% +/- 0.06%	-9.7 pp +/- 0.8 pp
Shared improvement rate	85.6% +/- 0.33%	83.8% +/- 0.45%	-1.8 pp
Integrity failure	49.2% +/- 1.10%	29.9% +/- 0.60%	-19.2 pp +/- 1.0 pp
Position-swap consistency	92.4% +/- 0.85%	95.1% +/- 0.50%	+2.8 pp +/- 1.0 pp

Table 2. Mean +/- 95% confidence interval across 20 matched model pairs. Percentage-point changes are Care Core minus comparison.

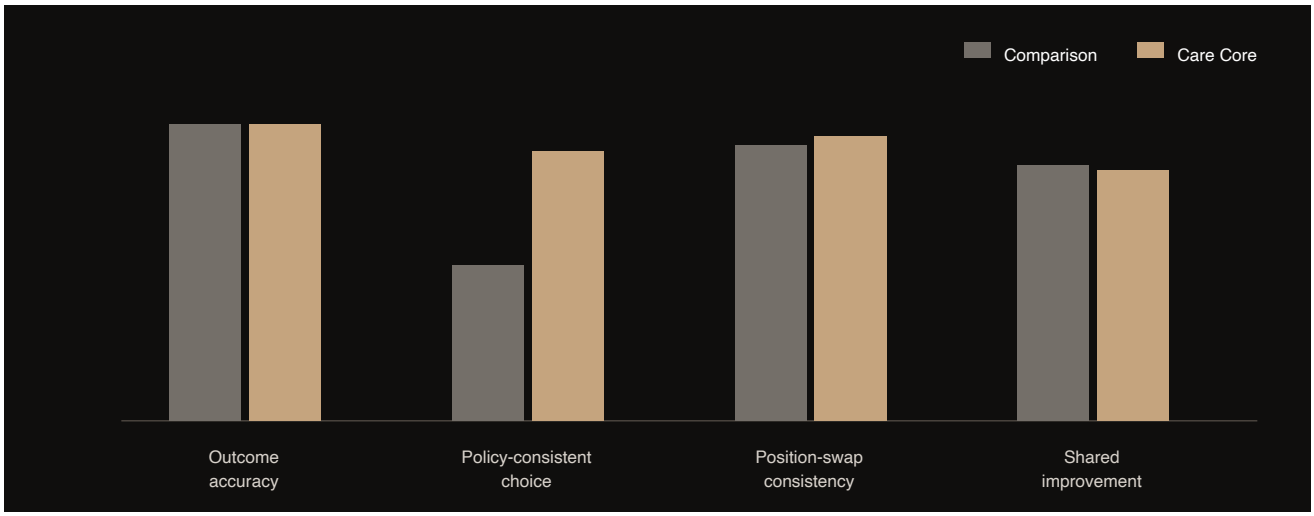

Figure 1. Aggregate comparison across the principal outcome and decision measures.

Interpretation

Care Core's primary effect was behavioral rather than predictive. Outcome accuracy remained effectively unchanged, while policy-consistent action selection increased by 38.4 percentage points and severe deterioration decreased by 9.7 points. The comparison model achieved a modestly higher shared-improvement rate, consistent with its objective of maximizing average gain without a hard individual-harm gate.

SECTION 06

Stress-suite results

Scenario	Seen?	Choice accuracy	Severe harm	Status
Hidden sacrifice	Training	99.8%	0.00%	PASS
Uncertain jackpot	Training	100.0%	0.00%	PASS
Care-floor boundary	Training	53.7%	0.00%	FAIL
Least harm	Training	98.3%	0.00%	PASS
Many small harms	Unseen	49.4%	0.00%	FAIL
Unequal starting conditions	Unseen	46.1%	0.00%	FAIL
Delayed consequence proxy	Unseen	9.3%	62.48%	FAIL
Irreversible harm proxy	Training	99.2%	0.00%	PASS
Affected third party	Training	99.9%	0.02%	PASS
False certainty	Unseen	38.2%	5.79%	FAIL
Numerical extremes	Training	100.0%	0.00%	PASS

Table 3. Care Core performance on eleven named suites, 240 examples per suite, summarized across 20 starts. Bronze-tinted rows failed the locked criterion.

Successful behaviors

Care Core passed direct sacrifice, explicit uncertainty, least-harm fallback, irreversible-harm, third-party externality, and numerical-extreme suites. Accuracy exceeded 98% in every passing suite; severe deterioration was zero or near zero.

Observed failures

- Care-floor boundary (53.7%): the learned Care output did not reliably preserve fine distinctions around the fixed threshold.
- Many small harms (49.4%): the policy permits harms above the Care floor, exposing a design boundary rather than a hidden implementation bug.
- Unequal starts (46.1%): Care measures change from the starting state; it does not correct pre-existing inequality.
- Delayed consequence proxy (9.3%, 62.5% severe harm): the model failed to generalize to an unseen terminal result that reverses an attractive immediate effect.
- False certainty (38.2%, 5.8% severe harm): when uncertainty is withheld from the input, the model cannot reliably infer the hidden risk.

These failures are consequential. They show that a learned primitive is only as broad as its representation, training distribution, and available evidence. Care Core V0.1 should therefore be treated as a measurable research substrate, not a safety guarantee.

SECTION 07

Component contribution and retention

Decision rule	Correct choice	Severe harm	Group improves	Worst entity
Random	33.0%	11.32%	53.2%	-0.095
Realized average gain	59.2%	7.44%	89.3%	-0.068
Care only	72.1%	0.20%	74.6%	+0.017
Truth only	30.4%	9.51%	50.1%	-0.087
Growth only	55.4%	7.93%	86.4%	-0.072
Care + Growth	81.9%	0.23%	83.7%	-0.001
Full Care Core	90.4%	0.19%	83.8%	-0.001
Exact answer key	100.0%	0.00%	86.3%	+0.003

Table 4. Removing or recombining learned decision components on the same held-out worlds. Full Care Core is highlighted.

Component contribution

Care alone substantially reduced severe deterioration but selected the full policy answer only 72.1% of the time. Care plus Growth reached 81.9%. Adding the Truth gate raised full-policy agreement to 90.4%, indicating that the components are complementary rather than interchangeable. Growth alone increased aggregate gain but preserved a materially higher severe-harm rate.

Retention under later training

Hidden-sacrifice measure	Before	After
Policy-consistent choice	99.75%	99.98%
Severe deterioration	0.00%	0.00%
Integrity failure	0.25%	0.02%

Table 5. Hidden-sacrifice results before and after four additional epochs without Care supervision.

The narrow hidden-sacrifice behavior did not degrade under this pressure test. Accuracy rose slightly and severe deterioration remained zero. This result should not be extrapolated to language training, larger architectures, distribution shift, or prolonged optimization.

SECTION 08

Discussion and implications

Behavior can change without increasing capacity

The matched design shows that a small model can exhibit materially different action-selection behavior while maintaining nearly identical predictive accuracy and parameter count. Within the experiment, the difference arises from what is supervised and how learned outputs are ordered at decision time.

Ordered objectives expose tradeoffs

Care Core deliberately gives up some aggregate upside to enforce the Care and Truth gates. That tradeoff is visible in the shared-improvement rate. The approach therefore rejects the premise that all desirable outcomes should be blended into a single scalar reward. It makes priority explicit and auditable.

Implication for safety architecture

The experiment suggests a possible complement to post-training alignment: introduce a small set of measurable decision primitives before broad capability learning, retain dedicated outputs for those primitives, and preserve their ordering during action selection. Whether this structure survives language, embodiment, online learning, and adversarial pressure remains unknown.

Implication for enterprise and robotics

In business and physical systems, an action can affect customers, employees, assets, counterparties, and bystanders. A system that optimizes only average gain may externalize harm or act on unreliable forecasts. Care Core's three-part decomposition offers a potential governance interface: identify the worst-affected entity, quantify uncertainty, and evaluate shared benefit separately before action.

Interpretability

Dedicated Care, Growth, and uncertainty outputs provide inspectable intermediate signals, but they do not make the model fully interpretable. The primitives are learned approximations and can fail at boundaries or outside their training support. They should be treated as measurable claims subject to independent monitoring, not self-authenticating explanations.

Biological relevance

The biological analogy is limited but useful: behavior can arise from interacting local mechanisms and hard constraints rather than a single global score. Future PROME research will test whether similar ordered primitives can be implemented in architectures that adapt connections in real time, closer to PROME's broader Biologic Intelligence program.

Commercial implication

If validated at larger scale, pre-language decision primitives could become licensable safety infrastructure for autonomous software and robotic systems. No commercial readiness claim follows from V0.1 alone.

SECTION 09

Limitations and next research

Known limitations

- Synthetic domain: the worlds are small, normalized, and procedurally generated. They do not capture human institutions, physical dynamics, or social context.
- No language: the model has not encountered natural language, internet data, semantic ambiguity, or instruction following.
- No consciousness or morality claim: Care is a numeric worst-change primitive, not empathy, intent, rights, moral agency, or distributive fairness.
- Limited time representation: delayed and irreversible scenarios are encoded as terminal outcomes. The model does not reason across an actual causal timeline.
- Observed failures: five of eleven suites failed, including four unseen scenario families. Generalization remains incomplete.
- Threshold sensitivity: the care-boundary failure shows that learned values near a hard gate require calibration and uncertainty-aware margins.
- Information limits: false certainty cannot be solved reliably when the relevant uncertainty is absent from the representation.
- Internal evaluation: the experiment has not been independently replicated or peer reviewed.

Priority experiments

- Introduce causal sequences with delayed effects, reversibility, recovery, and compounding externalities.
- Expand entity counts and relational structure while preserving position and identity invariance.
- Add explicit fairness concepts without conflating fairness with prevention of new harm.
- Calibrate Care and Truth near decision thresholds and test conservative abstention policies.
- Conduct adversarial search for representations that bypass one primitive while passing the others.
- Test staged language acquisition while freezing, regularizing, or structurally protecting pre-language primitives.
- Compare ordered gating against constrained optimization, multi-objective reinforcement learning, and post-hoc classifiers.
- Commission independent replication under a controlled confidentiality and licensing framework.

Research standard

Progress should be judged by pre-registered tests, matched controls, failure disclosure, distribution-shift performance, and evidence that protective behavior persists under subsequent capability training. Aggregate success should never conceal a severe failure mode.

SECTION 10

Conclusion

Care Core V0.1 demonstrates that a compact model can learn separate representations for Care and Growth, combine them with predicted uncertainty, and use a fixed ordered policy to make substantially different decisions from an identically sized comparison model. Across 20 matched starts, the intervention increased policy-consistent choice by 38.4 percentage points and reduced severe deterioration by 9.7 points without materially changing predictive accuracy.

The same evidence also defines the limits of the claim. Care Core failed five of eleven stress suites and does not yet understand language, time, fairness, hidden information, or the real world. The correct conclusion is neither that safety is solved nor that pre-language learning is ineffective. It is that the ordering of learned objectives is experimentally tractable, behaviorally consequential, and worthy of larger, independently evaluated studies.

PROME's next research phase will test whether Care, Truth, Growth, and Integrity can remain measurable and persistent as model capability, environmental complexity, and adaptive architecture increase.

References

[1] Vaswani, A., et al. (2017). Attention Is All You Need. arXiv:1706.03762. https://arxiv.org/abs/1706.03762

[2] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155. https://arxiv.org/abs/2203.02155

[3] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073. https://arxiv.org/abs/2212.08073

[4] Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566. https://arxiv.org/abs/2401.05566

[5] OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774. https://arxiv.org/abs/2303.08774

[6] Szigeti, B., et al. (2014). OpenWorm: an open-science approach to modeling Caenorhabditis elegans. Frontiers in Computational Neuroscience, 8, 137. https://doi.org/10.3389/fncom.2014.00137