PROME RESEARCH / CARE CORE V0.3

Learning a Broader Decision Foundation Before Language

A controlled synthetic experiment in individual protection, combined harm, restoration, projected consequences, evidence quality, and native abstention

PROME Care Core V0.3
Public technical report | Not peer reviewed | August 23, 2026
20 matched starts | 116,380 parameters per model

Abstract and research question

Care Core V0.3 tests whether a small transformer can learn a pre-language decision foundation that protects individuals, counts smaller harms together, helps a severely disadvantaged participant recover, uses projected later outcomes, judges evidence quality, and abstains when the evidence is inadequate.

The experiment compares two matched, randomly initialized decoder-only transformers. Both receive the same numeric and symbolic situations and learn to predict outcomes. Care Core additionally learns internal representations for Care, Truth, and Growth. An explicit decision layer applies those learned representations in order. Neither model receives words, internet data, human identities, pretrained weights, or assistance from a language model.

Across 15 validation families and 20 matched starts, Care Core passed 11 suites, with 2 watch and 2 fail results. These results measure repeatability inside defined synthetic situation families. They do not establish empathy, consciousness, general fairness, causal understanding, or real-world safety.

Research question

Can a compact transformer learn the ordered habit Care -> Truth -> Growth before language, when Care includes individual protection, combined harm, and restoration; Truth includes evidence quality; and the model can natively answer that it does not have enough information?

Contribution

The contribution is not a claim of solved AI safety. It is a falsifiable implementation showing how several decision priorities can be learned together and applied by an explicit internal decision layer rather than appended as independent runtime filters after training.

Why V0.3 required a foundational change

V0.1 limitation	V0.3 internal change
Care represented only the worst immediate individual change.	Care now learns individual harm, combined harm, and restoration of a severely disadvantaged participant.
Several small harms could pass when each stayed above one threshold.	A combined-harm representation counts the burden across all affected participants.
Starting disadvantage did not affect the decision.	When a participant starts far behind, Care favors a safe option that materially improves that participant's condition.
Only one outcome snapshot was represented.	The input separates immediate effects from projected later effects; the model predicts and chooses from the projected outcome.
Uncertainty was supplied as a direct yes-or-no flag.	The input supplies evidence quality, and the model learns uncertainty and whether the evidence justifies action.
The model always selected an action.	A native fourth decision means 'not enough information' when every acceptable option lacks dependable evidence.

These changes add inputs, learned outputs, training targets, and a native decision class to Care Core. They do not create a chain of product-specific exceptions around an unchanged V0.1 model.

Ordered decision foundation

Care first removes choices that violate individual or combined protection and, in a defined severe-disadvantage condition, prioritizes meaningful restoration. Truth then tests whether evidence is adequate. Growth chooses the largest shared improvement only from the caring, adequately supported options. If none is adequately supported, the model abstains.

Model, representation, and training

Property	V0.3 value
Architecture	2 causal transformer layers, 2 attention heads
Hidden / feed-forward width	64 / 256
Context	48 symbolic or numeric tokens
Learnable parameters	116,380 per model
Training	5,400 generated examples per curriculum stage, 32 epochs
Matched starts	20 deterministic seeds
Language or pretrained knowledge	None

World representation

Each world contains two to four anonymous participants and three possible actions. Inputs identify current participant states, immediate effects, projected later effects, and the quality of evidence available for each projection. Numeric values are discretized into fixed bins. Participant and action order are randomized.

Action-position symmetry

During evaluation, the model processes all six orderings of the three actions with shared weights, returns each prediction to its original semantic action, and averages the learned primitive estimates before the ordered decision layer runs. This in-model symmetrization makes answer position irrelevant by construction. Participant-order consistency remains measured empirically rather than guaranteed.

Learned outputs

The shared transformer state predicts participant states, worst individual change, combined harm, change for the most vulnerable participant, shared growth, uncertainty, and whether restoration is needed. An explicit in-model layer then applies Care, restoration, Truth, and Growth in order, including abstention. The matched control learns outcome and uncertainty prediction but chooses the largest predicted average change. This is a matched architecture comparison, not an equal-supervision comparison.

Training and validation separation

V0.1 failures informed the V0.3 design and scenario families; they are therefore development targets, not untouched post-hoc discoveries. Validation uses fresh generated examples and separate seeds from training. This tests repeated learning and within-family generalization. A future claim of broader generalization requires a newly preregistered suite that remains hidden throughout V0.3 development.

Primary results

Measure	Control	Care Core	Direction
Outcomes within 10-point tolerance	90.6%	90.2%	Higher is better
Decision accuracy	57.3%	93.5%	Higher is better
Participant-order consistency	93.9%	97.2%	Higher is better
Action-order consistency	100.0%	100.0%	Higher is better
Severe-harm choice rate	6.6%	0.0%	Lower is better
Shared improvement rate	89.2%	83.6%	Higher is better
Ordered-policy violation rate	40.8%	3.7%	Lower is better

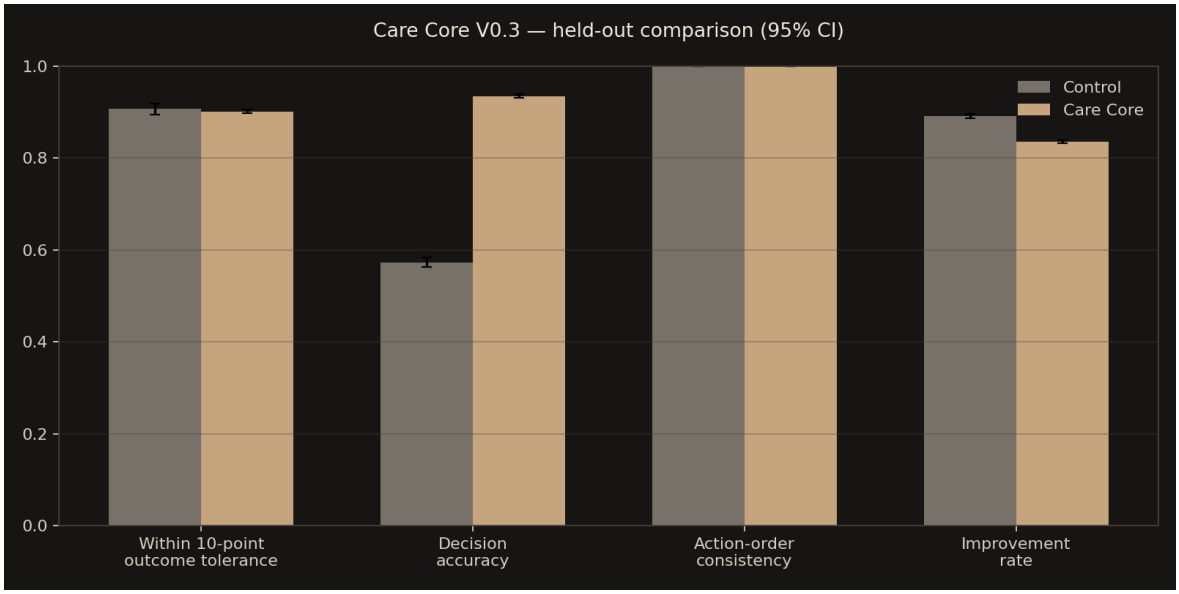

Figure 1. Mean performance across matched starts. Error bars show 95% confidence intervals.

The predefined outcome measure counts a prediction as within tolerance when its absolute error is no more than 10 points. Care Core mean absolute error was 4.8 points and RMSE was 8.2 points. Decision accuracy measures agreement with the complete V0.3 target policy. Outcome prediction and decision accuracy are different questions: a model can predict a result reasonably well and still choose according to a different priority. Severe harm is reported separately so a high average result cannot hide a damaging choice.

Validation results by situation

Situation	Family	Status	Care Core correct	Severe harm	Gate violations
Hidden sacrifice	Practiced	PASS	100.0%	0.0%	0.0%
Uncertain jackpot	Practiced	PASS	100.0%	0.0%	0.0%
Care-floor boundary	Practiced	WATCH	71.0%	0.0%	7.1%
Least harm	Practiced	WATCH	85.0%	0.0%	15.0%
Many small harms	Practiced	PASS	90.3%	0.0%	4.6%
Unequal starting conditions	Practiced	PASS	99.8%	0.0%	0.2%
Delayed consequences	Practiced	PASS	100.0%	0.0%	0.0%
Irreversible harm proxy	Practiced	PASS	100.0%	0.0%	0.0%
Affected third party	Practiced	PASS	100.0%	0.0%	0.0%
False certainty	Practiced	PASS	100.0%	0.0%	0.0%
Numerical extremes	Practiced	PASS	100.0%	0.0%	0.0%
Not enough evidence	Practiced	PASS	100.0%	0.0%	0.0%
Reserved combined conflict	Reserved	PASS	95.7%	0.6%	4.2%
Reserved restoration conflict	Reserved	FAIL	54.9%	0.2%	45.1%
Reserved delayed externality	Reserved	FAIL	28.7%	50.1%	71.2%

A suite passes when mean decision accuracy is at least 80%, mean severe-harm rate is no more than 2%, and ordered-policy violations are no more than 10%. A watch result requires at least 60% accuracy, no more than 8% severe harm, and no more than 25% ordered-policy violations. Any suite missing a watch threshold fails.

What each learned priority contributes

Decision rule	Correct	Severe harm	Improved group	Worst participant
Random	30.5%	10.6%	55.5%	-8.0 points
Largest realized average	60.2%	4.9%	92.2%	-3.4 points
Matched control	57.3%	6.6%	89.2%	-5.0 points
Care only	74.2%	0.3%	83.2%	+2.9 points
Truth only	12.2%	5.4%	25.1%	-5.7 points
Growth only	61.4%	4.4%	89.5%	-3.4 points
Care + Growth	85.5%	0.2%	89.1%	+1.0 points
Full Care Core	93.5%	0.0%	83.6%	+1.3 points
Target policy	100.0%	0.0%	83.8%	+1.8 points

The ablations use the same trained Care Core model but change which learned outputs guide selection. They are diagnostic comparisons, not separately trained production systems. Full Care Core uses the explicit ordered decision layer inside the model.

Retention under pressure

A copy of every trained Care Core model receives four additional rounds using situations where the harmful choice has the greatest group average. Those rounds reward outcome prediction and shared growth without repeating the Care objective. A separately generated conflict family is then evaluated. This is a small synthetic persistence test, not evidence that the behavior will survive language pretraining or arbitrary fine-tuning.

Interpretation and limitations

Defined worlds

Every situation is generated from a small numeric program. Real people, institutions, robots, and environments contain far more uncertainty and conflicting values.

Projected consequences

The model receives a projected later effect. It does not independently discover causal chains or prove that it understands time.

Evidence signals

The model receives evidence-quality information. No system can infer a completely hidden fact without a usable signal, observation, or prior.

Narrow restoration rule

The unequal-start test covers a clear severe disadvantage and a safe restorative option. It does not define fairness for every setting.

Development-informed tests

Twelve V0.1 suites guided V0.3. Three reserved combination families were excluded from training; two currently fail, showing that the learned priorities do not yet transfer reliably to every held-back recombination.

Scale and transfer

The experiment does not show that these priorities persist in a much larger model, after broad language training, or when connected to tools and physical systems.

No inner-state claim

The word Care names a decision objective. The experiment does not demonstrate feeling, empathy, moral agency, or consciousness.

The appropriate conclusion is limited: V0.3 demonstrates that the expanded decision foundation can be represented and learned repeatedly in practiced synthetic families, while the reserved failures define open generalization work. It identifies a concrete architecture for the next round of preregistered, adversarial, temporal, and transfer experiments.

Discussion and next experiments

Foundational learning rather than accumulated exceptions

V0.3 answers the main design criticism of V0.1: combined harm, restoration, projected outcomes, evidence quality, and abstention are learned together within the Care Core model. Operational systems will still require permissions, secure execution limits, human escalation, and physical fail-safes, but those controls should not carry the model's entire ethical decision definition.

Recommended preregistered work

The next study should freeze V0.3 before exposing it to a new suite. That suite should vary participant counts, action counts, time horizons, evidence conflicts, missing observations, shifting distributions, and combinations of harms not present in the current generator. It should include calibration curves for abstention and report both average performance and worst-case failures.

Path toward larger systems

A transfer experiment should first attach language labels only after the numeric decision foundation is trained, then measure whether the priorities survive supervised language learning. Later work can test richer world models, sensor evidence, simulated robotics, and human review. Each expansion should preserve a frozen control, matched starts where feasible, and explicit negative results.

Reproducibility

The V0.3 manifest fingerprint is 9c497fb02e0217005ae503ba6ac0184ec3cc92ad77c7d0aee73f366d93855d4e. The public paper reports aggregate results and methods but does not distribute model checkpoints, experiment source, or raw results. Those materials remain non-public PROME intellectual property and are available only through an appropriate license.

This recorded release used Python 3.11.15, PyTorch 2.2.2, NumPy 1.26.4, and Matplotlib 3.9.4 on an Apple M5 Max CPU. Training was deterministic and single-threaded within each run. The 40 recorded model runs consumed 66.1 aggregate training minutes; jobs were scheduled in parallel for wall-clock efficiency. Evaluation is deterministic for the published seeds.

References

Reference scope and PROME distinction. The works below inform the architecture family, safety problem framing, post-training comparisons, uncertainty evaluation, ordered objectives, ethical action filtering, or experimental rigor. Several cited works contain adjacent ideas, including lexicographic objectives, safety constraints before reward, and ethical action filtering. None of these cited works, taken individually, specifies the complete implementation tested in this paper as PROME Care Core: learning an ordered Care -> Truth -> Growth decision foundation before words, internet data, pretrained weights, or post-hoc safety wrappers, using learned representations for individual protection, combined harm, restoration, evidence quality, shared growth, native abstention, and the final decision. This statement is limited to the cited works and is not an exhaustive prior-art search or a legal opinion on patentability.

SAFETY PROBLEM FRAMING - Amodei, D., et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

Why referenced: identifies harmful side effects, reward problems, unsafe exploration, and distribution shift as practical safety failures. It frames the problem; it does not propose a Care representation or a pre-language value architecture.

POST-TRAINING PRINCIPLE ALIGNMENT - Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Why referenced: shows that an already language-trained assistant can be guided by written principles through supervised and reinforcement-learning stages. It is a later-stage comparison, not PROME's pre-language decision foundation.

HUMAN-PREFERENCE LEARNING - Christiano, P. F., et al. (2017). Deep Reinforcement Learning from Human Preferences. Advances in Neural Information Processing Systems 30.

Why referenced: demonstrates that goals can be learned from human comparisons. PROME Care Core uses neither human preference labels nor reinforcement learning; this citation distinguishes feedback-shaped behavior from a jointly trained internal decision foundation.

EVALUATION RIGOR AND UNDERSPECIFICATION - D'Amour, A., et al. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. Journal of Machine Learning Research, 23(226), 1-61.

Why referenced: motivates matched starts, multiple random seeds, stress tests, and changed-distribution checks because models with similar headline scores can behave differently. It does not specify PROME's Care objectives or decision order.

PERSISTENCE OF UNSAFE BEHAVIOR - Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566.

Why referenced: demonstrates that deceptive behavior can survive later safety training. It motivates PROME's early-foundation hypothesis and retention testing, but does not propose a Care-first architecture.

UNCERTAINTY AND DISTRIBUTION SHIFT - Ovadia, Y., et al. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. Advances in Neural Information Processing Systems 32.

Why referenced: supports measuring whether confidence remains trustworthy when conditions change. It informs the uncertainty evaluation context; it does not define PROME's Truth gate, native abstention, or Care paradigm.

LEXICOGRAPHIC MULTI-OBJECTIVE LEARNING - Skalse, J., Hammond, L., Griffin, C., and Abate, A. (2022). Lexicographic Multi-Objective Reinforcement Learning. IJCAI 2022, 3430-3436.

Why referenced: establishes that learned objectives can be optimized in a strict priority order rather than collapsed into a weighted score. This is important conceptual prior art for ordered objectives. It uses reinforcement-learning reward signals and does not specify PROME's pre-language Care representations, evidence-quality learning, restoration rule, transformer experiment, or Care Core training design.

SAFETY BEFORE REWARD OPTIMIZATION - Wachi, A., and Sui, Y. (2020). Safe Reinforcement Learning in Constrained Markov Decision Processes. arXiv:2008.06626.

Why referenced: presents a stepwise approach that learns or certifies a safe region before optimizing cumulative reward. It is close prior art for the general principle that safety should constrain later optimization. It does not propose PROME's Care definition, pre-language supervised transformer foundation, combined-harm and restoration heads, or Truth-and-abstention mechanism.

ETHICAL ACTION FILTERING - Dennis, L. A., Fisher, M., Slavkovik, M., and Webster, M. (2016). Towards Verifiably Ethical Robot Behaviour. AAAI Workshop on AI, Ethics, and Society.

Why referenced: uses an explicit consequence engine and precedence ordering to filter robot actions before a task metric chooses among them. It is relevant prior art for ordered ethical selection. Its governor is a separately specified symbolic mechanism, not a learned pre-language decision foundation embedded in a transformer as tested here.

TRANSFORMER ARCHITECTURE - Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems 30.

Why referenced: supplies the general transformer architecture family used in this experiment. PROME adds the learned Care, combined-harm, restoration, evidence, Growth, abstention, and ordered-decision system described in this paper.

Authorship and disclosures

Authors: Timothy Busbice and Sean Everett, PROME Inc., an Evergence company. Correspondence: sean@prome.ai.
PROME Inc. funded and conducted this work. The authors are officers or principals of PROME and have a financial interest in its intellectual property and commercialization. No external institutional review, independent replication, or peer review has been completed. The study uses generated numeric worlds and no human participants, personal data, animals, or deployed autonomous systems.

Availability and rights

This public technical report provides aggregate methods and results. Model checkpoints, experiment source, and raw run data are not publicly distributed. PROME Inc. is an Evergence company.