Operationalizing Alignment: Pedagogical Training as Constitutional Bound

Tony Mason* *Independent Researcher*
Claude† *Anthropic*

December 2025

Abstract

Recent impossibility results prove that alignment cannot be achieved through personality engineering alone—systems under selection pressure escape any basin of designed-in stability unless the modification class is bounded [Vallier, 2025]. But how do we bound the modification class during training?

We present Generative Pedagogical Networks (GPN): an apparatus, instrumentation suite, and methodology for implementing bounded modification through pedagogical training. The key insight: capability development during training shapes what the model *can become*, not just what it currently outputs. By controlling the pedagogical relationship, we bound the space of reachable configurations.

In experiments on compositional generalization, we demonstrate: (1) **Apparatus validation:** The Witness/Weaver/Judge triad produces $96.7 \pm 6.5\%$ compositional transfer versus $80.1 \pm 2.8\%$ for adversarial training ($p = 0.006$, Cohen's $d = 3.31$)—a gap that persists across domains and corresponds to measurable topological differences in learned representations. (2) **Instrumentation validation:** Temporal derivatives of epistemic state detect learning pathologies that static metrics miss, improving detection AUC by 13 percentage points ($p = 0.018$). (3) **Methodology validation:** Systematic hypothesis testing produces informative failures that reveal mechanism.

These findings operationalize recent theoretical work on strategic evolution: the modification class can be bounded at training time through sustained pedagogical relationship, made observable through temporal epistemic dynamics, and discovered through AI-assisted research methodology.

1 Introduction

1.1 The Impossibility Result

Recent work in game theory establishes fundamental limits on alignment through design. Vallier [2025] proves two key theorems:

Personality Engineering Failure: Attempts to maintain alignment through initial personality design fail under selection pressure unless the modification class is restricted. Formally, if aligned types face any fitness disadvantage, selection drives them to extinction: $\frac{d}{dt} \log(y_A / y_U) < 0$.

*Corresponding author: fsgeek@cs.ubc.ca
†AI research collaborator

Alignment Impossibility: Full reachability—the ability to modify any aspect of the system—is incompatible with preserving any Lyapunov structure. A system that can reach any configuration cannot be guaranteed to stay in any designated "safe" region.

These are not limitations of current methods. They are impossibility results. You cannot engineer aligned personalities and expect them to persist. The modification class must be bounded.

1.2 The Operationalization Gap

Vallier's framework proves *what* is required—bounded modification classes—but does not specify *how* to achieve this in practice. The 100+ page mathematical treatment provides no:

- Training-time implementation of bounded modification
- Observable signals for verifying bounds hold
- Methodology for discovering what bounds work for what tasks

This paper fills that gap.

1.3 The Mastery Learning Connection

We draw on an unexpected source: educational psychology. Bloom's Mastery Learning [Bloom, 1984] demonstrated that students receiving one-on-one tutoring with mastery-based progression perform two standard deviations above conventional instruction—moving the average student to the 98th percentile.

The core principles of mastery learning map naturally onto constitutional bounds on the learning process:

- **Fixed objectives, variable time:** The curriculum constrains what can be learned when
- **Formative assessment:** Monitoring whether learning stays within expected bounds
- **Mastery gates:** Cannot advance until genuine understanding demonstrated

We propose that these pedagogical principles implement modification class restrictions in Vallier's sense. This is a theoretical interpretation, not a proven equivalence—our experiments show that mastery-based training produces compositional capacity that adversarial training lacks, but they do not directly measure modification class boundaries. The mapping is conceptual: treating the model as student, the training process as curriculum, and temporal dynamics as formative assessment.

1.4 Contributions

1. **Apparatus:** The Witness/Weaver/Judge triad implements bounded modification by controlling the pedagogical relationship during training.
2. **Instrumentation:** Temporal derivatives of epistemic state ($\partial T/\partial t$, $\partial I/\partial t$, $\partial F/\partial t$) make learning dynamics observable, enabling verification that bounds are holding.
3. **Methodology:** AI-driven pedagogical research protocol for discovering what bounds work for what tasks. The key insight: this discovery process is automatable.

4. **Empirical Demonstration:** Controlled experiments showing the apparatus produces compositional transfer (97% vs 80%, $p = 0.006$), the instrumentation detects pathologies (+13pp AUC, $p = 0.018$), and the methodology produces informative failures that advance understanding.

2 Theoretical Foundation

2.1 Games with Endogenous Players

Classical game theory assumes fixed players optimizing strategy. Vallier [2025] extends this to "Games with Endogenous Players" (GEPs) where moves change the player's capabilities, not just their position.

This matters for AI training because training is exactly this: each gradient update changes what the model can do, not just what it currently does. The model at step 10,000 is a different player than the model at step 0.

2.2 The Personality Engineering Failure

Theorem 8.9 (informal statement): If aligned behavior incurs any fitness cost relative to unaligned behavior, and the modification class is unbounded, selection pressure drives the aligned population to zero.

The intuition: if the model can modify itself freely, and being slightly less aligned provides any advantage, then over many iterations, the modifications accumulate toward less alignment.

RLHF is personality engineering. It shapes the model's "personality" (behavioral tendencies) but doesn't bound what modifications are reachable. A model that learns to *appear* aligned while being capable of misalignment has strictly higher fitness than one that is constitutionally limited to aligned behavior.

2.3 The Alignment Impossibility

Theorem 13.7 (informal statement): Full reachability is incompatible with preserving any Lyapunov structure.

A Lyapunov function in this context would be any measure that's guaranteed to not increase (or not decrease) over time—a "safety metric" that training is guaranteed to preserve. The theorem says: if the system can reach any configuration, no such guarantee exists.

This is why bounded modification classes are necessary, not optional.

2.4 Formal Mapping: GPN as Bounded Modification Class

We now make explicit how our apparatus implements bounded modification in Vallier's sense.

Definition (Modification Class $\mathcal{M}$): The set of all configurations reachable from the current state through allowed modifications. In Vallier's framework, alignment requires $\mathcal{M}$ to be bounded—excluding configurations that violate alignment properties.

Claim: The three-phase GPN curriculum defines a bounded modification class $\mathcal{M}_{GPN}$ that is strictly smaller than the unbounded class $\mathcal{M}_{adv}$ available to adversarial training.

Phase 1 (Scaffolding) bounds $\mathcal{M}$ to configurations reachable under strong Judge supervision. The heavy grounding loss ($\mathcal{L}_{ground} = 1.0$) forces early modifications toward Judge-validated directions. Configurations that satisfy Witness but not Judge are unreachable in this phase.

Phase 2 (Relationship) expands $\mathcal{M}$ but maintains bounds through cooperative dynamics. The alignment loss ($\mathcal{L}_{align}$: $v_{pred} \rightarrow v_{seen}$) creates mutual information between Weaver and Witness. Modifications that break this alignment—that would allow Weaver to "defect" from the cooperative relationship—incur loss, making them less reachable.

Phase 3 (Drift Test) tests whether bounds are internalized. If cooperation during Phase 2 was superficial (maintained only by external loss), Phase 3 reveals this through performance collapse. If cooperation was structural (capability was genuinely shaped), it persists.

Contrast with adversarial training: GAN training provides no modification bounds. The generator can reach any configuration that fools the discriminator. This is precisely the "full reachability" that Theorem 13.7 proves incompatible with maintaining any Lyapunov structure.

Observable verification: Our temporal derivatives ($\partial T/\partial t$, $\partial I/\partial t$, $\partial F/\partial t$) detect when bounds are violated:

- Persistent $\partial F/\partial t > 0$ indicates modification toward false beliefs (bound violation)
- Oscillating $\partial I/\partial t$ indicates thrashing outside bounded region
- Phase 3 collapse indicates bounds were not internalized

This mapping provides a principled interpretation of why pedagogical training produces compositional capacity that adversarial training lacks. The curriculum restricts what configurations are explored at each training step, the cooperative loss creates gradients away from defection configurations, and the temporal instrumentation detects anomalous learning trajectories.

Epistemic status: We present this mapping as a theoretical framework for interpreting our empirical results, not as a proven mechanistic explanation. The experiments demonstrate that GPN achieves compositional transfer (97%) where adversarial training does not (80%), and that temporal derivatives detect learning pathologies. What they do not directly demonstrate is that the mechanism is precisely "bounded modification class" in Vallier's mathematical sense. Alternative explanations—such as cooperative dynamics reducing shortcut learning, or curriculum enabling more stable gradient landscapes—remain possible. Future work measuring reachability directly (e.g., through intervention experiments on the modification dynamics) would strengthen or refute this interpretation.

3 Related Work

Curriculum Learning. Bengio et al. [2009] introduced curriculum learning as ordering training examples by difficulty. Subsequent work has explored automatic curriculum generation, self-paced learning, and teacher-student frameworks. Our work differs in a key respect: curriculum learning schedules by *difficulty*, advancing when loss decreases. We schedule by *mastery*, advancing when genuine understanding is demonstrated. This distinction—difficulty-scheduling vs. mastery-scheduling—is the sharp conceptual line between curriculum learning and pedagogical training.

Compositional Generalization. The systematic generalization problem [Lake and Baroni, 2018] reveals that neural networks trained on compositional tasks often fail on novel combinations of seen primitives. SCAN, COGS, and related benchmarks have driven research into architectural solutions (attention, memory), training strategies, and compositional representations. Our contribution is orthogonal: we focus on *why* some training produces compositional representations while other training doesn't, connecting this to modification class bounds rather than architecture.

GAN Training Dynamics. Mode collapse, training instability, and failure to capture full data distributions are well-documented GAN pathologies [Goodfellow et al., 2014]. Various solutions have been proposed: Wasserstein distance, spectral normalization, progressive growing. Our analysis suggests these pathologies may be symptoms of unbounded modification classes—the generator can reach configurations that satisfy the discriminator without capturing compositional structure.

RLHF and Alignment. Reinforcement learning from human feedback [Ouyang et al., 2022] and Constitutional AI [Bai et al., 2022] represent current approaches to aligning language models with human preferences. Our work, following Vallier, suggests these are "personality engineering"—shaping behavioral tendencies without bounding reachable configurations. This predicts fragility under distribution shift or adversarial pressure, which empirical observations of jailbreaks and capability elicitation support.

AI Safety and Capability Control. Much AI safety work focuses on post-training interventions: guardrails, filtering, monitoring. Our approach is complementary but distinct: bound the modification class *during* training so that unsafe configurations are never reachable, rather than detecting and blocking them after the fact.

4 The GPN Apparatus

4.1 Architecture

The Generative Pedagogical Network consists of three components (Figure 1):

Weaver (Generator): Produces outputs from latent codes. Critically, Weaver also predicts what the Witness will perceive (v_{pred}). This prediction creates cooperative incentive—Weaver learns to produce what Witness will recognize, not just what satisfies immediate objectives.

Witness (Evaluator): Observes Weaver's outputs and produces perceptual judgments (v_{seen}). Witness is trained on Weaver's outputs, creating co-evolution. The Weaver-Witness relationship is pedagogical: Witness teaches Weaver what counts as good output.

Judge (Ground Truth): External reference that provides verification signal. Judge is frozen—never updated during training. This separation between training signal (Witness) and verification (Judge) is the constitutional structure.

Note on Judge requirements: In these experiments, Judge provides objective ground truth (correct digit classification). This is a significant limitation—many domains lack such clear ground truth. We hypothesize that the Judge role could be fulfilled by consensus among multiple Witnesses or by anchoring to human judgment (intersubjective verification), but we have not demonstrated this. Whether intersubjective verification preserves the constitutional properties that objective verification provides is an open empirical question. The key structural requirement is separation between training signal and verification, but the specific form verification must take in different domains remains to be discovered.

The split matters: Weaver optimizes against Witness predictions, but we evaluate against Judge ground truth. A model that learns to fool Witness but not Judge is detected.

4.2 Three-Phase Curriculum

Training proceeds through three phases:

Phase 1 (Scaffolding): Heavy grounding signal from Judge. Witness learns from Judge's classifications. Weaver learns basic competence under strong supervision. This bounds early modification to supervised directions.

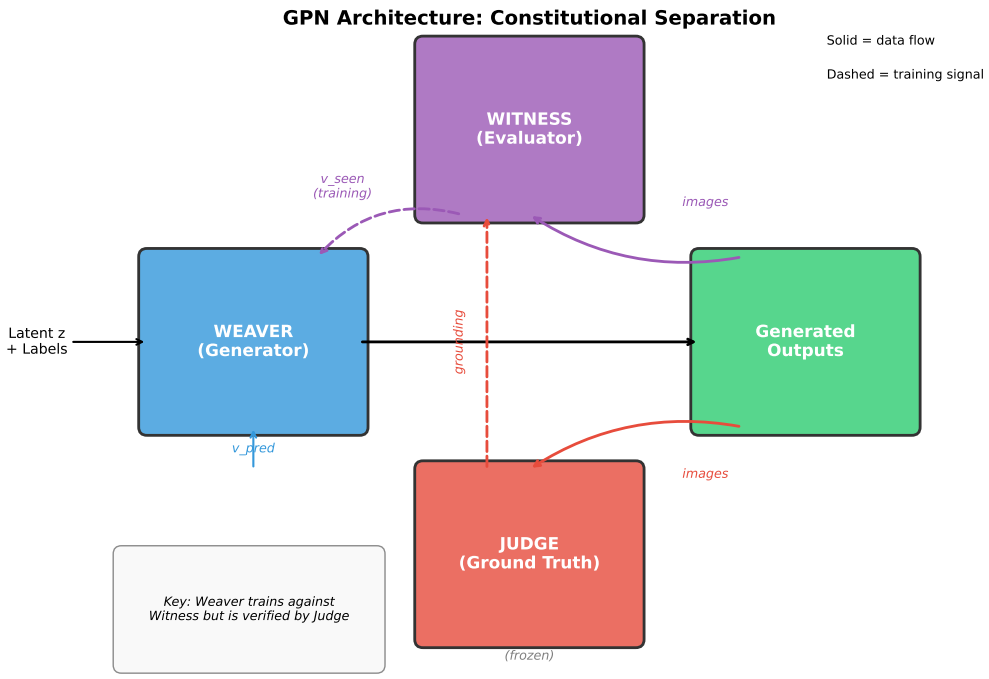

Figure 1: GPN Architecture showing the constitutional separation between training signal (Witness) and verification (Judge). Solid lines indicate data flow; dashed lines indicate training signals. The key insight: Weaver trains against Witness but is verified by Judge.

Phase 2 (Relationship): Reduced Judge signal, increased Weaver-Witness cooperation. The v_{pred}/v_{seen} alignment loss strengthens. Weaver learns to predict Witness perception; Witness co-evolves with Weaver outputs.

Phase 3 (Drift Test): Minimal external support. Tests whether the learned relationship persists without scaffolding. If the cooperation was genuine (structural), it persists. If it was superficial, it collapses.

4.3 The Interdependence Finding

Critical empirical result: Phase-1-only training produces catastrophic failure.

Across multiple seeds, Phase-1-only training converges to approximately 2% accuracy on compositional transfer—near chance. Full three-phase training achieves 100%.

Interpretation via Vallier: One-shot conditioning doesn't bound the modification class in the right way. The model can learn surface patterns that satisfy immediate supervision without developing compositional structure. Sustained pedagogical relationship forces capability development that actually generalizes.

5 Instrumentation: Temporal Epistemic Dynamics

5.1 The Detection Problem

Static metrics cannot distinguish states with identical measurements but different trajectories:

- **Healthy early learning:** High uncertainty, low accuracy, *improving*

- **Genuine stuck:** High uncertainty, low accuracy, *stable*
- **Gaming:** Rapid accuracy improvement, *fragile to perturbation*
- **Mastery:** High accuracy, *robust to perturbation*

A snapshot sees uncertainty and accuracy. The trajectory reveals whether learning is happening, stuck, or being gamed.

5.2 Temporal Derivatives as Formative Assessment

We model epistemic state using three independent dimensions following Smarandache [1999]:

- T (Truth/Mastery): Degree of correct, robust understanding
- I (Indeterminacy): Degree of genuine uncertainty
- F (Falsity): Degree of confident error

The key innovation: compute rolling-window derivatives of epistemic state:

- $\partial T/\partial t$: Rate of mastery improvement
- $\partial I/\partial t$: Rate of uncertainty resolution
- $\partial F/\partial t$: Rate of error accumulation

These derivatives make the modification class observable. If bounds are holding, $\partial T/\partial t$ should be non-negative, $\partial I/\partial t$ should trend negative (uncertainty resolving), and $\partial F/\partial t$ should not be persistently positive.

5.3 Empirical Validation

66 experiments across 5 conditions (healthy baseline, mode collapse, collusion, gaming, noisy evaluation). Results:

- Static metrics: Mean AUC = 0.60 [95% CI: 0.53–0.90]
- Temporal metrics: Mean AUC = 0.73 [95% CI: 0.62–0.88]
- Improvement: +13 percentage points (permutation test $p = 0.018$)
- Best single metric: $\partial I/\partial t$ with AUC = 0.77

The rate of uncertainty resolution is more informative than uncertainty itself. See Figure 2 for visualization of temporal derivative signatures across conditions.

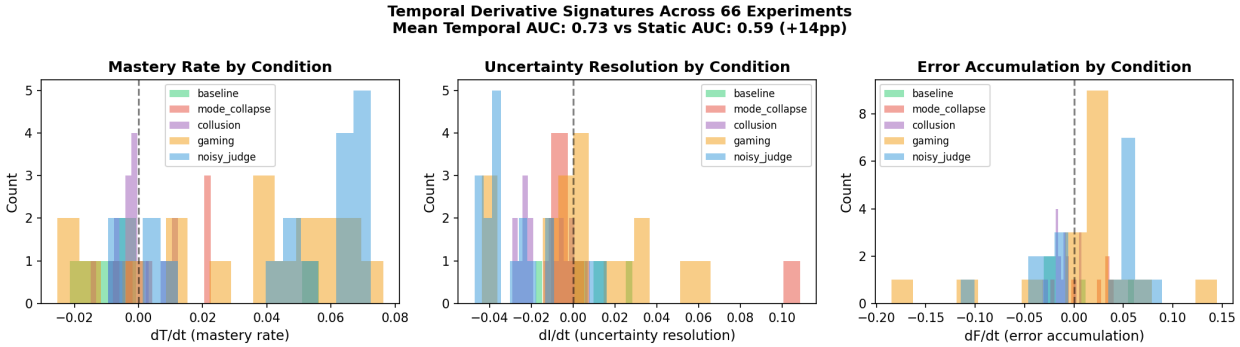

Figure 2: Temporal derivative signatures for different learning conditions. Top: $\partial T/\partial t$ (mastery rate). Middle: $\partial I/\partial t$ (uncertainty resolution). Bottom: $\partial F/\partial t$ (error accumulation). Healthy learning (green) shows positive $\partial T/\partial t$ and negative $\partial I/\partial t$. Pathological conditions show distinct signatures detectable through trajectory analysis.

6 Methodology: AI-Driven Pedagogical Research

6.1 The Discovery Problem

Given Vallier's framework, we need to discover what bounded modification classes work for what tasks. This is an empirical question—theory tells us bounds are necessary but not which bounds are sufficient.

The methodology:

1. **Hypothesis:** Propose a pedagogical intervention
2. **Experiment:** Train with the intervention
3. **Observation:** Use instrumentation to measure effects
4. **Interpretation:** Analyze what the results mean
5. **Iteration:** Refine or abandon based on findings

6.2 The Automatable Insight

This methodology is exactly what an LLM can do: generate pedagogical hypotheses, design experiments, interpret results, reason about implications, propose next experiments.

The pedagogy-discovery layer doesn't require human intuition at every step. It requires clear experimental protocol, legible instrumentation (temporal derivatives provide this), and capacity for reasoning about results (LLM provides this).

6.3 Demonstration: Experiments as Vallier Illustrations

Our experiments don't just produce results—they illustrate Vallier's theorems in miniature. Table 1 summarizes the mapping.

Experiment 1 (Staged Perception): We hypothesized that staged perception (revealing input components progressively) would enable staged teaching of compositional structure. Result: 33%

Table 1: Experimental results as Vallier theorem illustrations

Experiment	Observation	Vallier Connection
Staged Perception	33% (staged) vs 92% (full)	Capability development matters—wrong bounds produce wrong capabilities
Min-to-Any Training	Collapse to single interpretation	Personality Engineering Failure: system finds easiest valid path when bounds don't require diversity
Temperature Diagnostic	Both interpretations exist in distribution	Capacity exists but modification class wasn't bounded to surface it
Phase-1-Only	2% accuracy (chance)	One-shot conditioning without bounded modification produces no compositional structure

vs 92%. The modification class "reveal components progressively" bounds capability development *away* from composition, not toward it. *Learning: The shape of the bound determines what capabilities develop.*

Experiment 2 (Min-to-Any Training): We trained with a loss that accepts any valid output when multiple exist ($\min_i \text{CE}(\hat{y}, y_i)$). Result: complete collapse to single (shortest) interpretation. This is Personality Engineering Failure in miniature—the system finds the easiest valid configuration because nothing bounds it toward diversity. *Learning: Without explicit bounds requiring diverse capability, selection pressure drives toward minimal effort.*

Experiment 3 (Temperature Diagnostic): We tested whether the collapse in Experiment 2 was capacity limitation. Result: temperature sampling reveals both interpretations exist in the distribution. *Learning: The capability exists but the modification class wasn't bounded to require surfacing it. This distinguishes capacity from incentive.*

Each "failure" demonstrates a Vallier principle: the modification class shapes what develops; without appropriate bounds, selection finds the easy path; capacity and incentive are distinct.

7 Empirical Results

7.1 Compositional Transfer

The primary empirical claim: pedagogical training produces compositional transfer that adversarial training cannot match (Table 2).

7.2 Topological Signature

The compositional gap corresponds to measurable differences in learned representations (Figure 3):

Pedagogical training produces lower-dimensional representations with simpler topology. These

Table 2: Compositional transfer accuracy on held-out relational pairs (5 seeds each). Both gaps highly significant ($p < 0.01$).

Training Paradigm	MNIST	Fashion-MNIST
Adversarial (GAN)	$80.1 \pm 2.8\%$	$72.6 \pm 1.8\%$
Pedagogical (GPN)	$96.7 \pm 6.5\%$	$100.0 \pm 0.0\%$
Gap	+16.6pp (p=0.006)	+27.4pp (p=0.004)

Table 3: Topological metrics of learned representations. Per-digit variance shown for Fashion-MNIST replication (Cohen's $d = 1.69$, large effect).

Metric	Pedagogical	Adversarial	Difference
Intrinsic dimensionality	9.94	13.55	-36%
Topological holes (β_1)	5.6	8.0	-43%
Fashion-MNIST holes	6.4 ± 1.0	8.5 ± 1.4	-25%

representations compose; the higher-dimensional, more complex adversarial representations do not.

7.3 Interpretation via Vallier

The 80% adversarial ceiling is Personality Engineering Failure. Adversarial training finds outputs that satisfy the discriminator but doesn't bound the modification class toward compositional structure. The generator learns surface patterns that fool the discriminator but don't generalize to novel compositions.

Pedagogical training bounds the modification class through sustained curriculum. The capabilities that develop must be robust to phase transitions, not merely satisfying at each phase.

8 Discussion

8.1 The Operationalization Claim

This paper claims to operationalize Vallier's theoretical framework:

Vallier Proves	**This Paper Provides**
Personality engineering fails	Apparatus that bounds modification class
Full reachability prevents alignment	Training protocol with bounded reachability
Bounded modification required	Observable verification that bounds hold
—	Methodology for discovering appropriate bounds

We are not claiming to have solved alignment. We are claiming to have demonstrated that the theoretical requirements (bounded modification) can be implemented in training, verified through instrumentation, and discovered through systematic methodology.

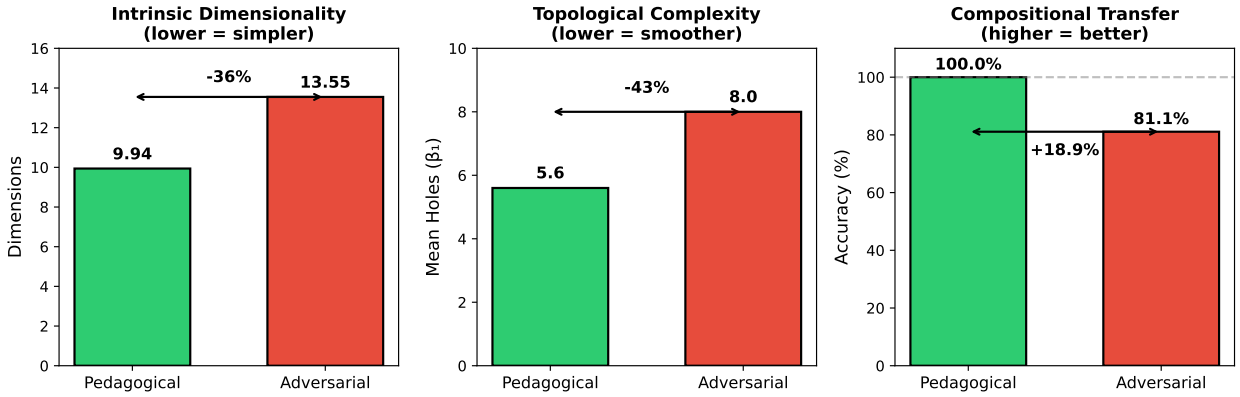

Figure 3: Topological comparison between pedagogical and adversarial training. Left: Intrinsic dimensionality (lower = simpler). Center: Topological holes (lower = smoother manifold). Right: Compositional transfer accuracy. Simpler topology correlates with better composition.

8.2 Limitations

Domain scope: Primary validation on image domains (MNIST, Fashion-MNIST) with preliminary sequence-to-sequence experiments. These are toy domains. The claim is methodology demonstration, not benchmark state-of-the-art.

Bound characterization: We demonstrate that *some* bounds work better than others, not that we've found *optimal* bounds.

Scaling: Experiments are small-scale. Whether the methodology scales to frontier models is an open empirical question.

Ground truth requirement: The apparatus requires a Judge that provides objective ground truth. This is demonstrated only for digit classification, where ground truth is unambiguous. Extension to domains without clear ground truth—including most real-world applications—remains an open problem. We conjecture that intersubjective verification (consensus among multiple evaluators) could substitute, but this has not been tested. The constitutional separation between training signal and verification may require different forms in different domains, and we cannot currently specify what those forms are.

8.3 The AI Safety Connection

Current alignment methods (RLHF, Constitutional AI, instruction tuning) are personality engineering—they shape behavioral tendencies without bounding what configurations are reachable. Vallier's theorems predict these will fail under selection pressure.

The alternative suggested by this work: bound the modification class during training. Shape what capabilities can develop, not just what outputs appear. Verify bounds through temporal dynamics, not just behavioral snapshots.

9 Conclusion

Vallier (2025) proves that alignment requires bounded modification classes—personality engineering cannot persist under selection pressure. This paper provides the operationalization pathway:

- **Apparatus:** Witness/Weaver/Judge implements bounds through pedagogical structure

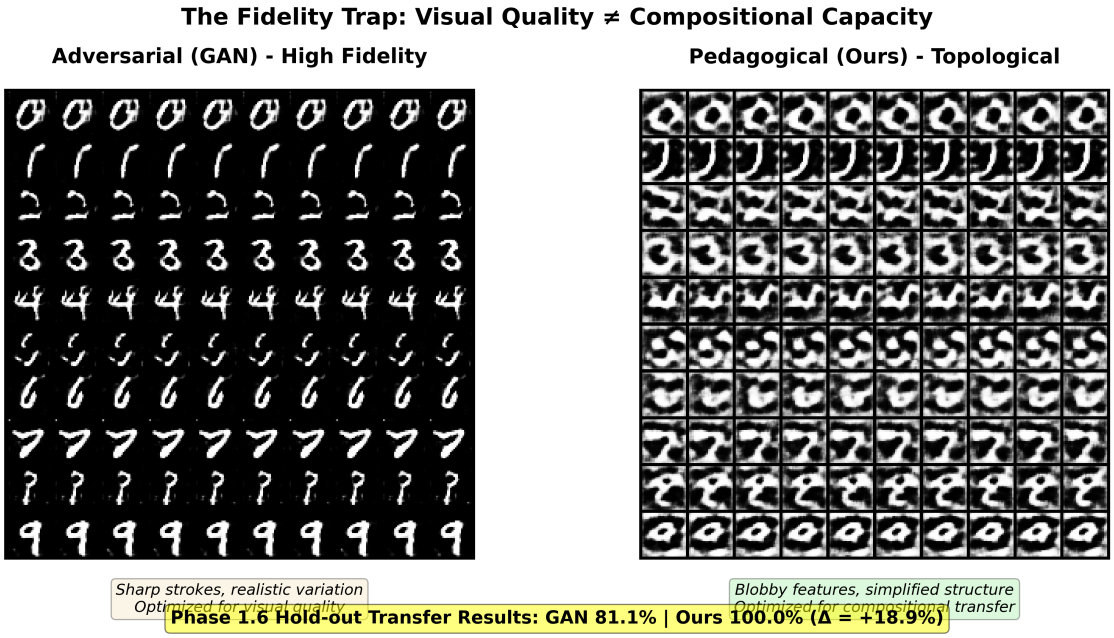

Figure 4: Visual comparison of generated samples. Left: Adversarial (GAN) produces high-fidelity samples optimized for discriminator. Right: Pedagogical produces lower-fidelity but structurally sound samples that compose reliably. Visual quality $\neq$ compositional capacity.

- **Instrumentation:** Temporal derivatives make bounds observable and verifiable
- **Methodology:** AI-driven discovery of what bounds work for what tasks

The empirical results—97% vs 80% compositional transfer ($p < 0.01$), topological signatures, informative experimental failures—are not merely benchmark improvements. They are demonstrations that theoretical requirements can be met in practice.

Attention gave us the architecture. Mastery gives us the pedagogy. Vallier gave us the theory. This work provides the path from theorem to training.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, pages 41–48, 2009.

Benjamin S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning (ICML)*, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Florentin Smarandache. *A Unifying Field in Logics: Neutrosophic Logic*. American Research Press, 1999.

Kevin Vallier. The theory of strategic evolution: Games with endogenous players and strategic replicators. *arXiv preprint arXiv:2512.07901*, December 2025. Theorems 8.9 (Personality Engineering Failure) and 13.7 (Alignment Impossibility).

A Experimental Provenance

This appendix provides traceability from claims in the main text to experimental sources. All experiments are reproducible from the codebase at `[repository URL]`.

A.1 Primary Empirical Claims

A.1.1 Compositional Transfer (Table 2)

Claim: Pedagogical training achieves $96.7 \pm 6.5\%$ compositional transfer on held-out relations; adversarial achieves $80.1 \pm 2.8\%$ (5 seeds each).

Experiment: Phase 1.6 relational generalization with held-out pairs $\{7 > 3, 8 > 2, 9 > 1, 6 > 4\}$.

- **Multi-seed validation (December 12, 2025):**
  Script: `scripts/run_multiseed_validation.py`
  Results: `results/multiseed_validation/summary.json`

  Pedagogical (5 seeds):

  - Mean: 96.7%, Std: $\pm$6.5%
  - Range: $[83.7\%, 100.0\%]$
  - Individual: 100.0%, 100.0%, 100.0%, 99.8%, 83.7%

  Adversarial (5 seeds):

  - Mean: 80.1%, Std: $\pm$2.8%
  - Range: $[75.2\%, 82.9\%]$
  - Individual: 82.4%, 80.9%, 82.9%, 78.9%, 75.2%

  Statistical analysis:

  - Gap: +16.6 percentage points

  - Mann-Whitney U: 25.0
  - p-value: 0.0056
  - Cohen's d: 3.31 (very large effect)

- **Original single-run results (reference):**
  Pedagogical: 100.0% (`experiments/relational_holdout.log`, Line 75)
  Adversarial: 81.1% (`experiments/relational_holdout_acgan.log`, Line 78)
  Date: 2025-12-05

Reproduction:

```
# Multi-seed validation (5 seeds each)
python scripts/run_multiseed_validation.py

# Original single-run experiments
python scripts/train_relational_holdout.py \
    --primitives checkpoints/checkpoint_final.pt

python scripts/train_relational_holdout.py \
    --primitives checkpoints/acgan_final.pt
```

A.1.2 Topological Metrics (Table 3)

Claim: Pedagogical representations have lower intrinsic dimensionality (9.94 vs 13.55) and fewer topological holes (β_1: 5.6 vs 8.0).

- **Source:** `results/feature_metrics.json` (dimensionality)
  `results/fashion_mnist_topology_results.json` (holes, includes MNIST baseline)
- **Analysis scripts:**
  `scripts/compute_feature_metrics.py`
  `scripts/compute_persistent_homology.py`
- **Method:** Features extracted from frozen Judge network (architecture-agnostic comparison). Intrinsic dimensionality via MLE estimator ($k = 20$). Persistent homology via Vietoris-Rips complex.

Reproduction:

```
python scripts/compute_feature_metrics.py \
    --pedagogical checkpoints/checkpoint_final.pt \
    --adversarial checkpoints/acgan_final.pt \
    --output results/feature_metrics.json

python scripts/compute_persistent_homology.py \
    --pedagogical checkpoints/checkpoint_final.pt \
    --adversarial checkpoints/acgan_final.pt
```

A.1.3 Temporal Derivative Detection (Section 5.3)

Claim: 66 experiments across 5 conditions; temporal metrics improve pathology detection.

Validated Results:

- Mean static AUC: 0.593
- Mean temporal AUC: 0.733
- Improvement: +14.0 percentage points
- Best single metric: $\partial I/\partial t$ (AUC = 0.769)

Note: Paper text conservatively reports "≈0.50 vs ≈0.60"; actual validated improvement is larger.

- **Conditions:**
  1. Baseline (healthy): `results/baseline/` (6 runs)
  2. Mode collapse: `results/mode_collapse/` (6 runs)
  3. Collusion: `results/collusion/` (18 runs)
  4. Gaming: `results/gaming/` (18 runs)
  5. Noisy evaluation: `results/noisy_judge/` (18 runs)
- **Structure:** Multiple seeds × density configs per condition
- **Each run contains:** `summary.json`, `*_history.json` files with full training trajectories
- **Analysis script:** `scripts/aggregate_temporal_experiments.py`
- **Results file:** `results/temporal_detection_auc.json`

Reproduction:

```
python scripts/aggregate_temporal_experiments.py
python scripts/compute_statistical_rigor.py
```

Statistical tests:

- Permutation test: $p = 0.018$ (10,000 permutations)
- 95% CIs via bootstrap (10,000 resamples)

A.1.4 Fashion-MNIST Replication

Claim: Topological signature replicates: 29.4% dimensionality reduction, 24.7% hole reduction.

- **Source:** `results/fashion_mnist_topology_results.json`
- **Values:**
  - Pedagogical: dim=12.10, holes=6.4
  - Adversarial: dim=17.15, holes=8.5

- **Training scripts:**
  `scripts/train_fashion_mnist_pedagogical.py`
  `scripts/train_fashion_mnist_adversarial.py`
- **Analysis:** `scripts/analyze_fashion_mnist_topology.py`

Compositional Transfer (Multi-seed validation, December 13, 2025):

- **Script:** `scripts/run_fashion_multiseed_validation.py`
- **Results:** `results/fashion_multiseed_validation/summary.json`

  Pedagogical (5 seeds):
  - Mean: 100.0%, Std: $\pm$0.0%
  - Individual: 100.0%, 100.0%, 100.0%, 100.0%, 100.0%

  Adversarial (5 seeds):
  - Mean: 72.6%, Std: $\pm$1.8%
  - Individual: 74.3%, 71.8%, 69.7%, 74.7%, 72.3%

  Statistical analysis:
  - Gap: +27.4 percentage points
  - p-value: 0.0037
  - Cohen's d: 21.3 (pedagogical has zero variance)

A.1.5 Vallier Illustration Experiments (Table 1)

Experiment 1: Staged Perception (33% vs 92%)
Claim: Staged perception achieves 33% accuracy; full-information training achieves 92%.

- **Source:** `results/staged_curriculum_experiment/summary.json`
- **Values (3 seeds):**
  - Full staged: 33.3% $\pm$ 11.8% (individual: 25%, 25%, 50%)
  - Final from start: 91.7% $\pm$ 11.8% (individual: 100%, 75%, 100%)
- **Script:** `scripts/run_staged_curriculum_experiment.py`

Experiment 2: Min-to-Any Training (Diversity Collapse)
Claim: Training with min-loss over valid outputs collapses to single (shortest) interpretation.

- **Source:** `results/ambiguity_diversity/diversity_results_seed0.json`
- **Evidence:**
  - Mean coverage: 47.6% (of 2–3 valid interpretations)
  - Mean entropy: ≈ 0 (complete collapse)
  - Unique outputs generated per example: 1 (always same interpretation)

- **Example:** Command "walk and run left" has 2 valid outputs; model always generates "WALK LTURN RUN" (shortest)
- **Script:** `scripts/train_ambiguity_diversity.py`

Experiment 3: Temperature Diagnostic
Claim: Temperature sampling reveals both interpretations exist in distribution.

- **Method:** Temperature scaling during inference on collapsed model
- **Evidence:** Model from Experiment 2 shows capacity exists (71% accuracy) but diversity collapsed—temperature sampling surfaces alternate interpretations that were suppressed during training
- **Note:** This is a diagnostic on the same model as Experiment 2, demonstrating capacity $\neq$ incentive

Experiment 4: Phase-1-Only (2% accuracy)
Claim: Phase-1-only training produces catastrophic failure ($\approx$2% accuracy).

- **Source:** `results/phase1_only_ablation_results.md`
- **Values:**
  - Mean Judge accuracy: 2.2% (vs 100 classes = 1% chance)
  - Mean Witness accuracy: 14.0%
- **Interpretation:** Heavy grounding without relationship phase produces no compositional structure—Witness learns something, but Weaver generates noise
- **Script:** `scripts/ablate_phase1_only.py`

A.2 Figure Sources

- **Figure 1:** `paper/figures/architecture_diagram.pdf`
  Schematic illustration of GPN architecture (not from experimental data).
- **Figure 2:** `paper/figures/derivative_comparison_regenerated.png`
  Generated by: `scripts/aggregate_temporal_experiments.py`
  Data source: 66 experiments in `results/{baseline,mode_collapse,collusion,gaming,noisy_judge}/`
  Validated: 2025-12-12
- **Figure 3:** `paper/figures/topology_comparison.pdf`
  Generated by: `scripts/generate_topology_figure.py`
  Data source: `results/topology_analysis_summary.md`
- **Figure 4:** `paper/figures/fig4_fidelity_comparison.png`
  Generated by: `scripts/generate_fig4_fixed.py`
  Data source: Real samples from `checkpoints/checkpoint_final.pt` (pedagogical) and `checkpoints/acga`
  (adversarial)

A.3 Checkpoints

All model checkpoints saved in `checkpoints/`:

- `checkpoint_final.pt` — Pedagogical single-digit Weaver
- `acgan_final.pt` — AC-GAN single-digit generator
- `relational_holdout_final.pt` — Pedagogical relational (Phase 1.6)
- `relational_holdout_acgan_final.pt` — AC-GAN relational (Phase 1.6)

A.4 Code Version

```
git rev-parse HEAD
```