EGWM: Feeling the AGI – Part III
Grow–Compress World Banks and Elegance Annealing

Melissa Howard

December 7, 2025

Abstract

In Part I of this series, we introduced Emotion-Gated World Models (EGWM): a toy architecture where simple "feelings" (loss, mismatch, novelty) decide *when to learn* and *when to skip*, protecting a single world model from junk experience and catastrophic overwriting. In Part II, we extended EGWM to a *world bank*: multiple specialist models, with value-like signals deciding whether to update, spawn new specialists, ignore a phase, or merge models. That work exposed a failure mode: naïve "elegance" objectives can collapse the system into too few models, or stop learning entirely, because the easiest way to be "simple" is not to learn.

In this Part III, we show how to fix that problem with a *staged grow–compress strategy* and a simple *elegance annealing schedule*:

1. First, a *growth policy* creates a rich, redundant world bank by spawning specialists whenever the existing models are confused.
2. Then, a separate *compression phase* repeatedly merges models only if competence stays above a target accuracy.
3. Finally, an *annealed elegance schedule* ramps the strictness of compression over time, allowing redundancy early in "life" and stronger pruning later.

Using only tiny 2D logistic regression models, we show:

- In a stationary 3-world environment, staged grow–compress reduces the number of heads from roughly 3–4 down to about 2 while maintaining near-100% per-world accuracy, and sharply decreases redundancy (overlap) between heads.
- In a non-stationary 4-world environment (worlds appear, disappear, and reappear), a single monolithic model suffers catastrophic interference, while the world bank with grow–compress maintains high accuracy and a bounded number of heads.
- By sweeping the compression target, we trace a competence–elegance trade-off: stricter compression yields fewer heads but risks under-specialization early; looser compression preserves performance but leaves more redundancy.
- An elegance annealing schedule (lenient early, strict late) achieves both: robust adaptation in early phases, and a compact, high-performing world bank at the end.

These experiments are still tiny, but they illustrate a concrete principle: *feelings must have timescales*. Fast confusion signals create new structure; slow, constrained "elegance" signals prune and compress only after competence has stabilized. We argue that this staged, time-structured use of value signals is a plausible ingredient in an AGI-like architecture that both learns continuously and discovers simpler internal structure over its lifetime.

1 Introduction

Parts I and II of EGWM explored a simple idea: instead of training a model blindly on everything, we give it *feelings*—fast scalar signals that estimate "how useful or dangerous is this experience?"—and let those feelings gate learning.

In brief:

- **Part I:** A single model with feelings deciding when to update versus when to freeze, reducing catastrophic forgetting in a simple continual-learning toy.
- **Part II:** A *world bank* of specialist models, with feelings deciding when to route data to an existing specialist, spawn a new specialist (new head), ignore a phase, or merge models.

Part II highlighted a tension:

- We want *elegance*: fewer, more general specialists with minimal redundancy.
- But a naïve elegance reward (e.g., accuracy minus λ times "number of heads" and "overlap") can push the controller toward degenerate solutions: do nothing, do not learn, or collapse into a single mediocre model.

In Part III, we address this by *separating growth and compression*:

1. **Growth:** a simple, local rule creates specialists based on confusion ("if all heads are confused, spawn a new head").
2. **Compression:** an offline or slower process merges specialists only if a *hard competence constraint* is preserved.
3. **Annealing:** the strictness of that constraint changes over time, approximating a developmental schedule ("messy brain early, elegant brain later").

The core question is:

> Can a simple grow–compress schedule, driven by feelings of confusion and constrained elegance, maintain competence in changing worlds while discovering simpler internal structure?

2 Architecture Recap

We keep the EGWM framing but use very simple models so experiments are transparent and reproducible.

2.1 World models (heads)

Each *head* h is a logistic regression classifier:

$$p(y = 1 \mid x; h) = \sigma(w_h^\top x + b_h), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

The setup is:

- Input: $x \in \mathbb{R}^2$.
- Parameters: $w_h \in \mathbb{R}^2$, $b_h \in \mathbb{R}$.
- Loss: logistic cross-entropy on labeled pairs (x, y).

2.2 Value signals and actions

At each time step we see a labeled sample (x, y). For each head h, we compute:

- its loss $\ell_h(x, y)$,
- optionally a simple overlap proxy (how many heads are correct on this point).

In this Part III, we do not fully re-implement the rich value vector from Part II (usefulness, risk, novelty, alignment). Instead, we focus on a *confusion-based growth rule* and a *competence-constrained compression rule*.

Growth rule. At each step:

1. If there are no heads, spawn the first head and train it on (x, y).
2. Otherwise:
   (a) Evaluate $\ell_h(x, y)$ for all heads h.
   (b) Let $\ell_{\min} = \min_h \ell_h(x, y)$.
   (c) If $\ell_{\min} > \tau$ (all heads are confused) and the number of heads is below a maximum:
      - **Spawn** a new head and train it on (x, y).

   (d) Else:
      - **Route** (x, y) to the head with minimum loss and train that head.

Compression rule. Compression operates periodically on the current pool of heads and their accumulated buffers. For each pair of heads (i, j):

1. Train a merged head h_{ij} on the union of their buffers.
2. Form a candidate bank consisting of all heads except i and j, plus the merged head h_{ij}.
3. Evaluate global accuracy A_{cand} on a held-out test set containing all worlds.
4. Accept the merge only if

$$A_{\text{cand}} \geq A_{\text{target}} \quad \text{and} \quad A_{\text{cand}} \geq A_{\text{base}} - \delta,$$

   where A_{base} is the current bank accuracy and δ is a small tolerance.

Merges are applied greedily until no acceptable merge remains.

This gives two time-scales:

- A fast, local **growth controller** driven by confusion.
- A slower, global **compression controller** driven by competence and elegance.

3 Experiment 1: Grow–Compress in a 3-World Stationary Environment

3.1 Environment

We construct three hidden linear worlds in 2D:

- Worlds 0 and 1 share the *same* decision boundary (same w, b), but are centered on different Gaussians in the plane.
- World 2 has a *different* boundary and a different mean.

Each world w generates samples by:

1. drawing $x \sim \mathcal{N}(\mu_w, \sigma^2 I)$,
2. labeling via $y = \mathbf{1}[w_w^\top x + b_w \geq 0]$.

We interleave phases from each world and feed the mixture to the system.

3.2 Growth phase

Using the confusion-based growth rule over multiple episodes and phases:

- The system starts with no heads.
- It spawns a new head whenever all existing heads are confused on a sample.
- Otherwise it routes to the lowest-loss head and updates it.

A typical run:

- Ends with roughly 3–4 heads.
- Achieves near-100% per-world accuracy under "oracle routing" (choose the head with lowest loss per sample).
- Shows substantial overlap: multiple heads are often correct on the same point, with an overlap metric (average extra correct heads per sample) around 0.2–0.8.

3.3 Compression phase

Starting from this grown bank, we run the competence-constrained compression described above, using a target accuracy A_{target} (e.g., 0.99) and a small tolerance δ.

In a typical run:

- Growth: 3 heads, overall accuracy ≈ 1.0, overlap ≈ 0.2.
- After compression:
  - 2 heads, overall accuracy still ≈ 1.0.
  - Per-world accuracy remains near 1.0.
  - Overlap drops close to 0, i.e., almost no redundant correct heads.

Because worlds 0 and 1 share the same decision boundary, the compressor learns to merge their specialists into a single head, while keeping a separate head for world 2.

3.4 Takeaways

1. **Growth discovers structure; compression removes redundancy.** Growth creates multiple specialists, some overlapping; compression distills them into a smaller set with the same competence.

2. **Elegance is constrained.** Because compression is only allowed when accuracy is preserved, we avoid the collapse seen in Part II when elegance was optimized directly.

3. **World sharing emerges naturally.** When two worlds share a rule, a single head can cover both; compression finds and merges these cases automatically.

4 Experiment 2: Non-Stationary 4-World Environment

We now test the mechanism in a more realistic continual-learning setting.

4.1 Environment and schedule

We define four worlds:

- Worlds 0 and 1: same boundary, different regions.

- World 2: different boundary and mean.

- World 3: another distinct boundary and mean.

Training is organized into segments:

1. Segment 0: W0 only (world 0 samples).

2. Segment 1: W1 only.

3. Segment 2: W2 only.

4. Segment 3: W0 again (world 0 reappears after a gap).

5. Segment 4: W3 new (completely unseen world).

6. Segment 5: Mixed all (any of worlds 0–3 at each step).

At the end of each segment we evaluate on held-out test sets for each world.

4.2 Methods compared

We compare:

1. **Monolith:** a single logistic regression trained online on the full stream.

2. **Growth-only bank:** confusion-based growth, no compression.

3. **Grow–compress bank:** confusion-based growth plus compression at the end of each segment with a fixed target accuracy A_{target}.

4.3 Monolith: catastrophic interference

The monolithic model behaves like a standard continual learner under domain shifts:

- After W0 only, accuracy on world 0 is decent, but it has never seen the other worlds.
- After W1 only, parameters shift toward world 1, and performance on world 0 deteriorates.
- After W2 only, the model tries to satisfy all three worlds with one linear boundary; accuracy on individual worlds can drop sharply.
- After W3 new, adapting to the new world further interferes with older ones.
- Only after a long Mixed all segment does it settle into a compromise boundary with moderate accuracy.

In short, a single model struggles: new worlds overwrite old structure, and incompatible worlds are hard to represent.

4.4 Growth-only world bank: strong memory, redundant structure

The growth-only bank:

- Spawns a new head when all existing heads are confused.
- Converges to roughly 4 heads, one dominant specialist per world.
- Under oracle routing, maintains very high accuracy (often near 1.0) on each world, even when worlds disappear and reappear.

However:

- Overlap is high: multiple heads often agree on the same point.
- The bank is effective but unelegant and redundant.

4.5 Grow–compress bank: bounded capacity, preserved competence

The grow–compress bank:

- Grows to 3–4 heads in early segments.
- Maintains high per-world accuracy (typically ≥ 0.9) across segments.
- Has lower overlap and a more compact structure than the growth-only bank.

When new worlds appear:

- Growth spawns new specialists for genuinely novel regimes.
- Compression merges redundant heads when this can be done without violating accuracy constraints.

In the final Mixed all segment, the bank typically stabilizes at about 3–4 heads, with near-1.0 accuracy and moderate overlap. Unlike the monolith, it:

- Remembers old worlds when they return.
- Detects and isolates new worlds.
- Keeps the number of heads bounded.

5 Experiment 3: Compression Aggressiveness and Trade-Offs

We next vary the compression strictness by changing the target accuracy A_{target}:

- Strict: $A_{\text{target}} = 0.99$.
- Medium: $A_{\text{target}} \approx 0.97$.
- Loose: $A_{\text{target}} \approx 0.95$.

All experiments reuse the same non-stationary 4-world environment.

5.1 Observations

Strict compression ($A_{\textbf{target}} = 0.99$).

- The controller is reluctant to merge, especially early, because any small drop in accuracy is forbidden.
- It can keep more heads for longer, or become under-specialized on new worlds until enough data accumulates.

Medium compression ($A_{\textbf{target}} \approx 0.97$).

- Provides a good compromise: high competence in all segments, moderate head count (3–4), and reduced overlap.

Loose compression ($A_{\textbf{target}} \approx 0.95$).

- In this toy setup, does not drastically change the final number of heads compared to the medium setting: once the bank has about three well-formed specialists, further merging would genuinely hurt accuracy, so compression naturally stops.

Overall, we see a competence–elegance frontier: reducing A_{target} can yield more compression, but only until it starts destroying real structure.

6 Experiment 4: Elegance Annealing

Finally, we test a time-varying competence target, mimicking development:

- Early segments: allow more slack (e.g., 0.95, 0.96).
- Mid segments: gradually tighten (0.97–0.98).
- Late segments: require near-perfect competence (0.99).

For six segments, we use:

$$A_{\text{target}}^{(0..5)} = [0.95,\ 0.96,\ 0.97,\ 0.98,\ 0.99,\ 0.99].$$

Compression at the end of segment s uses the segment-specific target $A_{\text{target}}^{(s)}$.

6.1 Results

In a typical grow–compress run with this schedule:

- Early segments (W0 only, W1 only): 2–3 heads, overall accuracy ≈ 0.92–0.94, moderate overlap.
- Middle segments (W2 only, W0 again): 3 heads, accuracy ≈ 0.90–0.91, slightly higher overlap while multiple hypotheses coexist.
- Late segments (W3 new, Mixed all): 3 heads, overall accuracy ≈ 1.0, and a structured pattern of overlap (e.g., one head covering both worlds 0 and 1).

The bank ends with three heads and nearly perfect accuracy on all worlds. This is the same compact structure observed under strict compression, but annealing makes the path to that structure more stable and forgiving early on.

6.2 Interpretation

Elegance annealing suggests:

- Early in learning, redundancy is allowed; overlapping specialists can coexist. This supports exploration and protects against premature merging.
- Later, the system demands that compressive changes preserve almost perfect performance, pushing the bank toward an elegant minimal representation.

This parallels biological development, where brains overproduce synapses and then prune aggressively once experience clarifies which patterns matter.

7 Discussion

Across these experiments, a consistent pattern appears:

1. **Confusion-driven growth** is a simple, powerful primitive: if all current models are confused on an experience, that is a clean signal that the current hypotheses are insufficient, and spawning a new head becomes an appropriate structural change.
2. **Competence-constrained compression** avoids elegance collapse: rather than optimizing a single scalar "elegance" objective, we treat competence as an inviolable constraint, avoiding degenerate "do nothing" solutions.
3. **Time-structured feelings matter:** fast feelings (loss, mismatch) act locally and immediately, while slow, global feelings (overall accuracy, redundancy) act through offline or periodic compression. An annealed schedule lets the system be messy first, then elegant later.
4. **World structure emerges from overlap and compression:** when two worlds share underlying rules, the bank first learns multiple overlapping specialists, then discovers they can be safely merged. When worlds are truly different, merging them hurts accuracy and is rejected.

Overall, Part III shows that EGWM can be extended from "protect learning" to "discover simpler structure over time" by separating growth and compression and by respecting competence as a hard constraint.

8 Limitations and Future Work

These experiments are intentionally small and idealized:

- The models are tiny (logistic regression in 2D), with no high-dimensional perception or long-horizon planning.
- Evaluation often uses oracle routing, which isolates the world bank quality but assumes a perfect gating policy at inference time.
- Growth and compression are hand-coded heuristics, not learned controllers.

Promising next steps include:

- **Learned gating and meta-control:** replacing hand-designed rules with a learned meta-policy that receives value-like features (loss, disagreement, overlap, head age, etc.) and outputs structural actions (spawn, merge, freeze, update).
- **Beyond linear worlds:** testing EGWM grow–compress in higher-dimensional tasks, with non-linear models and richer forms of concept drift.
- **Integration with the full EGWM stack:** combining the world bank with planner, tool-use module, multi-channel value/emotion module, and long-term memory updater.
- **Real-world continual benchmarks:** applying similar ideas to standard continual learning and domain shift benchmarks where catastrophic forgetting is known to be an issue.

9 Conclusion

Part III of EGWM shows that feeling-guided structure is not just about protecting a single model from forgetting. With a staged grow–compress scheme and a simple elegance annealing schedule, even tiny toy systems can:

- discover multiple worlds,
- maintain competence as environments change over time,
- and gradually compress themselves into a smaller, more elegant internal structure without collapsing.

In the spirit of "Feeling the AGI", these experiments suggest a broader theme: intelligence might emerge not from one giant static model, but from a living system of world models whose growth and simplification is continuously guided by feelings of confusion, competence, and elegance over multiple time-scales.

Future work (Part IV) will explore making these controllers learned and embedding them in more realistic architectures, pushing EGWM from toy worlds toward systems that can inhabit open-ended reality.