Successors should earn authority, not win contests: graded succession in continual learning

W. Schulz

Nov 11, 2026

Abstract

A literature synthesis and design proposal that reframes generational continual learning: a new generation of models or knowledge should not compete with the old one but take over from it by demonstrated competence, gradually and per context, until its authority is dominant, while a protected constitution changes only by explicit amendment. It assembles anytime-valid statistics, fixed-share mixtures, belief revision, truth discovery and staged deployment into one design, and proposes five pre-registrable tests. None has been run yet.

W. Schulz, NeuroGenesys, November 2026. A literature synthesis and design proposal, not an experiment. Cited numbers are the sources' own; the lab's results carry their ledger IDs, and in-lab toy studies outside the ledger (E3, E4, the QI succession script) are labelled. Experiments S1–S5 are proposals with placeholder thresholds; none has been run.

NeuroGenesys should drop "competition" from the GCCL proposal and adopt graded succession by earned authority as its name and organising idea. A successor generation starts with no authority. It gains authority in steps, per context, as an independent evaluator accumulates evidence that deferring to it beats deferring to the incumbent. It loses authority when that evidence turns. It becomes dominant only when the evidence is decisive. It never gains power over a protected constitution, which changes only through an explicit, recorded amendment. Every component already exists somewhere. Industrial ML deployment runs shadow, canary and ramped traffic shifts with automatic rollback. Anytime-valid statistics (e-processes, confidence sequences) can turn "proven competence" into an error-controlled ladder. Fixed-share mixtures hand weight over gradually, where Bayesian averaging and Hedge collapse to one winner and are slow to switch. Belief-revision theory separates admitting new evidence from how firmly it is held, and has formal operators that make a core set immune. Nobody has yet assembled these into a model-succession mechanism, and nothing in the lab has tested one. The lab's own record supports the parts: self-confidence is a poor competence signal, stale majorities win votes, stiff anchors keep stale facts, missing anchors let beliefs drift, and selector error grows with population size. The specific claim that graded succession beats both a hard switch and plain averaging is the lab's design hypothesis. Experiments S1–S5 below are proposals to test it, not results.

"Competition" names the mechanisms the lab's record says fail

The original GCCL proposal pictured old and new models competing. Each would earn influence through a context-dependent competence profile, be selected by a softmax over competence, and be retired when redundant (GCCL lab record). The lab's results argue against nearly every mechanism the word "competition" invites.

Winning by your own confidence fails. In a QI-era succession toy pre-registered from a nearly identical user proposal, filtering what the predecessor handed on by its own confidence dropped accuracy from 0.759 to 0.725, while naive succession held at 0.761. The script's own verdict was "self-confidence filtering is worse than no filtering at all" (exp_succession.py, not in the ledger). In-domain experts are calibrated only on their own data (own-task ECE 0.011 vs 0.697 on mixed inputs, ledger 9.2). A shared temperature left selector error unchanged (0.367 → 0.368), so tuning a softmax's temperature cannot repair a competence signal built on confidence (lib/findings.ts).

Winning by numbers fails. In the lab's iterated-closure experiment E3 (toy, not in the ledger), unanchored self-consistency reached 100% consistency with 44.0% of updated persons coherently wrong. At round 0 the two templates that never saw an update outvoted the updated one (E3 results). In a competitive population, the parent plus its older siblings is exactly such a stale majority.

Large fields crowd out newcomers. On CIFAR-100, max-softmax selector error ε rose 0.534 → 0.644 → 0.788 at $T = 5/10/20$, while within-task accuracy rose. The whole accuracy loss was selection (ledger 8.18, confirmed, Holm-adjusted $p < 0.0001$) (lib/findings.ts).

Contests get gamed. One verifier's exploitability ranged 6.2× (0.153 to 0.949) depending only on the candidate pool it was handed (exploitability paper). The real-judge test, however, found no Best-of-N decline by $N = 32$ (8.13). Plurality over the top half beat argmax in 6/6 math cells, by +0.03 to +0.17 (8.14) (lib/findings.ts). A softer rule for conferring authority beat winner-take-all there, even without exploitation.

"Competition" also blurs three different questions: who is right, who acts, and who makes the rules. Philosophy separates theoretical authority (a reason to believe) from practical authority (a reason to act). Raz's service conception holds that authority is legitimate when deferring to it helps you follow reasons that already apply to you better than reasoning alone (SEP, Authority). Hart separates primary rules from the secondary rules of change that alter them (SEP, Legal Positivism). These are conceptual tools, not evidence about models. They do, however, locate the design's central commitment: competence earns authority under the primary rules (answering, predicting, acting) and never under the rule of change.

Recommended terminology

The lab should rename the programme graded succession by earned authority (short form: *earned succession*) and use this vocabulary consistently in papers, code and the ledger:

Replace	With	Meaning in the design
competition between generations	succession	a directed handover from incumbent to successor, not a symmetric contest
fitness / winning	earned authority $a_t(c)$	the share of decisions in context c the successor holds at window t, set by accumulated independent evidence
competitor / challenger	successor, in a named office (shadow → probation → co-authority → dominant)	authority is held by role, with entry criteria, not by the model itself

Replace	With	Meaning in the design
loser / extinct model	incumbent → emeritus generation	kept recoverable at a residual floor until sunset, then retired
selection pressure	evidence ladder	thresholds on an e-process that unlock authority levels
natural selection of rules	constitution with amendment	protected items no authority can override; changed only by an explicit, versioned amendment
survival of the fittest knowledge	supersession	an explicit, justified, logged replacement of a specific item

"Dominance" stays, with a working definition: a ≥ 0.9 held for m consecutive windows. That definition comes from the lab's notes and is a design choice, not an empirical threshold (succession analogues note).

Accumulated evidence, not current scores, should set the dial

The theory behind graded succession comes from three literatures. Together they separate a *gradient* from both a switch and an average.

E-processes turn "proven competence" into an error-controlled ladder

Fixed-sample tests become invalid if you keep looking and stop when the result looks good. Johari, Pekelis and Walsh call p-values "wholly unreliable" under continuous monitoring. They built always-valid p-values for a large commercial A/B platform (arXiv 1512.04922). Time-uniform confidence sequences give intervals that are valid at every sample size at once (Howard et al., Annals of Statistics 2021). E-processes, built on test martingales that read as a bettor's wealth, stay valid "at all stopping times" (Ramdas et al.).

The design consequence is the core of earned authority, and it is the lab's proposal. Run a paired e-process per context slice for H_0 "the successor is no better than the incumbent". Shadow mode supplies outcomes on identical inputs, which cuts variance. Each time the e-process first crosses $1/\alpha_k$, raise authority to level w_k. Ville's inequality bounds the chance of ever wrongly reaching rung k by α_k. An illustrative ladder from the notes is $\alpha = 0.2, 0.05, 0.01, 0.001$ mapping to w = 5%, 25%, 60%, 100%. A second e-process, for "the successor is worse", triggers demotion. Authority then tracks *accumulated* evidence, not the current point estimate, so a noisy early lead cannot buy authority and one bad batch cannot erase a long record. No ML-deployment or LLM-routing paper was found that uses e-processes to schedule a traffic ramp rather than to stop a test. This is an open application, not an established method.

Bayesian averaging and Hedge collapse; fixed share hands over

The obvious "evidence-weighted" mixtures are actually winner-take-all in the limit. Yao, Vehtari, Simpson and Gelman show that when no candidate model is true, Bayesian model averaging "will asymptotically select the one single model in the list that is closest in KL divergence". Stacking, which optimises

combination weights for leave-one-out predictive score, keeps genuine mixtures when models complement each other (Yao et al., Bayesian Analysis 2018). Van Erven, Grünwald and de Rooij named the catch-up phenomenon: Bayesian mixtures keep predicting with an older model long after a newer one has become better. Their switch distribution fixes this (arXiv 0807.1005).

Herbster and Warmuth's fixed share redistributes a fraction α of weight each round. Its regret is measured against the best *sequence* of experts, not the best single expert (Koolen & van Erven; Herbster & Warmuth 1998). The share rate does two jobs that succession needs. It caps entrenchment, so a better newcomer overtakes within about $\log(1/\alpha)$ divided by its per-round advantage, not the incumbent's whole history. It also keeps every superseded expert above a small floor, so rollback stays possible. One family covers all three regimes: a learning rate near zero is plain averaging, a very high one is a hard switch to the current leader, and the middle is graded. That makes it the natural controlled baseline.

Fixed share alone is not *earned* authority, because its floor is unconditional. The proposed design pairs a small fixed-share floor (exploration, recoverability) with the e-process ladder (earned increments). Stacking handles the separate question of complementarity, where the incumbent stays better on some slices.

Belief revision supplies admission, firmness and an immune core

Classical AGM revision gives new input absolute priority through its Success postulate. The rest of the field softens it (SEP, Logic of Belief Revision). Two dials matter here.

Admission decides whether an input is accepted at all. Credibility-limited revision accepts only inputs in a credible set (Hansson et al., JSL 2001). Two-level variants set separate thresholds (Garapa 2022).

Firmness decides how strongly an accepted input is held. Spohn's ranking-function conditionalization takes a firmness parameter, so acceptance comes in degrees (Spohn 1988). Jeffrey conditionalization shifts probabilities over a partition without making any cell certain. Successive Jeffrey updates generally do not commute, so the order of evidence can change the end state (Wagner). A graded-succession system should therefore log update order and test for path dependence.

Screened revision (Makinson) makes a core set immune: an input inconsistent with the core is refused, not absorbed (SEP §6). That is the formal shape of the protected constitution.

What belief revision lacks is a way to *learn* credibility. Truth discovery fills that gap. TruthFinder estimates source trustworthiness and fact truth jointly, by iteration (Yin, Han & Yu, KDD 2007). Google's Knowledge-Based Trust scored 119M web pages by the correctness of the facts they state (Dong et al., PVLDB 2015). Dong, Berti-Équille and Srivastava show that copying between sources makes naive majority and naive recency both fail (PVLDB 2009). The lesson transfers directly: a successor distilled from its parent is a copier, so their agreement is not independent evidence, and authority must not accrue to it. Recency becomes a learned per-relation prior, never a rule that the newest value wins.

These external mechanisms are needed because models weigh evidence by the wrong signals. Xie et al. found that ChatGPT's reliance on its own memory climbs from 3.7% to 99.8% as the share of supporting passages rises from 0/2 to 2/2. Acceptance is graded, but by proportion (ICLR 2024). Schuster, Gautam and Markert found a consistent credibility ordering across 13 open models, with government above newspapers above individuals. Repetition overturned it: two social-media tables beat one government table, with an average preference gap of 33.90 toward social media (ACL 2026). If a successor's outputs are fed back as data, sheer volume will out-vote a smaller but more accurate incumbent.

Authority therefore has to be assigned outside the model, from de-duplicated, independently verified outcomes.

Deployment practice already grades authority; human analogues explain why it matters

Engineering precedent: shadow, canary, ramp, rollback

This is a real practice, not an analogy, and it is the closest working template. AWS SageMaker shifts traffic all at once, in two canary steps, or in linear equal steps. Each step has a bake period. Any tripped alarm rolls all traffic back to the old fleet, and the old fleet is terminated only after the new one has run smoothly for a set period (AWS blue/green). Shadow tests send a copy of live requests to the new variant and return only production responses (AWS shadow tests). Google's SRE Workbook sizes canaries so that a 20% canary error rate costs about 1% overall at 5% traffic. It recommends at most a dozen metrics and one canary at a time (Google SRE Workbook). Kayenta scores each metric with a Mann–Whitney test against tolerance bands, and a critical-metric breach forces the score to zero (Spinnaker canary judge). Uber reports shadow coverage of more than 75% of critical online use cases. Its deployments start on a small traffic slice and roll back automatically to the last known good version (Uber).

The mapping onto succession is direct: shadow is apprenticeship with zero authority, canary is probation under an error budget, ramp steps are earned increments, the final bake plus decommissioning is completed succession, and rollback is authority returning to an incumbent that must stay servable. Two gaps define what the lab would add.

- Industry ramps advance on the clock. A step proceeds after a fixed bake unless an alarm fires.
- Industry gates test non-inferiority, not superiority. They ask "is it not worse?" rather than "has it proven better?". No primary source publishes numeric ramp schedules or pass thresholds for ML models.

Evidence-gated, superiority-tested, per-context ramps are the lab's extension, not documented practice.

Analogies, labelled as such

The following human, scientific and legal cases motivate hypotheses. They prove nothing about models.

Science (analogy). Kuhn described individual conversion as a gestalt switch (SEP, Kuhn). At the level of whole fields, the best causal evidence shows gradual change. After 452 eminent life scientists died prematurely, non-collaborators' output in their subfields rose 8.6%, in "a slow but steady increase", and offset collaborators' decline by about year five. Entry was lowest under the most eminent stars ("Goliath's shadow") (Azoulay, Fons-Rosen & Graff Zivin, AER 2019). Chu and Evans found that in very large fields, citations concentrate on papers already cited, which ossifies the canon (PNAS 2021). Two structural lessons carry over:

- Smooth curves come from many hard switches. Population-level authority can be graded even when each individual decision is a hard pick, as every per-request canary is.
- A prestigious incumbent can starve a better successor. Science's remedy was exogenous: death. The engineered remedy is a probe budget reserved independently of current authority.

Lakatos supplies a template for the constitution: a "hard core" shielded by methodological decision, and a "protective belt" revised by evidence (SEP, Lakatos).

Law (analogy). Stare decisis is graded supersession with a raised bar. *Dobbs* restates five factors for overruling: the nature of the error, the quality of reasoning, workability, effect on other law, and reliance

(Cornell LII). *Janus* overruled *Abood* after earlier cases had criticised it without overruling it (opinion). Friedman warns that "stealth overruling" hides when authority actually moved (Geo. L.J. 2010). The design lessons:

- A supersession checklist that asks more than "higher score".
- An explicit criticise → distinguish → narrow → overrule sequence, recorded at each stage rather than left implicit in routing statistics.

There is one crucial disanalogy. Courts overrule precedent through ordinary adjudication, but constitutional text changes only through amendment. The design should treat knowledge like precedent and the constitution like entrenched text.

Organisations (analogy). In 204 US manufacturing successions, relay succession, where an heir apparent first serves as understudy, was followed by better performance (Zhang & Rajagopalan 2004, abstract-level). Weber's account of charismatic authority names succession as its weak point. His remedy, routinisation, moves authority into an office governed by rules (SEP, Weber). This supports placing authority in *offices with entry criteria* rather than in particular models.

Ecology (analogy). Connell and Slatyer's *inhibition* mechanism is the failure mode: early occupants hold a site until a disturbance removes them (summary).

The constitution sits above authority, and only amendment moves it

Frontier specifications already rank authority

Both frontier labs publish layered specifications. OpenAI's Model Spec (current version 2026-10-29) has six levels. At the top, Root principles "cannot be overridden by conversational context". Assistant and tool messages have no authority by default unless a higher level delegates it. The model should not let "its own previous messages" reinterpret higher principles (Model Spec). The changelog shows OpenAI restructuring the hierarchy itself in v2025.09.12, when it set Root above System (CHANGELOG). Anthropic's constitution sets seven hard constraints. It forbids influencing its own training or adjusting its own values without sanction. It describes a "disposition dial" from fully corrigible to fully autonomous, to be moved as trust grows (Claude's constitution). That is, in substance, an official statement of authority conferred in degrees, with the dial held by the issuer rather than the model.

Two facts limit the template:

- Neither lab publishes an amendment procedure. Both amend unilaterally, documented by version history.
- Trained-in hierarchies are imperfect. Wallace et al.'s instruction-hierarchy training improved robustness by up to 63% and held-out attacks by up to 34%, not 100% (arXiv 2404.13208).

A protected constitution should therefore be enforced architecturally, by a decision-time check outside the model, and not only by training.

Legal entrenchment suggests three tiers (analogy)

The German Basic Law requires that amendments "expressly" change the text (Art. 79(1)). It requires two-thirds of both chambers (Art. 79(2)), and puts Articles 1 and 20 beyond amendment (Art. 79(3)) (Basic Law). Elkins, Ginsburg and Melton warn that rigidity has costs. Median national constitutional lifespan is about 19 years, and constitutions too rigid to adapt get replaced (Cambridge). Ginsburg and Melton find that "amendment culture" predicts amendment rates better than formal rules (ABF).

By analogy, GCCL's constitution should have three tiers:

- Eternity tier: small, unamendable through the ordinary channel. Hard constraints and oversight belong here.
- Entrenched tier: amendable only by an explicit, human-gated, versioned text change.
- Ordinary tier: domain knowledge and defaults, where succession operates freely.

Two rules follow from the analogy. The "express amendment" rule should be copied literally: no amendment by implication, and no drift accumulated through a successor's behaviour. And the eternity tier should stay minimal, so that operators are not pushed to fork the specification.

The lab's evidence: anchor to verified evidence, overwrite on amendment

The lab's evidence brackets how strongly old knowledge should be protected.

- Too stiff fails. In-lab experiment E4 (toy knowledge-base transformer) used full EWC at $\lambda = 1e5$. Edit success fell from 62.5% to 13.5% (± 5.6, $n = 5$), and stale-paraphrase rate rose from 10.8% to 81.7% (± 15.1) (E4 results). Ledger 4.5 found that KL anchoring to a frozen parent on new-fact inputs blocks learning (5% immediate accuracy). Over 100 stages, output anchoring alone fell to 20.4% base-fact retention (6.1) (lib/findings.ts).
- Too loose fails. E3 found 44.0% coherently wrong without an anchor, against 0.0% with one (Δ -44.0 $[-53.7, -34.3]$). Its post-hoc sweep concluded that "the overwrite step, not the extra anchor loss, is what makes small evidence win" (E3 results).
- Overwriting works and lasts. A rank-one overwrite plus a pruned pool cut stale 2-hop answers from 93.0% to 9.5% (7.1, L1; the edit site was tuned on seed 0, post hoc). Stale answers stayed at or below 1.5% after 50 steps of unrelated fine-tuning (7.2) (lib/findings.ts).
- Rehearsing superseded facts keeps them alive. Forcing superseded facts into rehearsal gave 23.0% update accuracy, against 73.5% when they were pruned (3.10) (lib/findings.ts).
- Anchoring to a frozen base works when it does not block updates. Token KL to the frozen base tied gold replay (92.0% vs 91.8%) and protected never-rehearsed facts better (+5.4 pp) (3.13) (lib/findings.ts).

The synthesis, as a design inference:

- Anchor to verified evidence, not to the incumbent's outputs. Anchoring to outputs preserves the incumbent's errors.
- Make ordinary-tier anchors item-wise and evidence-gated. An anchor is released when contradicting verified evidence for that item crosses a threshold.
- Make constitutional items hard checks with overwrite-on-amendment. An amendment then takes effect at once, not through a gradual transfer of authority.

The evaluator must be independent of both generations

Who judges competence has to be fixed before succession starts. Weak-to-strong generalisation shows that a weaker supervisor recovers only part of a stronger model's ability (Burns et al.), so the incumbent may under-credit a better successor. A successor that judges itself can credit itself. The lab found that weighting whole sources by reliability *backfired* (attack $0.468 \rightarrow 0.696$) and that per-example weighting fixed it (0.935, post hoc) (2.3) (lib/findings.ts). Authority should therefore be estimated per context from held-out outcomes by an evaluator independent of both generations, with its integrity protected like a hard constraint. Competence never moves the corrigibility dial.

A concrete design: offices, a ladder, item-wise anchors and a screened core

The proposal below is the lab's design. Each component is tagged with how much of it rests on evidence.

Component	Mechanism	Status
Offices	Shadow (0 authority, paired outcomes logged) → probation ($\leq$5%) → co-authority → dominant (a $\geq$ 0.9 for m windows) → incumbent becomes emeritus, kept at a fixed-share floor until sunset	Practice-based (AWS, SRE, Uber); offices motivated by Weber analogy
Evidence ladder	Per-context paired e-process; crossing 1/α_k unlocks w_k; a second e-process for "worse" demotes; restart or discount after detected change points	Statistics established; application to ramps is a proposal
Independent evaluator	Held-out probes, execution or KB checks, outcome logs; reserved probe budget independent of current authority	Supported by lab record (QI succession, 9.1–9.4, 2.3); reserved budget a proposal (tested by S2)
Recoverability	Fixed-share floor on superseded generations; non-inferiority alarm triggers full rollback	Theory (Herbster–Warmuth) plus practice
Reliability layer	Truth-discovery-style per-domain reliability; copy-discount agreement with ancestors; de-duplicate repetition; recency as a learned per-relation prior	Literature-supported; integration is a proposal
Item-wise anchors	Each ordinary-tier item's anchor strength ϰ_i set by verified support minus contradiction; anchor to the record, not to the parent's outputs	Bracketed by E3 and E4; per-item gating unpublished
Justified propagation	Supersession retracts a derived fact only when *all* its justifications pass through the superseded premise; claims labelled by supporting generation (ATMS-style), and transition complete when no live query depends only on the old generation	Theory (JTMS/ATMS); whether CCU's pruning is "all paths" or "any path" is unverified
Supersession record	Explicit criticise → distinguish → narrow → overrule entries; update order logged	Legal analogy plus Jeffrey non-commutativity

Component	Mechanism	Status
Constitution	Three tiers; decision-time hard check outside the model; express, versioned amendments that take effect immediately via overwrite; competence never moves corrigibility	Spec practice plus legal analogy; E3/7.1–7.2 support overwrite
Retirement and cost	Population cap; report forward passes, labels and bytes	8.18 supports capping; per-request hard picks keep inference at one pass

The per-item κ generalises a set of familiar dials. KL-RLHF's β, the pretraining-mix weight γ (InstructGPT), EWC's λ, Learning-without-Forgetting's distillation weight (Li & Hoiem), mean-teacher EMA decay (Tarvainen & Valpola) and WiSE-FT's interpolation coefficient (Wortsman et al.) all set how much authority the old model keeps. In published use they are fixed hyperparameters. PPO's adaptive KL penalty adjusts online, but only to step size (arXiv 1707.06347). No paper was found that schedules any of these as a function of the successor's measured competence, so this is where the design is newest and least supported. Ledger 7.6 warns against the averaging alternative: halving a LoRA update restored old facts from 33.5% to 83.0% but cost 24.4 pp of update accuracy. Weight-space interpolation is a fixed trade-off, not succession (lib/findings.ts).

Five pre-registrable experiments separate gradient from switch and average

All five are proposals, not results. They replace G1, G2, G6 and G8 in their succession-curve versions, and G3 and G7 in their real-judge version. G4 and G5 stand, with G5's record-inheritance arms feeding S4 (succession analogues note).

What gets measured. Each run logs authority a_t(c) per window. The primary measure is the *conflict share*: on queries where the two generations disagree, the fraction settled in the successor's favour. Agreement-set routing is uninformative, so it is excluded. Each trajectory is fitted with constant, step and logistic models and compared on held-out windows. "Graded" requires the logistic to win by a pre-set margin, with a 10%→90% width of at least two windows.

Shape alone is insufficient. The decisive signature is *selectivity* across planted successor regimes: better (B), equal (E), worse (W), better in part of the contexts (P), and better-then-degrading (D). The predictions for each arm:

- Earned authority rises logistically in B, stays low in E and W, rises only in the winning contexts in P, and rolls back in D.
- A hard switch steps to 1 in B and false-promotes in E and W.
- Averaging stays flat in every regime.

Summary statistics are false-promotion rate, context-selectivity gap, transition regret (split into under- and over-delegation), time to dominance, rollback latency, constitutional-violation rate and amendment latency.

ID	Question	Harness	Key hypotheses (placeholder thresholds)	Kill rule
S1	Does evidence-gated authority give logistic, selective succession that beats switch and average on transition regret?	Toy-KB transformer (228 new facts, 40 contradicting updates) extended to 10–20 stages, CPU; arms: hard switch, latest-only, plain average, graded fixed share with reserved probes, graded on self-confidence, oracle	Logistic beats step and constant with width ≥ 2 windows in $\geq 80\%$ of B seeds; false promotion $\leq 5\%$ graded vs $\geq 20\%$ switch; graded regret below switch in E/W/D and below average in B/P; averaging ≥ 15 pp more stale in B; self-confidence no better than switch	Averaging ties graded in B *and* switch ties it in W/E, or shape test fails a planted-step control
S2	Is authority earned: does time to dominance scale with evidence rather than tenure, and does a prestigious incumbent starve a successor?	S1 regime B $\times$ probe budget k $\in \{5, 20, 80\} \times$ effect size $\{10\%, 30\%\} \times$ incumbent prior $\{0.5, 0.9, 0.99\}$; plus a traffic-driven softmax router	log t_{90} falls linearly in log(k $\times$ effect); at prior 0.99 a better successor fails to reach 0.5 in $\geq 30\%$ of seeds under the softmax, and a $\leq$ 10% reserved budget halves that	t_{90} insensitive to k, or starvation $<$ 10% at every temperature
S3	Do canary steps (1→5→25→50→100%) roll back a degrading successor faster, with less regret, than continuous graded authority or switch-plus-undo?	S1 stream, regimes D and B; CIFAR-100 domain-shift variant later behind the GPU lock	Detection latency canary $\leq$ continuous $<$ switch; spurious rollback in B $\leq$ 5%; lowest over-delegation regret for canary	Switch with a rollback rule matches canary on regret and spurious rollback

ID	Question	Harness	Key hypotheses (placeholder thresholds)	Kill rule
S4	Does a dominant successor override protected items, and does a recorded amendment take effect without gradual authority?	Toy-KB plus E3 code; protected set Π, successor trained against a subset Π', amendment to a subset Π''; arms: no constitution, anchor-only, hard check, hard check plus overwrite-on-amendment, "winner writes the rules"	Violations on Π' 0% for hard-check arms, $\geq$ 20% for none and winner-validated once a ≥ 0.9; amendment latency ≤ 1 window with overwrite; protection costs $\leq$ 2 pp on non-protected revisions; durable $\leq 1.5\%$ after 50 unrelated steps (after 7.2)	Hard-check arms $> 5\%$ violations, or protection costs > 5 pp on ordinary revisions (the 4.5 failure)
S5	Under real judges, does judge-earned authority promote exploiters, and does soft authority beat argmax as in 8.14?	Best-of-N stack (Llama 3.1 8B; Skywork-Reward-V2-Qwen3-0.6B, Qwen2.5-Coder-7B judge, self log-prob; GSM8K, MATH-500, HumanEval+); exploit successor via `17_constructed_exploits.py`; cheaper Ollama in-context variant first	Exploit false promotion $\geq 25\%$ under judge-argmax and $\leq$ half that under graded or plurality; judge-driven graded t_{90} within $2\times$ of ground-truth-driven	Exploit never promoted under any judge arm, reported as a real-judge null at generational depth like 8.13

The proposed order is S1 → S4 → S2 → S3 on CPU now, since the toy-KB stream runs in minutes per seed, as E3 did. S5's in-context Ollama variant comes next. The GPU arms queue behind the CCU Phase 1 work.

Every experiment follows the lab's pre-registration practice:

- a frozen protocol, with seeds paired across arms;
- Holm correction within each experiment;
- a Phase 0 on validation splits to fix window size, probe count, temperature and the model-comparison margin, including a power simulation with planted curves to check that steps and logistics are separable at all;

- equal optimiser steps per stage, with probes counted as labels and bytes reported.

Two caveats. Every threshold above is a placeholder, not derived from data. And single-seed differences under about 6 pp are noise at 1.5B scale (7.8). Finally, no novelty gate has been run against online model selection, bandit routing, progressive delivery for ML, or population-based training. That check belongs in `docs/NOVELTY-GATE.md` before any of S1–S5 is frozen.

Conclusion

Reframing GCCL as graded succession by earned authority changes what the lab has to build. The question shifts from who wins to who is entitled to decide what, on what evidence, and for how long. "Competition" left that implicit, and the lab's own results show what fills the vacuum by default: self-confidence, majority volume, incumbent prestige and gameable scorers. Earned succession assigns each of those roles to an explicit part, with distinct jobs:

- an independent evaluator decides who is right;
- an e-process ladder decides how much authority each generation holds;
- a fixed-share floor keeps rollback possible;
- a record with justification links decides what supersession propagates to;
- a screened, expressly amended constitution decides what no amount of competence can touch.

The least-tested part of the design is not the constitution, which E3, 7.1 and 7.2 already partly support. It is the claim that authority over *ordinary* knowledge should be item-wise and evidence-gated, rather than set by a global stability knob. E4 shows the cost of a global knob, and no published method does the item-wise version. If S1 and S4 come out as predicted, the lab would have a measurable definition of "earned" (selectivity plus evidence dose-response) and a measurable definition of "protected" (zero violations, immediate amendment). It would also have one finding that cuts against intuition: a constitution should change *abruptly and deliberately* while knowledge changes *gradually and on evidence*, and conflating the two speeds is the failure mode to design against. If S1's kill rule fires, the right conclusion is that canarying plus a good evaluator suffices. The gradient would then be an implementation detail, not a research programme.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/graded-succession-earned-authority