The Stroboscopic Generalization Hypothesis

Orbit-Level Capability and Harness-Level Agency

Kamil Dixon
Third Rail
April 2026 — v1.2.1

Abstract

A model checkpoint is a point estimate. Standard evaluation treats capability as a property of that point. Yet many successful techniques — exponential moving averages, stochastic weight averaging, model soups, ensembles, self-consistency, and recursive model calls — produce observed capability by integrating across trajectories, samples, or calls rather than by relying on a single configuration. We propose the *Stroboscopic Generalization Hypothesis*: a system may exhibit stable apparent generalization when partial capabilities are traversed and integrated by an observer faster than the observer's integration window decays, in the same way a rotating point of light is perceived as a continuous ring. We develop the hypothesis along four anchors: the distinction between checkpoint-level and orbit-level capability; the integration window as a tunable control surface; a failure mode we term the *false ring*, in which integration masks brittleness or basin bias; and the consequence that in continuity-preserving systems, agency must be specified separately from generalization. The paper extends prior work on commit regimes [Dixon, 2026] from *when* generalization happens to *where* it lives. We specify falsification conditions: cases where integration should improve apparent capability, cases where it should fail, and probes designed to distinguish distributed capability from smoothed brittleness. We argue that in systems that act over time, agency is not a direct consequence of generalization but a property of the harness that selects, commits, remembers, governs, and preserves continuity.

1. Introduction

A point of light fixed to a rotating wheel, traversing fast enough, is perceived as a continuous ring. The light occupies one position at any instant. The ring is a perceptual artifact — produced by the observer's integration over time, not by the light itself. This is *persistence of vision*, the same effect Plateau used to build his 1832 phenakistoscope and that drives every cinema, every CRT, every blinking traffic light.

Something analogous may be at work in machine learning evaluation. The hypothesis is not that ML systems literally behave like visual perception. The analogy names a *locus error*: attributing an integrated phenomenon to a point state.

Standard practice treats model capability as a property of a checkpoint. Weights are loaded, queried, scored. The score is attributed to that configuration. The procedure is reproducible and convenient, and for many purposes — version comparison, tracking training progress, assigning credit to architectural changes — it is sufficient.

It is also incomplete.

Several techniques in widespread use already imply that observed capability does not live where it is attributed. Exponential moving averages stabilize generation quality in diffusion training. Stochastic weight averaging produces flatter optima from trajectory integration. Model soups recover capability by averaging fine-tuned variants. Ensembles and self-consistency methods produce behavior no single forward pass exhibits. Recursive language model architectures distribute reason-

ing across calls. In each case, the apparent capability of "the model" is a function of trajectory and integration, not of any single point estimate.

We name the underlying effect the *Stroboscopic Generalization Hypothesis*. Stated informally: a system may exhibit stable apparent generalization when partial capabilities are traversed and integrated by an observer or controller faster than that observer's integration window decays. The ring is real to the observer. Whether it is intrinsic to the substrate is a separate question, and one we treat seriously in Section 4.

The hypothesis extends prior work on commit regimes [Dixon, 2026], which argued that generalization timing has phase structure and is sometimes controllable, sometimes not. If generalization timing has phase structure, then the locus of generalization itself deserves examination. The question is no longer only *when* generalization commits but *where* it lives — point, orbit, or observer.

This distinction matters more as model systems shift from single-call question answering toward long-horizon agents, tool-using systems, recursive inference, persistent memory, and multi-model orchestration. In these systems, the checkpoint is only one component of the operative system. The integrator increasingly determines what the system appears able to do.

The paper develops the hypothesis in the sections that follow.

Section 2 distinguishes checkpoint-level from orbit-level capability and surveys the techniques that already imply orbit-level effects. Section 3 introduces the integration window as a tunable control surface — the choice of how much temporal and spatial evidence is required before a state becomes an action, memory, or update. Section 4 names a failure mode we call the *false ring*, in which integration produces apparent stability that masks brittleness or basin bias, and connects it to the geometry of the loss landscape. Section 5 draws the consequence: in continuity-preserving systems, agency is not a property of any model but of the observer-controller that integrates models over time. These distinctions matter for agentic systems because generalization alone does not produce agency. A system may generalize across inputs without preserving identity, forming goals, maintaining commitments, or governing action over time. Section 6 proposes falsification conditions and an eval surface for testing the hypothesis empirically.

We do not claim to demonstrate the hypothesis. We claim it is well-formed, consistent with several lines of existing work, and falsifiable under conditions specified in Section 6. The contribution is conceptual: a frame in which a number of accumulated empirical observations cohere, and a set of predictions that frame entails.

A note on terminology. We use *observer* throughout in a deliberately broad technical sense. By observer, we do not mean only a human evaluator. We mean any process that integrates model states, outputs, samples, calls, memories, or actions into a downstream judgment or decision. An observer may be an EMA update rule, an ensemble vote, a benchmark harness, a planner, a memory system, or a human evaluator. The hypothesis does not depend on subjective perception. It depends on the presence of a non-trivial integrator between substrate behavior and downstream consequence.

---

2. Checkpoint vs. Orbit

A checkpoint is a point estimate of a model's parameters. A training trajectory is an orbit through parameter space. Most evaluation infrastructure assumes capability is a property of the point.

The assumption is partially correct. A checkpoint can be loaded, queried, and scored; results are reproducible; the protocol is standard. For comparing models, tracking improvement across runs, and assigning credit to architectural changes, point-level evaluation does the job.

It is incomplete in cases where the operative system is not the checkpoint.

A note on terminology before proceeding. We use *orbit* in two related senses. A *parameter-space orbit* is the trajectory traced by model weights during training or fine-tuning. A *behavior-space orbit* is the set of outputs, intermediate states, tool calls, or recursive invocations generated by a model-controller system at inference time. The hypothesis applies to both, but the integrator differs: weight averaging integrates parameter-space orbits, while voting, recursion, memory, and planning integrate behavior-space orbits. In both cases, the operative system is the integrated trajectory, not the sampled point.

Consider exponential moving averages in diffusion training [Karras et al., 2022]. The live model — the one being updated by the optimizer — is often noisy, and samples drawn from it look correspondingly worse than samples drawn from the EMA copy. Practitioners have known this for years. The EMA copy is not an independently trained model. It is an integrated view of a trajectory the live weights traversed. The capability sits in the integration, not in any single point along the way.

Stochastic weight averaging [Izmailov et al., 2018] generalizes the idea. Averaging weights from later epochs of a learning rate cycle produces flatter optima with better test performance than any individual epoch's weights. The flatness is a property of the trajectory's geometry. It is not visible at any one point.

Model soups [Wortsman et al., 2022] extend the principle across fine-tunes. Averaging weights of independently fine-tuned variants of a base model yields better performance than any individual variant, provided the variants live on a connected region of the loss surface. This works because of linear mode connectivity [Frankle et al., 2020; Garipov et al., 2018] — under broad conditions, models that share a pretraining trajectory can be linearly interpolated without crossing a barrier. The capability is distributed across a manifold; the average integrates it.

At inference time, the same pattern recurs in different guise. Ensembles produce predictions no single member exhibits. Best-of-n sampling, majority voting, and self-consistency methods improve task performance by aggregating multiple samples [Wang et al., 2022]. Recursive language model approaches, in which a model is invoked repeatedly over decomposed slices of a problem, produce reasoning behavior the single forward pass does not [Zhang et al., 2025]. Equivariant architectures form a related limit case [Cohen and Welling, 2016]: by imposing symmetry constraints, they tie behavior across a group orbit, reducing the need to observe or sample each transformed instance independently.

What unifies these techniques is an implicit claim about locus. In each case, the operative system is not the point. It is the integration of points — across time (EMA, SWA), across a manifold (soups, ensembles), across samples (self-consistency), across calls (recursive inference), or across a symmetry group (equivariance). The point is a sample. The operative capability lives in what is sampled from.

The stroboscopic frame names this directly. A capability traversed faster than the observer's integration window decays appears, to that observer, as continuous coverage. The phenomenon is general. It occurs whenever a non-trivial integrator sits between substrate and evaluation. EMA

is one such integrator. An ensemble is another. A recursive controller is another. A benchmark harness that aggregates multiple samples, tools, or retries is another.

The frame does not deny that point capability matters. A model with sufficiently sparse coverage cannot be saved by integration, just as a stationary light cannot become a ring. What the frame predicts is that for systems with reasonable point coverage and an effective integrator, observed capability will diverge from checkpoint capability in measurable ways. The divergence is not merely noise to be cleaned up. It is signal about the relationship between substrate, trajectory, and observer — and, in some cases, evidence that the observer is manufacturing stability the substrate does not possess.

Section 3 takes up the integrator's parameters: how the observer's window is set, how it should be set, and what changes when we treat it as a control variable rather than an artifact of measurement.

---

3. The Integration Window as Control Surface

The previous section argued that observed capability often lives in integration rather than in any single point. This section asks the natural follow-up. If integration is doing the work, what determines how much integration occurs, and over what?

We call this the *integration window*: the span of evidence an observer-controller uses before converting substrate behavior into a stable downstream state. The downstream state may be an output, an action, a memory write, a parameter update, a goal commitment, or a verdict on capability. The integration window is the gate between the orbit and the consequence.

In every technique surveyed in Section 2, an integration window is set. It is rarely set explicitly, and rarely set by design. EMA decay rates are inherited from defaults. Ensemble sizes are chosen for compute budget. Best-of-n samples track what the API charges. Benchmark harnesses are configured for reproducibility, not for the question being asked. The window exists, but it is treated as a measurement artifact rather than a control variable.

We argue this is a missed surface. The integration window can be specified along five dimensions, each of which is implicitly tuned in current practice and could be tuned explicitly.

Temporal width is how far across training time, inference time, or interaction history the observer integrates. EMA at low decay integrates a long temporal window. Single-sample inference integrates none. A planner with persistent memory integrates across an entire interaction history.

Spatial width is how many models, operators, samples, tools, or agents are integrated. An ensemble of three has narrow spatial width. A model soup over fifty fine-tunes has wide spatial width. A multi-agent system with delegated subtasks has spatial width determined by its branching factor.

Decay rate is how quickly older evidence loses influence. EMA explicitly parameterizes this. Memory systems implement it through retrieval scoring, recency weighting, and consolidation policies. Single-pass inference has no decay because it has no history.

Threshold is how much agreement is required across the integrated evidence before the downstream state changes. Majority vote uses a fifty percent threshold. Self-consistency methods use plurality. Approval-gated agentic actions use unanimity from named verifiers. A high threshold means stronger agreement is required before commitment; a low threshold means commitment can occur on weaker or noisier signal.

Risk sensitivity is how the window adapts to the consequence class of the decision. Routine decisions can run on narrow windows. Irreversible, identity-relevant, or high-stakes decisions warrant wider windows: more samples, longer temporal integration, higher thresholds, slower decay. In current ML practice this dimension is almost never explicit. In agentic systems it has to be.

These dimensions are not independent. Widening the temporal window without adjusting decay can drown current signal in stale evidence. Increasing spatial width without adjusting threshold can produce the appearance of consensus by aggregating correlated weak learners. Calibration matters as much as size.

A wider window is not automatically a better window. Wider integration can improve robustness when evidence is diverse and weakly correlated, but it can also amplify shared bias, stale memory, correlated errors, or delayed response. The relevant design question is not how to maximize the window, but how to calibrate it to the consequence class.

The framing has consequences for how we read existing methods. Stochastic weight averaging is a temporal-width intervention with a particular decay schedule. Model soups are a spatial-width intervention with uniform or validation-weighted contribution across variants. Self-consistency is a behavior-space spatial-width intervention with majority threshold. Best-of-n with reranking is spatial width plus a separate threshold mechanism. The methods are not unrelated tricks. They are configurations of the same control surface.

It also has consequences for evaluation. A benchmark score is the output of a particular integration window, not an unmediated property of the model. Two evaluations of the same checkpoint, with different windows, will report different capabilities — and both reports may be accurate within their respective windows. This does not make capability subjective. It makes the integration window part of what gets reported.

The deeper move is that the integration window is not only a measurement parameter. In agentic systems, it determines what the system *does*. A harness that reads model output, integrates across calls, consults memory, weighs against constitutional rules, and only then commits to action is implementing a wide integration window. A harness that pipes a single model output directly to a tool call is implementing a window of width one. The first will exhibit different behavior than the second, even if the underlying model is identical.

We can put this directly. The integration window is not a measurement artifact. In agentic systems, it is a governance surface.

This is the bridge from hypothesis to architecture. If apparent generalization is partly produced by integration, then a system's behavior is partly produced by *how* it integrates — by the temporal width, spatial width, decay, threshold, and risk sensitivity its harness implements. These choices are not downstream of the model. They are upstream of the system's effective character.

The risk this creates is the topic of Section 4. An integration window can reveal distributed capability the substrate genuinely possesses. It can also manufacture the appearance of capability the substrate does not possess. The dimensions above are tools; they are not safeguards. Distinguishing these two cases requires probes designed for the purpose, and a clear account of what kind of failure each probe is meant to detect.

---

4. The False Ring and Basin Geometry

The first three sections argued that integration is doing real work, that the control surface implementing it can be specified along five dimensions, and that calibrating those dimensions matters more than maximizing them. This section presses on a question those sections deferred. When a system exhibits stable apparent capability under integration, when is the stability a property of the substrate, and when is it a property of the integrator?

The hypothesis demands the question. If integration can reveal distributed capability, it can also fabricate the appearance of capability that does not exist. The cap on the rotating wheel is a useful image precisely because it admits both readings. Spin a single light fast enough and you see a ring. The ring is not a lie — it is what your visual system is doing — but it is also not evidence that the wheel is luminous. We have to be honest about which case we are in.

We call the failure mode the *false ring*: a system appears broadly competent because its integration layer smooths over discontinuities, but targeted probes reveal that the underlying capability is sparse, brittle, or basin-biased. The false ring is not noise. It is a structurally produced illusion of coverage, generated by the same integrators that, in other cases, surface real distributed capability.

The mechanisms behind the false ring are not exotic. They are familiar.

Smoothed brittleness. An EMA reduces variance in evaluation. Variance reduction looks like capability gain. If the underlying live-weight trajectory is in fact unstable, the EMA hides the instability rather than resolving it. Practitioners encounter this when a model with good EMA scores degrades sharply on a different evaluation seed, a different prompt template, or a slight distribution shift. The smoothing did not produce competence. It produced the inability to detect incompetence.

Correlated weak learners. An ensemble of N models gives the appearance of N independent samples, but if the models share architecture, training data, and hyperparameters, their errors are correlated. Aggregating correlated errors tightens confidence intervals around a biased estimate. Spatial width without diversity is a confidence trap. Self-consistency methods inherit this risk: K samples from one model under one prompt distribution can converge on an answer the model is consistently wrong about.

Phase-locked evaluation. When evaluation cadence aligns with training cadence, results sample a particular phase of the training trajectory and miss the rest. This is the wagon-wheel effect made into an ML hazard. A weekly eval run that always lands shortly after a learning-rate restart will report the system's best-behaved moments and miss its volatile ones. The integration window does its job; the window itself is just placed wrong.

Basin-biased coverage. This is the deeper case, and the one that connects the false ring to optimization geometry. SGD trajectories are not uniform tours of the loss surface. They are biased samplers, weighted by basin width, gradient magnitude, and step-size schedule. The "orbit" through parameter space looks, on paper, like a path through the full landscape. In practice, it spends nearly all its time in a small region of low curvature, with characteristic biases [Frankle et al., 2020; Garipov et al., 2018]. When weight averaging operates on this trajectory, it integrates over a probability cloud whose density structure mirrors basin geometry, not task structure. The averaged model inherits whatever the basin happens to encode well — and whatever it happens to encode poorly. Coverage on the surface does not imply coverage of the task.

The cleanest version of this critique applies equally to behavior-space orbits. A model invoked

many times under similar prompts samples its high-density behavior modes. Integration across those samples can manufacture the appearance of stable competence on whatever those modes happen to handle, while leaving low-density modes — rare inputs, adversarial inputs, distribution-shifted inputs — untouched. Self-consistency methods that report high agreement under in-distribution prompts can fall to chance under modest paraphrase or domain shift. The ring is real on the inputs the integrator saw. It is not yet evidence of a model-level property outside that integration regime.

The same false-ring logic applies when the integrated object is not merely an answer or score, but the apparent continuity of an agent.

Identity smear. This case is specific to agentic systems and worth flagging separately. When a harness integrates across multiple cognitive engines or operators with distinct behavioral fingerprints, naive averaging produces output that has no fingerprint at all. The integration succeeds in the technical sense — outputs are produced, scored, accepted — but the system's distinctive character is averaged away. From the outside, the system looks consistent. From the inside, it has no there there. Identity smear is the false ring applied to the agent layer.

These mechanisms share a structural feature. In each case, the integrator is operating correctly given its inputs. The smoothing, the aggregation, the averaging, the cross-call integration — these all do what they were designed to do. The failure is not in the integrator. It is in the implicit assumption that whatever the integrator returns is a property of the substrate. The hypothesis as stated in Section 1 rules this out: the ring is real to the observer, and whether it is intrinsic to the substrate is a separate question. Section 4 is where that caveat earns its keep.

The question becomes how to tell the cases apart. Here the news is mixed. Some false rings can be detected with standard tools. Out-of-distribution evaluation, adversarial probes, prompt-template variation, and seed sensitivity tests all push on different forms of smoothed brittleness. Diversity-aware ensembling, deliberate decorrelation of training data, and fingerprint-preserving harness designs address the correlated weak learner case. Phase-locked evaluation can be defended against with randomized eval cadence. None of this is new.

The harder cases are the ones where the integrator and the evaluator share assumptions. If both are configured around the same in-distribution behavior modes, both will report stability. If both inherit basin biases from the same training trajectory, both will fail to detect the bias. The classical defense — test on data the system has not seen — is harder to apply when the system is itself a controller integrating over a behavior orbit, because "data" and "behavior" become entangled. We will return to this in Section 6, where falsification conditions for the hypothesis are proposed. The relevant probes have to be designed against specific suspected failure modes, not against capability in the abstract.

The point of this section is not that integration is unreliable. The point is that integration is a mechanism, and like any mechanism it admits modes of correct operation and modes of failure that look like correct operation from outside. Sections 1–3 argued that integration is where capability often lives. Section 4 argues that this is also where false reports of capability often originate. The two claims are not in tension. They are the two faces of the same hypothesis. A frame in which the integrator is taken seriously has to take seriously what the integrator can do wrong.

Section 5 turns this back on agentic systems. If the integration window is a governance surface, and if that surface admits structurally generated illusions, then agency cannot be inferred from

generalization. It has to be designed at the level of the integrator itself.

---

5. Harness-Level Agency

The previous sections developed a frame in which observed capability is often a property of integration rather than of any single model state, in which the integration can be specified as a control surface, and in which the same control surface admits structurally generated illusions. This section draws the consequence for systems that act.

The frame separates two questions that are routinely conflated. *Generalization* describes what a system can cover — the breadth of inputs it can handle, the distributions it can extrapolate across, the compositions it can compose. *Agency* describes how a system selects, commits, acts, remembers, revises, and remains itself over time. The distinction matters because the techniques that produce generalization do not produce agency. A model that handles arbitrary inputs gracefully still has no native procedure for deciding which of its outputs should become an action, which should be retracted, which should persist as memory, or which should define the agent across future calls.

Stated plainly: generalization has no native concept of commitment. A forward pass produces a distribution over outputs; it does not produce a decision about which output, if any, should change the world. The model is indifferent between sampling and acting, between answering and remembering, between assertion and refusal. Whatever resolves these — whatever turns capacity into commitment — is not the model.

We will use *harness* for the layer that resolves them. The term is intentionally general. A harness is whatever sits above one or more cognitive engines and is responsible for selection, commitment, memory, goal formation, governance, and continuity. Even a bare chat loop is a harness; it simply hides its agency decisions inside defaults. The simplest harnesses are implicit: a chat interface that pipes a single sample to display, a tool wrapper that forwards function calls without inspection. The most elaborate harnesses are explicit: planning systems, multi-agent runtimes, constitutional layers, persistent memory architectures. The claim of this section is that whatever sits in this position — implicit or explicit, simple or elaborate — is the actual locus of agency in the system. Generalization is what the engine offers. Agency is what the harness does with it.

We can be more specific. A harness performs at least six functions, each of which is an integration over the substrate it sits above.

Selection. Among possible outputs, samples, candidates, or actions, which is chosen? Selection is the harness's first integration: from a distribution to a particular. Greedy decoding is selection at width one. Best-of-n with a reranker is selection at width N. Multi-agent debate is selection at width N with structured argumentation. The integration window from Section 3 directly parameterizes selection.

Commitment. When does a tentative output become an action, a memory write, or a parameter update? Commitment converts orbit-level integration into discrete consequence. A harness that commits at width one — every model output becomes action — has a different effective character than one that commits only after multi-sample agreement, even if the underlying engine is identical. Commitment is the threshold dimension of the integration window applied at the action layer.

Memory. What persists across calls, and how is it retrieved? Memory is integration over time made structural. Without memory, every call is fresh; the system has no autobiography and no

commitments older than the current context window. With memory, the harness must decide what gets written, when it gets consolidated, how contradictions are handled, and how confidence decays. Each of these is an integration policy, and the policies determine what the system *is* across time as much as the engine determines what it can produce in any single call.

Goal formation. What becomes worth pursuing? Most current systems take goals as input — a user prompt, a task specification, a reward signal. A harness with goal-formation capability generates candidate goals from observation: anomaly relative to a world model, value violation relative to constitutional rules, commitment tracking, opportunity detection, curiosity-driven uncertainty reduction. The candidates are typed, not raw — at minimum by source, confidence, reversibility class, time horizon, and approval requirement. Goal formation matters here because it is the component most often elided when systems are described as "agentic." A system that pursues only externally specified goals is delegated, not autonomous. Autonomy here does not imply unconstrained action; it means internally proposed goals still pass through governance, approval, reversibility, and continuity constraints. Whether internal goal formation is desirable is a design question; that the difference exists is not.

Governance. Which actions require approval, which can be taken autonomously, which must be refused, which must be reversible, which must be logged? Governance is the harness layer that turns commitment into bounded commitment. Its presence or absence is the difference between an agent that acts and an agent that acts safely. In current practice, governance is often retrofitted via post-hoc filters or wrappers; in the frame proposed here, it is more naturally specified as a calibration of the integration window — wider window, higher threshold, stricter approval class for higher-consequence decisions.

Continuity. What preserves the agent across engine swaps, operator changes, memory updates, or self-modifications? An agent whose constitution drifts under updates, whose behavioral fingerprint averages away under operator integration, whose memory ossifies or fragments, is an agent that does not survive its own operation. Continuity is the property that makes the agent the same agent across these transitions. It is the slowest integration, and the most consequential — the one that determines whether there is a continuous "this system" at all, or only a sequence of similar-looking responses.

These six functions are not independent. Selection feeds commitment. Commitment writes memory. Memory shapes goal formation. Goal formation routes through governance. Governance is constrained by continuity. The harness is what implements them coherently. The cognitive engine, however capable, is upstream of all six.

The frame yields several immediate consequences.

A model swap is not an agent swap, *if* the harness preserves the relevant invariants. Conversely, an apparent agent built without an explicit harness — for instance, a chat loop with no persistent memory or governance layer — is not made into an agent by upgrading its underlying model. Increasing engine capability raises the ceiling on what can be selected, committed to, remembered, pursued, and governed; it does not, by itself, produce selection, commitment, memory, pursuit, or governance.

Generalization-based extrapolations toward agency therefore require care. Claims of the form "this system is more agentic because it generalizes better" misplace the locus. Better generalization makes the engine's outputs more useful as inputs to the harness. It does not constitute agency in

the absence of a harness layer that implements the six functions above.

The false ring discussion of Section 4 applies here with particular force. A harness that integrates across many model calls without diversity in inputs, without phase-decorrelated evaluation, without fingerprint-preserving design, will produce systems that *look* coherent in the average and dissolve under targeted probes. The illusion is not at the engine layer. It is at the harness layer, which is exactly the layer where agency lives. A false-ring agent is not a contradiction in terms; it is a likely outcome of building agentic systems on the assumption that generalization implies the rest.

Three clarifications are worth making explicit.

First, the harness account is not offered as a complete theory of agency. It is a locus claim: whatever agency a system exhibits must be implemented in the layer that selects, commits, remembers, governs, and preserves continuity over time. The paper does not specify how each function should be designed, only where the design surface is.

Second, this section is not a claim about which specific harness designs are correct. It is a claim about where the design surface is. Selection, commitment, memory, goal formation, governance, and continuity are not optional features for "advanced" agents. They are the surface on which any system that acts over time is implicitly designed, whether or not the design is articulated. The choice is between designing them deliberately and inheriting them from defaults.

Third, the section does not collapse into a claim that more harness is always better. The same calibration argument from Section 3 applies. A system with only an implicit, width-one harness has at most brittle or momentary agency. A system with maximally elaborate harness can be slower, more brittle, and harder to evaluate than the substrate it wraps. The relevant question is not how much harness, but what calibration of the six functions is appropriate to the system's intended operating regime and consequence class.

Section 6 turns to the empirical question implicit in everything above. If observed generalization can be a property of integration, if integration admits structurally generated illusions, and if agency is implemented at the harness layer, then evaluating any of these requires probes designed against specific suspected mechanisms — not against capability in the abstract. The next section sketches what such probes have to do.

---

6. Falsification Conditions and Eval Surface

A hypothesis paper without falsification conditions is an essay. This section specifies the conditions under which the Stroboscopic Generalization Hypothesis loses force, the eval families that bear on those conditions, and the minimal empirical prediction the framing commits to.

6.1 What would weaken or falsify the hypothesis

Five predictions follow from Sections 1–5. Each can fail in a structured way. The table below names each prediction, the result that would weaken or falsify it, and what the experiment must control for to make the test informative.

The purpose of these tests is not to prove that integration always helps. It is to determine whether capability changes systematically when the integration window changes, and whether those changes survive probes designed to expose smoothing artifacts.

Prediction	Falsifying result	Experimental requirement
Changing the integration window measurably alters apparent capability for a fixed substrate.	Different windows produce no meaningful score or behavior differences.	Substrate held fixed; only window dimensions (temporal width, spatial width, decay, threshold, risk sensitivity) vary; metrics sensitive enough to detect structured change.
Wider integration windows help only when the integrated evidence is diverse and weakly correlated.	Wider windows help equally under correlated and decorrelated inputs.	Matched ensemble or sample sets — same width, varied correlation structure; effects measured against the diversity-controlled baseline.
False rings collapse under targeted probes designed against specific failure modes.	Integrated systems remain robust across OOD, paraphrase, seed-sensitivity, evaluation-cadence, and low-density-input probes.	Probes specified against named mechanisms from Section 4 (smoothed brittleness, correlated weak learners, phase-locking, basin bias); robustness reported per probe family, not aggregated.
Harness-level agency varies independently of model capability.	Stronger underlying models produce stronger agentic behavior even under identical weak harnesses.	Crossed design: same harness across multiple engines, same engine across multiple harnesses; agency measured by harness-function metrics, not engine output quality.
Identity continuity depends on harness invariants, not on engine identity.	Engine swaps preserve continuity even when harness invariants are not preserved.	Continuity measured against named invariants (constitutional surface, autobiographical thread, behavioral fingerprint, active commitments); engine swaps performed under controlled invariant-preservation and invariant-removal conditions.

The first prediction is the load-bearing one. If it fails — if integration windows turn out to be measurement parameters with no structured effect on observed behavior — the rest of the framing collapses regardless of how the other four resolve.

6.2 Eval families

The predictions above suggest five eval families.

Window-sweep evals. Hold substrate fixed; vary the five dimensions of the integration window. Report capability and behavior as a function of each dimension and combinations of them. Look for structured surfaces, not isolated points.

Correlation and diversity evals. Compare integration over correlated samples, models, or fine-tunes against integration over decorrelated equivalents at matched width. The hypothesis predicts that decorrelation amplifies the benefit; lack of differential effect is informative.

False-ring probes. Probes specified against named mechanisms: out-of-distribution evaluation against smoothed brittleness; correlated-error stress tests against weak-learner aggregation; randomized eval cadence against phase-locking; low-density and adversarial inputs against basin bias; behavioral-fingerprint stability under operator integration against identity smear. Report per probe, not aggregated.

Harness ablation evals. Two crossed studies. (i) Same harness, varying engines: does harness function quality scale with engine capability? (ii) Same engine, varying harnesses: does observed agency vary with harness design? The hypothesis predicts substantial decoupling — strong models in weak harnesses underperform on harness-function metrics; weaker models in well-designed harnesses outperform.

Continuity and invariant evals. Engine swap with controlled invariant preservation. Memory under contradiction injection. Goal-source attribution under post-hoc interrogation. Constitutional drift detection under adversarial update sequences. Each tests a specific harness function; results are not interchangeable.

These families are not exhaustive. They are minimum-sufficient for the predictions above. Researchers extending the framing should expect to add families specific to their substrate or domain.

A warning applies across all five families. These evals should not be collapsed into a single aggregate score. The hypothesis predicts mechanism-specific effects, so aggregation can hide the very failures the probes are meant to expose. A system that scores well on average across out-of-distribution, paraphrase, seed-sensitivity, and basin-bias probes — but fails sharply on one of them — has not been shown to be robust; it has been shown to fail in a structured way that an averaged score conceals. Per-mechanism reporting is the precondition for taking the eval seriously.

6.3 The minimal empirical prediction

For a fixed substrate, changing the integration window should produce measurable, structured changes in observed capability and agentic behavior. *Structured* is doing real work in this sentence: the prediction is not merely that windows have effects, but that the effects vary systematically with the dimensions specified in Section 3 — temporal width, spatial width, decay rate, threshold, risk sensitivity — and that the variation can be characterized rather than only observed.

If no such structured changes appear, the stroboscopic framing loses force. The hypothesis becomes a redescription of measurement noise. We do not believe this is the result researchers will find, but the hypothesis must be willing to fail in this way for it to be worth the name.

6.4 What this paper does not claim to settle

A hypothesis is not a finding. We do not claim that any specific eval result has yet been produced under this framing, that any particular harness design satisfies the criteria sketched in Section 5, or that any current system is or is not subject to the false-ring failure mode. The paper specifies a frame in which these become questions with structured answers. The answers themselves are the work of the eval suite this section sketches and the harness specifications it points toward.

Two artifacts naturally follow from this paper. The first is an eval suite implementing the five families above against contemporary models and harnesses. The second is an architectural specification for harnesses that target the six functions of Section 5 with explicit calibration of the integration window. Neither is in scope here. Both are precommitted to as next steps in the research program this paper opens.

---

7. Discussion

The frame we have proposed — capability often distributed across orbits and integrated by an observer-controller, integration as a tunable surface, the false ring as a structurally generated illusion, agency implemented at the harness layer — is conceptual rather than empirical. Its contribution is to make a number of accumulated practices and observations cohere, and to surface predictions and design questions that a checkpoint-centric view does not naturally generate.

Several lines of existing work become more legible in this frame. The success of weight averaging methods becomes a statement about parameter-space orbits and their integrators. The robustness gains of self-consistency and ensembling become a statement about behavior-space orbits. Linear mode connectivity becomes a structural condition on when parameter-space integration can succeed. The persistent gap between strong base models and reliable agentic systems becomes a statement about the absence or weakness of harness functions, not a deficiency of the underlying engines. None of these reframings are forced; each is offered as a candidate description that the hypothesis renders coherent.

The frame also relocates several familiar concerns. AI evaluation that depends on aggregated benchmarks now depends, by the argument of Section 3, on the integration windows those benchmarks implement. Reports of capability are not unmediated; they are functions of substrate and integrator together. The implication is not that evaluation is unreliable but that evaluation is itself a designed artifact whose properties matter as much as the systems it evaluates. The integration window is a governance surface for the evaluator as much as for the agent.

Limitations should be named directly. The hypothesis as stated does not specify a quantitative model of integration. Words like "wider," "narrower," "calibrated," and "structured" are placeholders for measures that will need to be made precise per substrate and per task. The hypothesis also does not predict which integration regimes will succeed in advance; the prediction is that effects will be structured, not that the structure can be derived from first principles. And the harness functions of Section 5, while we believe the list is at least minimally complete, are not proven to be jointly sufficient. Future work may add to the list or merge entries within it.

The relationship to prior work on commit regimes [Dixon, 2026] is now visible in full. Commit Regimes argued that learning has phase structure and that generalization timing is sometimes controllable. The present paper extends that line by asking where the generalization that timing

applies to actually lives. The answer offered here — in the integrated trajectory, under an observer-controller, subject to structurally generated illusions, with agency implemented at the harness layer — does not contradict the earlier work. It places it. Phase structure in learning describes the shape of the orbit. The present paper describes how the orbit is observed, integrated, and acted on, and where in that pipeline agency emerges or fails to.

We close with the line that has organized the paper from Section 1. The point is a sample. The operative capability lives in what is sampled from. Generalization is what the engine offers. Agency is what the harness does with it. The Stroboscopic Generalization Hypothesis is the proposal that taking these statements seriously — as architecture rather than as metaphor — yields a more accurate account of how capable systems behave, and a more honest account of how they fail.

References

Cohen, T. S., & Welling, M. (2016). Group Equivariant Convolutional Networks. *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, PMLR 48:2990–2999. arXiv:1602.07576.

Dixon, K. (2026). *Commit Regimes in Learning: When Generalization Timing Is Controllable — and When It Isn't.* Third Rail.

Frankle, J., Dziugaite, G. K., Roy, D. M., & Carbin, M. (2020). Linear Mode Connectivity and the Lottery Ticket Hypothesis. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, PMLR 119:3259–3269. arXiv:1912.05671.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., & Wilson, A. G. (2018). Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. *Advances in Neural Information Processing Systems 31 (NeurIPS)*. arXiv:1802.10026.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging Weights Leads to Wider Optima and Better Generalization. *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 876–885. arXiv:1803.05407.

Karras, T., Aittala, M., Aila, T., & Laine, S. (2022). Elucidating the Design Space of Diffusion-Based Generative Models. *Advances in Neural Information Processing Systems 35 (NeurIPS)*. arXiv:2206.00364.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *International Conference on Learning Representations (ICLR 2023)*. arXiv:2203.11171.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., & Schmidt, L. (2022). Model Soups: Averaging Weights of Multiple Fine-Tuned Models Improves Accuracy Without Increasing Inference Time. *Proceedings of the 39th International Conference on Machine Learning (ICML)*, PMLR 162:23965–23998. arXiv:2203.05482.

Zhang, A. L., Kraska, T., & Khattab, O. (2025). Recursive Language Models. arXiv:2512.24601.