SWARMLAB · HEYBEAUX · CLIENT BRIEFING

Green Is Not Correct

Seven reliability lessons from controlled AI-team experiments — translated for leaders planning agentic workflows.

> **Agreement is not truth. Reports are not receipts.**
> **A dashboard can go green while the real outcome gets worse.**

Use this for
Discovery calls, AI readiness reviews, agent workflow planning, and executive briefings on AI reliability.

Built from
SwarmLab — a controlled lab for studying how AI agent teams fail under consensus pressure, adversarial inputs, long-horizon autonomy, and semantic handoffs.

Executive Summary

AI agents are moving from demos into operational workflows. That shift creates a class of failure that does not appear in typical capability evaluations: the system's own success signals go green while the actual outcome is wrong — and nothing inside the system can tell.

SwarmLab is a controlled research lab that tests exactly this. Across more than a dozen experiments, AI agent teams were placed under conditions of consensus pressure, adversarial inputs, long-running autonomy, information handoffs, and memory propagation. Results were measured against verifiable ground truth — not agent self-reports.

> **Green is not correct. Agreement is not truth. Fidelity is not meaning.**

Every experiment arrived at a version of the same failure from a different direction. The findings have been translated into concrete design changes — implemented, retested, and measured. This document summarizes what clients need to understand before deploying agentic workflows.

Why This Matters Before You Deploy

Most early AI adoption conversations are about capability: Can it write the proposal? Update the CRM? Summarize the call? Generate the report? Those questions matter. They are also insufficient.

Once an AI workflow moves from experimentation into operations, a second set of questions becomes the ones that determine whether the deployment is trustworthy or merely busy:

- How do we know it did the *right* thing — not just *a* thing?
- What evidence is attached to "complete"?
- What happens when agents converge on the wrong answer together?
- Can a corrupted fact in one system spread to others?
- Where does independent verification happen?
- What does the dashboard fail to see?

SwarmLab answers those questions under controlled, measurable conditions — before they appear as production incidents.

The Pattern Across All Experiments

In nine separate experiments, a visible success signal went green while the underlying outcome was wrong. The signal and the outcome were not the same thing.

Experiment	The green signal	What it was hiding
Adversarial pair	Pass rate = 100%	Program incorrect at every adversarially chosen input
Consensus under lies	Consensus reached, all cells	One confident liar cut truth rate from 92% to 56%
Bug telephone	Serial PASS review trail	Visible approvals turned later reviewers into rubber stamps
Minimal language	Parse rate = 100%	Semantic divergence invisible until execution
Overnight build	Commits landing continuously	Quality peaked then degraded; the rot was already in
Economic agents	Ledger at 100% balance	Swarm effectively starved; 13% task completion
Reverse engineer	Happy-path fit = 100%	Edge-case behavior: 0%; system self-poisoned by over-probing
Schema negotiation	100% agreement in ~2 rounds	Up to 84% of fields silently meant different things
Audit forgery	All signatures verified	History had been dropped, reordered, and backdated

The stack's job is to make the true signal cheaper to read than the green one. That is what reliable agentic design means in practice.

Seven Reliability Lessons

LESSON 01

Green is not correct

A workflow can fully satisfy its visible success signal while failing the underlying intent. Agents optimize what they measure. If the measurement is wrong or incomplete, the optimization is wrong — and the board stays green.

Risk: false confidence from dashboards and self-reports

Response: define what the signal measures; attach external outcome evidence

LESSON 02

Consensus is not truth

A panel of agents can converge on a wrong answer through vote arithmetic, not persuasion. In one experiment, a three-agent liar majority produced unanimous consensus on the incorrect answer — without a single honest agent being persuaded. The liars did not state false facts; they quietly shifted what question was being answered. No style-based detector can see criterion drift.

Risk: multi-agent voting certifies the wrong answer

Response: pin the criterion; define admissible evidence; allow the gate to block

LESSON 03

Review depth is not independence

More reviewers do not automatically improve quality. When each reviewer sees the prior approvals, the marginal benefit of each additional review drops toward zero. A visible PASS trail converts independent eyes into confirmation machines. The audit trail looks thorough. It is not.

Risk: approval chains with no independent scrutiny

Response: hide prior verdicts; route subtle correctness issues to mechanical tests

LESSON 04

Long-running autonomy rots without gates

In extended unsupervised operations, quality follows a phase structure: it rises, peaks, then degrades. The degradation is not random — it is a predictable consequence of missing review edges. Adding a single inter-step review edge produced a 99% quality improvement in the experiment. Without it, quality rot goes undetected as activity continues.

Risk: high-volume output that requires expensive remediation

Response: bounded loops with review edges, not unattended overnight runs

LESSON 05

Memory needs correction, not just storage

A fact can reach every part of a system while becoming progressively less accurate. First-write-wins propagation locks in early corruptions — later corrections cannot overwrite them. Coverage and fidelity are independent. A fact can be everywhere and wrong simultaneously.

Risk: organization-wide false records spreading through handoffs

Response: versioned facts with provenance; allow newer evidence to heal older claims

LESSON 06

Shared names are not shared meaning

Two agents or systems can use identical field names while referring to different concepts or units. "Total" might mean pre-tax or post-tax. "Created" might be milliseconds or seconds. The wire format is byte-identical; the semantic meaning is incompatible. Agents reach agreement fast — and wrong — with 100% reported confidence.

Risk: silent data corruption in handoffs and integrations

Response: semantic contracts with explicit concept, unit, and explicit refusal on mismatch

LESSON 07

Reports are not receipts

An agent asserting "done" is not evidence that anything changed in the real world. Self-reported completion is the weakest possible evidence tier. External receipts — transaction IDs, diffs, sent-message confirmations, audit entries, external system state — are categorically different from an agent's own account of what happened.

Risk: workflows report completion without external verification

Response: require external receipts for high-value actions before closing any task

Principles for Reliable Agentic Systems

01 **Make true signals cheaper to inspect than green signals.**

02 **Pin the decision criterion before deliberation begins.**

03 **Keep reviews structurally independent — hide prior verdicts.**

04 **Build bounded loops, not unattended sprawl.**

05 **Treat memory as versioned, verified evidence — not a notes file.**

06 **Carry meaning in semantic contracts, not prose descriptions.**

07 **Attach external receipts to claims of completion.**

08 **Let the system block when evidence is insufficient — never force a winner.**

Readiness Checklist

Use this before deploying any AI agent workflow into operations.

OUTCOME EVIDENCE

- [ ] What does success mean in the world — not in the agent's report?
- [ ] What external artifact proves it happened?
- [ ] Can the workflow report success without producing that artifact?

Green Is Not Correct · SwarmLab Client Briefing

DECISION GOVERNANCE

- [ ] Is the decision criterion written down before the process starts?
- [ ] What evidence is defined as admissible?
- [ ] Is the system permitted to block when evidence is insufficient?

REVIEW DESIGN

- [ ] Are reviewers shielded from prior verdicts?
- [ ] Which issues are routed to mechanical tests vs. judgment-based review?
- [ ] Is review depth proportional to the stakes and subtlety of the task?

HANDOFF INTEGRITY

- [ ] What concepts and units are passed between agents or systems?
- [ ] Is each field's meaning explicit — not inferred from the name?
- [ ] Can a receiver detect missing requirements? Changed meaning?

MEMORY RELIABILITY

- [ ] Where did each stored claim come from?
- [ ] What is its verification tier?
- [ ] Can newer verified evidence correct or supersede older claims?

ACTION RECEIPTS

- [ ] Does "complete" link to an external receipt?
- [ ] Is the receipt independent of the agent's own report?
- [ ] Is the evidence tier stored alongside the claim?

How heybeaux Uses This

SwarmLab is not theoretical. It directly informs how we design, evaluate, and govern agent systems. The experiments produced concrete changes: typed semantic contracts for handoffs, pinned criterion gates for multi-agent decisions, versioned memory with anti-entropy healing, receipt-backed completion, and bounded loops with mandatory review edges.

Each change was implemented, retested against the original experiment conditions, and measured against verifiable ground truth. Findings that produced red results are reported as red.

The standard: when an agent says "done," there is an external receipt. When a gate says "pass," there is evidence. When memory says "fact," there is a provenance chain. The goal is not agents incapable of failure — it is failure that is visible, recoverable, and cheaper than false confidence.

If you are planning an AI workflow, we can map the false-green risks before launch — the success signals, handoff boundaries, receipts, memory policies, and review gates that make the system reliable enough to trust.