Issue No. 002 NEWSLETTER June 2026

The Broken Feedback Loop

How AI removes the evolutionary correction mechanism that kept systems alive

THE ARGUMENT

Systems that survive don't avoid failure. They detect it early enough to correct it. Pain, price signals, unrest, exodus, these aren't symptoms of dysfunction. They're a feedback layer that converts a small failure into something the system can act on before it becomes terminal. The biological case is documented directly rather than by analogy: people with a rare congenital condition that leaves them completely insensitive to pain have an apparently normal nervous system, normal intelligence, and good health, apart from accidents and undetected injuries or infections (Cox et al., 2006). Markets run on the same logic one level up: prices function as an emergent information-aggregation mechanism, coordinating economic activity across participants who each hold only local, partial knowledge, with every price change signaling a shift in someone's circumstances elsewhere in the market, to which others need to adapt in turn (Hayek, 1945). Institutions run on a third version of it: when members of an organization perceive a decline in quality, they respond either by exiting or by voicing complaint (Hirschman, 1970). This isn't Darwinian selection, though it borrows its name. Selection needs nothing to be felt, a species can go extinct without ever sensing it. It works across generations, blindly, through differential survival alone. The feedback layer is faster: it operates within a single lifetime, a single business cycle, a single election. It exists because organisms and institutions that had it once outcompeted ones that didn't. It's downstream of selection, not the same thing.

Optimization systems are built to find the most efficient path to a target, which means they're also built to treat any deviation from that path as noise to be minimized, including the deviation that would have served as a warning. This is the layer AI puts at risk. Minsky (1992/1993) named the underlying mechanism decades before AI existed: stability breeds the very instability it seems to disprove, because the system's own reactions amplify small movements instead of damping them. Good times make lenders and borrowers progressively bolder, until the financing holding everything up depends entirely on conditions that never stop improving (Minsky, 1992/1993). Put plainly: calm doesn't just allow this drift, it causes it, every year without a crisis reads as proof the model is safe, which licenses a bigger bet the following year. 2008 ran this exactly, inside the risk models themselves. The Gaussian Copula model used to price mortgage-backed securities match the correlation and the individual default probabilities of the underlying bonds — but matching those two figures alone left the

Charles Wilform, PhD, MBA

likelihood of extreme, simultaneous defaults almost completely unconstrained. Two models with identical correlation and identical individual default risk could disagree by a factor of five on how often the worst-case event happens (Donnelly & Embrechts, 2010). The model wasn't sabotaged. It was optimized against the metrics everyone agreed to track, correlation, individual default probability; until those metrics had quietly defined the real risk, the tail, out of the picture entirely. This isn't a warning about AI replacing labor. It's a structural claim: a system loses its capacity to be corrected by its own failures exactly when the optimization making it look healthiest is the thing quietly removing the signal that would have caught it.

THE DARWINIAN FRAME

Darwin's insight was not about survival of the strongest. It was about survival of the most responsive. The organism that detects environmental change and adjusts has an advantage over the organism that cannot sense the signal at all. Fitness is a feedback function, not a static attribute.

Stuart Kauffman extended this logic to complex adaptive systems — organizations, economies, ecosystems. His models proposed that evolution maximizes adaptive capacity near a boundary between order and chaos, where fully ordered systems are rigid, fully chaotic systems are incoherent, and the most adaptive systems sit at the edge, with variation preserved and feedback loops intact (Kauffman, 1993). The claim is influential but not uncontested: when Mitchell, Hraber, and Crutchfield (1993) directly tested the related hypothesis that computational capacity peaks at this boundary, their results diverged sharply from the original findings, leading them to conclude the original interpretation did not hold. The structural claim survives the caveat even so — a system needs enough order to hold together and enough disorder to keep learning — but it's a working hypothesis, not an established law.

Natural selection is a mechanism for accumulating information about the environment. When that mechanism is disabled, the system stops learning. It does not stop operating. It stops learning.

Gould and Lewontin's spandrels paper add a critical nuance: not every structure in a complex system exists because it was selected for; some are necessary architectural byproducts of an unrelated design decision, not adaptations chosen for their own utility (Gould & Lewontin, 1979). Organizations accumulate analogous structures — processes, hierarchies, approval chains that were never adaptive; they were never tested. They persist only because nothing forced them to justify themselves. AI-enabled systems amplify this dynamic. They can sustain such structures indefinitely, because the friction that would have exposed them as costly is automated away.

Fitness is a feedback function. When the feedback stops, fitness becomes a memory — not a measurement.

HOW AI BREAKS THE LOOP

The evolutionary correction mechanism operates through four stages: variation generates options, selection pressure eliminates non-viable options, feedback encodes the lesson, and adaptation updates the system. The earlier sections of this piece focused on one of these stages — the feedback layer. This section widens the claim: AI intervenes at every stage, and in each case the intervention is marketed as an improvement.

Stage	What AI does to it
Variation	Models are trained on a necessarily finite slice of the world, and rare conditions are structurally underrepresented in that slice simply because training data can never capture every real-world possibility (Holder et al., 2023). Algorithmic optimization converges on what the training distribution already rewards. Edge cases, experiments, and anomalies — the raw material of adaptation — are filtered out as noise before they can generate information.
Selection Pressure	Automation absorbs costs that would otherwise register as pain. This is a mechanized, faster version of a well-documented organizational pattern: repeated deviations that don't produce immediate catastrophe get gradually reclassified as acceptable, not through any single decision but through the institutional structures meant to catch them quietly losing the capacity to do so (Vaughan, 1996). Inefficiencies that should trigger structural review are instead handled by adding a process layer. The system survives its own failures.
Feedback Encoding	Decades before generative AI, Bainbridge (1983) identified the core irony of automation: the more reliable a system becomes, the less its human operators practice the skills they'd need for the rare moments when manual judgment is required, because automation has removed the everyday occasions to use those skills. AI-managed systems extend this directly — outputs are produced without human exposure to the conditions that generated them. Operators lose tacit knowledge. The institutional memory of what it felt like to fail erodes by disuse, exactly as Bainbridge predicted four decades before the systems doing the eroding looked anything like AI.
Adaptation	When correction finally becomes unavoidable, the system lacks the vocabulary to respond. This is the documented signature of brittleness: a system performs well right up to its boundary, then collapses rapidly rather than degrading gracefully, precisely because it never built or practiced the capacity to extend itself past that boundary (Woods, 2015). It has no recent experience of structured failure. It has no practiced pathway for recovery. It fails discontinuously.

ASHBY'S LAW AND THE REQUISITE VARIETY PROBLEM

W. Ross Ashby formalized this dynamic in 1956 with the Law of Requisite Variety: a regulator can only control a system if it possesses at least as much variety, distinguishable states, response capacity as the disturbances it's trying to manage (Ashby, 1956). The often-quoted summary is Ashby's own: only variety can destroy, or absorb, variety. A controller with too few possible responses cannot guarantee stability against an environment that can generate many different kinds of disturbance.

AI-managed systems appear to increase variety. They process more inputs, respond faster, handle more edge cases. But they reduce variety in the one dimension that matters for evolutionary correction: the variety of human responses to novel failure conditions. The operators become less varied. Their exposure to the system's failure modes narrows. Their capacity to respond to conditions the AI was not trained on — the definition of a novel disturbance, atrophies.

Nassim Taleb's work gives this outcome a name, though the name needs to be stated precisely. Taleb's framework has three categories, not two: fragile, robust, and antifragile (Taleb, 2012). Fragility is harmed by volatility; robustness merely withstands it without change; antifragility is the one that *gains* from it. Fragility's true opposite in Taleb's own system isn't robustness, it's antifragility. What matters for this argument is a different part of his claim: depriving any complex system of the small,

regular volatility it would otherwise face causes it to weaken, and eventually produces a larger, more cascading instability than the small disturbances it was shielded from ever would have (Taleb & Douady, 2012), the same mechanism Minsky described for financial stability, restated as a general property of complex systems rather than a claim specific to credit markets.

That's the precise risk in optimizing operators out of regular exposure to failure. The system has been tuned to perform under known conditions. Unknown conditions, which are the only conditions that matter for survival over time, find it not merely unprepared but structurally deprived of the very stressors that would have kept it prepared.

THE GREEK PARALLEL

Plato's cycle of political decay in the Republic runs on a related but more specific logic than simple feedback loss. Each regime falls because the value it organizes itself around eventually consumes its own restraints (Plato, trans. 1992, 545a–569c). The timocracy prizes honor until its rulers start secretly craving wealth they can't admit to wanting — and the concealment becomes the seed of oligarchy. The oligarchy prizes wealth until rich and poor become, in Plato's own phrase, two cities sharing one place, permanently scheming against each other rather than ruling as one. The democracy prizes freedom until freedom answers to nothing, and that internal anarchy is what a demagogue exploits to become a tyrant.

What unites all three falls isn't a disabled sensor. It's narrower: each regime is undone by the very thing it built its legitimacy on, pursued past the point of self-discipline. The value worked exactly as advertised, it was just taken further than the system could survive.

The parallel to AI-managed institutions isn't decorative, but it sits at a specific point. Plato's regimes don't fail because they stopped feeling pressure. They fail because the value they over-pursued became indistinguishable from the disorder it produced, and nothing inside the system was positioned to call that disorder by its name before it was irreversible. That's the same shape as an optimization system smoothing away a deviation by treating it as more of the same successful pattern. The mechanism differs, Plato's is moral, ours is computational — but the outcome doesn't. A system that can no longer recognize its own failure as failure will not feel the pressure building. The collapse will look sudden. It will not have been sudden. The capacity to name the problem was lost long before the problem was visible.

WHAT THIS MEANS

The question is not whether to use AI. The question is whether the deployment of AI preserves the feedback mechanisms that generate adaptation. Three diagnostics follow from the evolutionary frame — the first two restate this newsletter's argument as a test; the third names a related risk this piece hasn't fully argued for, but that follows naturally from it.

Diagnostic	Question
Signal Preservation	Are the failure signals AI is smoothing still being recorded and reviewed? Efficiency gains that eliminate data are evolutionary liabilities.
Variation Maintenance	Is the system maintaining deliberate variation — processes that are not optimized, experiments allowed to fail, operators with unmediated exposure to system conditions?
Feedback Latency	How long does it take for a novel failure condition to reach a human with the authority and knowledge to respond? Every layer of automation extends that latency — a separate cost from losing the signal outright, but one that compounds it.

Revised June 30, 2026

Charles Wilform, PhD, MBA

References

Ashby, W. R. (1956). *An introduction to cybernetics*. Chapman & Hall.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Cox, J. J., Reimann, F., Nicholas, A. K., Thornton, G., Roberts, E., Springell, K., Karbani, G., Jafri, H., Mannan, J., Raashid, Y., Al-Gazali, L., Hamamy, H., Valente, E. M., Gorman, S., Williams, R., McHale, D. P., Wood, J. N., Gribble, F. M., & Woods, C. G. (2006). *An SCN9A channelopathy causes congenital inability to experience pain*. Nature, 444(7121), 894–898. https://doi.org/10.1038/nature05413

Donnelly, C., & Embrechts, P. (2010). *The devil is in the tails: Actuarial mathematics and the subprime mortgage crisis*. ASTIN Bulletin, 40(1), 1–33. https://doi.org/10.2143/AST.40.1.2049222

Gould, S. J., & Lewontin, R. C. (1979). *The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme*. Proceedings of the Royal Society of London. Series B, Biological Sciences, 205(1161), 581–598. https://doi.org/10.1098/rspb.1979.0086

Hayek, F. A. (1945). *The use of knowledge in society*. American Economic Review, 35(4), 519–530. https://oll.libertyfund.org/titles/hayek-the-use-of-knowledge-in-society-1945

Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard University Press.

Holder, C. J., Khonji, M., Dias, J., & Shafique, M. (2023). *Building resilience to out-of-distribution visual data via input optimization and model finetuning*. *IEEE Access*. https://arxiv.org/pdf/2211.16228

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford University Press.

Minsky, H. P. (1993). *The financial instability hypothesis*. In P. Arestis & M. Sawyer (Eds.), Handbook of radical political economy (pp. 153–158). Edward Elgar. (Original work published as Working Paper No. 74, Levy Economics Institute, 1992). https://www.levyinstitute.org/pubs/wp74.pdf

Mitchell, M., Hraber, P. T., & Crutchfield, J. P. (1993). *Revisiting the edge of chaos: Evolving cellular automata to perform computations*. Complex Systems, 7, 89–130.

Plato. (1992). *Republic* (G. M. A. Grube, Trans.; C. D. C. Reeve, Rev.). Hackett Publishing. (Original work composed c. 375 BCE)

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder*. Random House.

Taleb, N. N., & Douady, R. (2012). *Mathematical definition, mapping, and detection of (anti)fragility*. *Quantitative Finance*, *13*(11), 1677–1689. https://doi.org/10.1080/14697688.2013.800219

Vaughan, D. (1996). The Challenger launch decision: Risky technology, culture, and deviance at NASA. University of Chicago Press.

Woods, D. D. (2015). *Four concepts for resilience and the implications for the future of resilience engineering*. *Reliability Engineering & System Safety*, 141, 5–9. https://doi.org/10.1016/j.ress.2015.03.018