The Potentiality Field

Attribution and Observability in Non-Deterministic Agentic Systems

---

> *"Give me a place to stand, and a lever long enough, and I shall move the world."*
>
> *Archimedes*

§1. The Potentiality Field

An engineering team developing an agentic system makes a change to fix a defect, runs the evaluation again, and the score moves. This is the moment the trouble begins, though it rarely announces itself as trouble. The number went up, or it went down, and the natural next act is to reach for the change that was just made and credit it or blame it. The reach is a reflex, and the reflex assumes a thing that is not true: that in a system like this, a moved number has a single owner.

Three things can move the number, and on any given day more than one of them has. There is the change the team made, in the part of the system they own and can inspect at will, near at hand. There is the model, the light source they neither own nor can see into, which may be re-routed or re-versioned with no notice given. And there is chance, the same input resolving differently on two runs for no reason that exists to be found, because that is simply what these systems do. This third issue is a land mine. It is the one that sends a careful developer chasing a cause that was never there, hardening the system against a statistical wobble disguised as a structural defect, potentially creating more sources for anomalies. The three arrive knotted into a single figure on a dashboard, and the figure does not say which of them, or in what measure, produced it. This is the tanglement, the ordinary daily condition of building on non-determinism: several causes fused into one effect, with no seam visible to pull.

The knot tightens when the team goes looking for the vendor's hand in it. A change to the model overhead is a driver regardless of anyone being told, and the silence around it is the ordinary condition, not the exception. But the evidence a team reaches for is almost always the wrong kind. Someone notices that the provider shipped a new version the same week the score slipped, and the coincidence hardens into a story. Proximity in time is not attribution. It is a guess disguised as a finding. To attribute a movement to the vendor honestly, one must capture the identity of what answered the call and carry it into the decision matrix, not read it off a changelog and a calendar. The uninformed operator reaches for the nearest change in time, but the potentiality field does not respect chronology.

Set against all of this is the quiet possibility that runs through the whole practice. A developer working this way may not be fixing bugs. They may be seeding them. The

change that raised the score in the cases anyone looked at may have lowered it in a dozen cases no one thought to look at, and the run that would have exposed the harm did not happen to draw those cases. And here lies the uncomfortable core. No one can be sure the change did not cause more problems, and the one thing that can be said with confidence is that the developer is not sure either. This is not a catastrophe, and the claim is not that the system is broken. It is something stranger and harder to sit with: that no one can tell whether the change helped or harmed, and that working faster and running more tests does not make the knowing arrive.

It helps to stop picturing the system as a thing that has a behavior and to start picturing it as a thing that has a field of them. A non-deterministic system does not sit at a single point where a change nudges a neighboring point. It occupies a wide space of behaviors it could realize, a potentiality field, and every run is one draw from that space. A change does not move the system from one behavior to another. It reshapes the field, quietly, across regions outside the suite of run samples. When a defect finally surfaces, it is not a new fact arriving from nowhere. It is a draw from a region of the field that was always reachable and simply never lit. Today that defect is handled as a context-free point, a fresh ticket with no lineage, and the team's best hope is that whoever last touched the relevant part remembers touching it, wrote down what they changed, and that someone thinks to read the note. What is missing, named here only by its absence, is a defect that arrives with coordinates. Charting does not light the dark; the unsampled regions stay unsampled. What changes is narrower and more honest: nothing comes from the darkness anonymous. A defect, when it surfaces, arrives trailing the record of where it was drawn from, placed against everything seen before it, recognized as a return to charted ground or located as a new point upon it, rather than as a context-free surprise.

It is tempting to believe that more data closes this gap. It does not. Running more cases samples more of the same lit region and tightens the team's confidence around the tangled number, which only feels like progress. Confidence in a figure that three drivers produced is confidence in a figure that still means three things at once. Breadth illuminates more of what was already visible; it never reaches into the dark.

There is one more factor at play, and it is the subtlest, because it is the instrument teams increasingly reach for to see the others. The now-standard way to evaluate one of these systems at scale is to have another model grade its outputs, and that grader is itself a non-deterministic model, subject to the very drift it is meant to detect. This is not a separate problem so much as the same one reached from another direction: the instrument a team uses to read the field belongs to the field. Like the first model, it is a source to be captured, attributed, and calibrated over time. The potentiality field and a shifting ruler together create a particular unease, felt by any team in this situation: having changed something, watching a number move, unable to say what moved with it.

§2. Two Fallacies

Faced with the tanglement, a team does not usually despair. It adapts, leaning on two beliefs that carry the weight of seasoned engineering judgment and are, in this setting, false. They are defined here, turning a vague sense of futility into a pair of specific claims to be examined and, in the end, refuted. Each belief fails for a reason that belongs to non-determinism, each driving false convergence.

The first is the belief that this is simply how engineering works: that the effects of a change can be traced and contained, as they always have been. Call this the deterministic-transfer fallacy, because it carries a habit that is sound in deterministic engineering, unexamined, into a place where it does not survive.

The transfer is not a failure of intelligence, and it should not be described as one. The instincts it carries were earned. A generation of engineers learned, correctly, that a system is knowable, that its faults can be reproduced and cornered, and that discipline consists in mapping the consequences of a change before shipping it. Those lessons built nearly everything that works. The error is not in holding them. It is in assuming they survive a move into a setting where the same input does not produce the same output, and where the map has no fixed territory to describe. The fallacy is a good instinct pointed at the wrong world.

The assumption in question is that the blast radius of a change can be mapped. In a deterministic system this is very nearly true, and decades of good practice rest on it. A change to one function alters a knowable set of downstream behaviors, and a side effect, if it exists, can be reached by a test written to catch it, because the surface of what might break can be enumerated and walked. This is what gives a regression suite its meaning: the same code on the same input fails the same way, so a test that passes today is a promise that holds tomorrow. Non-determinism dissolves the promise. A change to a prompt, a tool, or a retrieved document does not alter a fixed set of downstream behaviors; it reshapes the field, and the field rearranges between draws. A developer can edit a single skill and silently break an unrelated response three cases away. A failure to write the appropriate test case is not the fault of the developer, but inherent, because the broken case was never drawn. Worse, even if they happen to test that exact case, a solitary pass is a false promise, a single draw from a region that remains volatile. The regression suite becomes a theater of safety. The blast radius here is not merely large. It is unmappable, since there is no fixed map on which to mark it, and every blind fix ships without sight of what it moved, silently reshaping the field. The unobserved region stays dark, and the team learns to call the darkness noise.

The second belief is quieter, and it tends to arrive only after the first has done its work. It is the conclusion a team settles into once it has accepted the field as a kind of weather: the variance is irreducible, the tanglement is just the texture of these systems, and the honest response is to stop trying to take it apart and learn to live with it. This is the

inseparability fallacy, and it is the load-bearing falsehood of the whole practice, the belief that quietly converts a hard measurement problem into an impossible one. It is the claim this paper refuses. This paper establishes the conditions under which the tanglement comes apart, and stops there. How much of it comes apart is a separate question, one of measurement rather than reasoning.

Precision is the key to resolving what the team is looking at when determining the tanglement is permanent. What a developer experiences as a tanglement, three drivers knotted into a single moved number, is not a mystery peculiar to language models. It is a shape the empirical sciences have lived with for a century and given a plain name. A confound: two or more sources of variation that move together in the data, such that the effect of any one of them cannot be determined until the others are accounted for. And a confound, Gordian as it may be, is not a wall. It is a problem with known handling. Variation that moves together can be pulled apart when its sources are held still one at a time, or arranged so that their contributions can be distinguished, and the reasoning that allows this is old, well understood, and not in dispute. The tanglement was never irreducible. It was unaddressed, because the observational position does not provide the perspective necessary to separate the sources.

This paper is precise about what it claims, and what it does not. It does not claim that the noise can be removed, or that a non-deterministic system can be made to hold still, or that chance can be argued out of existence. The field remains a field. The claim is narrower: the variance has structure, the structure can be recovered, and a large part of what the team has been calling irreducible noise is attributable variation it never had the position to separate. The unobserved region does not shrink. It is charted, one recorded observation at a time, until a fresh draw arrives already recognizable as coming from a region that has been seen before. This is characterization, not elimination, and the distance between those two words is the distinction between a discipline resigned to its noise and one that has learned to read it.

§3. The Familiar Fixes

A team that has begun accepting the tanglement rarely does so all at once. Before it settles into the second fallacy, it works the problem, and it works it with the tools it already trusts. What follows is not a catalogue of mistakes. It is the sequence a capable engineering team would move through, in roughly the order it would do it, each step reasonable and prudent. This surfaces an important pattern: not one of these moves is foolish, not one of them separates the tanglement, and they all miss it for reasons from the same source.

The first move is to achieve visibility. When a system misbehaves and no one can see inside it, the reflex of every mature team is to instrument it: wire in tracing, stand up an observability stack, put a probe on everything that moves. The instinct is sound, and in most of software it is the right first step. Here it runs into two walls. The first is that

instrumentation embedded in the system moves with the system. A probe built into the substrate is versioned along with it, so when the substrate changes the probe changes too, and there is no longer a fixed vantage from which to read what moved. An instrument that travels with the thing it measures cannot hold still long enough to establish a baseline. The second wall is plainer and harder. One can only embed a probe in code one controls. A team evaluating a third-party agent or working inside a regulated or air-gapped environment cannot reach in to place the instrument at all. For a large and growing class of the systems that most need this, the first move is simply not available.

The second move is to reproduce the problem, and it is the most seductive of the four, because it looks exactly like control. Pin the model version, pin the code to a commit, re-run the case, and hold everything fixed until the behavior returns the same way twice. In deterministic software this is the whole game. Here it fails quietly, because a commit hash does not pin the thing that actually moved. The substrate is not a single artifact. It is the prompt and the tool definitions and the retrieval index and the accumulated state the system has gathered along the way, elements which, in the interest of resolving the tanglement, must be considered strictly separate from the model's non-deterministic light. Any of these substrate columns can drift underneath a frozen commit without changing a tracked line of code. A team can believe it has held everything still while it is measuring against a substrate that shifted when no one was watching. The version is pinned. The substrate is not.

The third move is to out-measure the noise. If one run is unreliable, run the case a hundred times and take the average; if the average is unsteady, gather more cases still. This is the statistical instinct, and it is not wrong so much as aimed at the wrong quantity. More runs of the same kind sample more of the region already lit, and they tighten the team's confidence around the number it already had. But that number is the tangled one, produced by three drivers together, and a sharper estimate of a confounded quantity is still an estimate of a confounded quantity. Breadth buys precision about what was already visible and buys nothing about what was not. Telemetry collected without a stationary frame merely measures the uncertainty more precisely. The deeper trouble is that averaging is not free of time. A hundred runs, or a thousand, are gathered across a window, and across that window the light source can move on its own, sometimes mid-evaluation, so the figure the team ends with is not a steady reading. The longer the window, the more assured the number looks, because the interval around it keeps narrowing while the confound it carries goes untouched. And when a longitudinal number finally drifts far enough to alarm someone, the team does what experience has taught them: reach for the nearest dated event, a release note, a version bump, a status page, sitting near the week the average turned, and the nearness in time feels like an answer. It is not one. Nothing in the accumulated runs records what actually answered each call, so the team is reading a calendar rather than a cause. It is the same mistake as before, proximity taken for attribution, only now with far more data standing behind it.

The fourth move belongs to teams already far enough along to suspect their own instrument, and it presupposes that the earlier moves have already been made. If a model is grading the outputs, perhaps the grader has drifted, so calibrate it: pin it to a reference set, check it for agreement, correct it where it strays. This is careful work and provides value to the overall system efficacy. It also does not reach the problem. Calibrating the grader is still calibrating a second, non-deterministic model in the system, so what has been added is a mirror, not a fixed point. And even a perfectly steady grader would leave the central question unanswered a significant number of times, because the grader stands downstream of the tanglement rather than inside it. Whether the substrate or the light source moved the score is a question that must be answered at the point of impact, and no amount of care spent on a downstream ruler can decode a cause it never witnessed.

The four moves fail for one reason wearing four coats. Not one of them establishes a fixed point to measure against. Not one separates the system from the model it runs on. Not one records where in the field an observation fell, so that the next one can be set beside it. Instrumentation moves with the substrate; the pin does not hold the substrate; averaging sharpens a confounded number; calibration polishes an instrument that was never the confound. In every case the field stays dark and the tanglement stays whole, and the team is returned to where it began, with a number that moved and no way to say what moved it, or moved with it. What has been missing through all of it is not effort, and it is not rigor. It is position: somewhere to stand from which the sources come apart.

§4. A Place to Stand

Archimedes asked for two things, and this paper thus far has named one of them. A place to stand is the first requirement, but positioning moves nothing on its own. What moves the Archimedean world is a lever, specifically a rigid lever. A lever that bends under load transmits no force; the push goes into the bending, and the world stays where it is. The vantage is the fulcrum, the instrument the lever. What follows requires a lever that does not flex when bearing down on the object it means to move.

Consider the vantage first, because it is the simpler half. The operator's change and the vendor's response can be read together, paired and live, in only one place: the aperture, where the system hands its request to the model and takes the answer back. Off that boundary, one side or the other falls out of view. From inside the agent, the vendor's reply has already resolved into a finished answer, past the moment its share could be told from the operator's. Once the call is done, the two survive only as separate records, the input in one place and the output in another. They can be correlated after the fact, but correlation after the fact is reconstruction, not observation, and the live pairing that briefly held them together is gone. Only at the aperture, in the instant of exposure, do the substrate's state and the light source's reply meet at a single point that can be read as one thing.

But the aperture is not enough, and this is the objection that undoes every instrument built on position alone. Every eval harness already sits near the call boundary. Standing there is common. Positioned there alone changes nothing, because the vantage is only a fulcrum without a rigid lever. What the vantage needs is an instrument that does not move when the thing it measures moves, and that requirement, not the position, is what makes observing the confound possible.

Here the logic forces a conclusion. An instrument embedded in the substrate is versioned with the substrate: when the thing under measurement changes, the instrument changes with it, and there is no fixed reading left to compare against. That is a lever bending under load. A fixture that lives inside the substrate drifts with the substrate for the same reason, so it cannot serve as the still sentinel against which a shift is detected; it shifts too. The conclusion is not a preference and not a matter of design taste. It is forced. The only way to obtain an instrument that holds still is to observe by position, to embed nothing in the system observed, and to hold the fixtures outside the substrate entirely. Position, no instrumentation, external fixtures: these are not three features to weigh against one another. They are the three conditions that together produce a reference that does not move when the object under observation moves.

It is the composition that matters, and it is the composition the position-only instruments miss. Position without separation leaves a probe that travels with the substrate or a fixture that drifts with it, flexing the frame. The three are load-bearing as a set. Remove any one and the lever bends. Hold all three, and for the first time there is something stationary against which a movement can be measured.

Nothing in this argument is statistical. It is entailment, resting on the plain fact that a thing versioned with the substrate cannot be a fixed reference for the substrate. What follows depends upon this, as attribution is measurement, and measurement is impossible against a ruler that moves with the measured object. Before a single number is computed, before variance is decomposed or a source is named, there must be a frame that holds still. Without it, no attribution is possible at all.

That is the whole of the first requirement: a rigid frame, established by position and composition rather than by instrumentation, holding still while the field it observes does not. The lever must also be long enough, and its length is the resolution of the data. With the frame fixed, what can be told apart reaches exactly as far as the observations reach, and no further, which is the honest boundary and the only one the instrument is entitled to claim. Set the instrument here, holding only these things, and what was a knot of moving parts becomes, at last, something a fixed eye can begin to resolve.

§5. What the Frame Makes Possible

With the frame held still, the tanglement ceases to present as a single knot. It was never one thing. It is two couplings that call for opposite treatments; a recurring error is to reach for one and apply it to both.

The first coupling is between the substrate and the light source, the agent and the model it runs on, fused in every output because the one can act only through the other. This coupling is not to be removed. The attempt sends teams averaging when they should be separating. It is to be characterized. The substrate's state is captured and held fixed. The identity of the light source is captured alongside it, so that a re-route or a version change becomes a recorded fact rather than a coincidence read off a calendar. The same fixed request is put to the model repeatedly, at varying intervals over time, such that the way the vendor's output moves while the operator's input holds still allows for direct observation rather than inference. Because the request never changes, the accumulated record of the vendor's replies to it becomes a reference the vendor never agreed to and cannot revoke. When a silent re-route or a quiet re-version shifts the replies, the shift stands against that fixed history, and because nothing on the operator's side moved, the movement is not the operator's. The vendor's silence stops being an obstacle: the fixture remembers what the changelog never mentioned. This directly answers the paper's initial question: when the number moved, did the change belong to the operator or the vendor?

The second coupling involves the evaluator, the model that grades the outputs, and it is handled in the opposite spirit. The evaluator is not exempt from measurement; it grades the output the substrate and light source produce together, so it reads the tanglement, not the substrate alone, and its own drift is one more thing that can move a score. It is not to be characterized and kept. It is accounted for: held to a known identity and a fixed version, arranged so that its variation can be separated from the substrate's rather than mistaken for it.

The idea that the substrate and the light source can be separated by holding one still and capturing the other, and that the evaluator's drift can be distinguished rather than confused for the system's, follows as a logical cascade from the frame. How far the separation goes, how much of the variance resolves into named sources and how much remains a floor that no arrangement drives to zero, is a question of measurement, and measurement is a burden of demonstration, not assertion. This paper asserts that, with a stationary frame, the sources become separable in principle, and what has been called irreducible noise stands revealed as attributable variation waiting to be measured.

§6. Nothing Exempt

The method holds only if nothing is exempt from it, the instrument least of all. It reads the field with a model of its own, because grading a non-deterministic output at scale

takes another model, the only thing that can read what crosses the call boundary, and that grader drifts like anything else built on a light source. This is not a crack in the method but the recursion it already accounts for: the instrument does not stand outside the field it measures. It takes its place inside it, reads its own drift alongside everything else, and reports the limit that place imposes rather than claiming an objectivity it does not have.

§7. One Instantiation

Everything to this point describes a class of instrument, not a product, and that class is not without its members. At least one implementation, described separately, already meets these requirements together rather than trading them against one another: a versioned representation of the substrate that holds still, the light source captured under a vendor-agnostic identity, position held at the aperture by a mediating gateway that embeds in nothing it observes.

One consequence follows from where such an instrument sits rather than from anything it computes. Because it observes by position and embeds nothing, it runs where instrumented tools cannot reach: on local hardware, inside regulated environments, behind an air gap, wherever the system to be measured is not one its observer is free to open.

This is not, however, a paper about any one instrument, and the argument does not rest on any. It is a paper about a blind spot in the code stacks and eval harnesses of agentic systems, and about the events that surface from it and are dismissed as noise. Those events are not noise. They are draws from an uncharted potentiality field, defects arriving without coordinates, a tanglement of substrate, light source, and chance read as a single confounded number. The frame that separates them is a matter of implementing the logical requirements. That such an instrument already exists is evidence these requirements are real and compatible.

§8. The Honest Instrument

Something opens once the sources come apart. An operator can finally answer the only question that mattered, whether the number moved because of their own hand or that of the vendor, and answer it with a record rather than a guess. A defect stops arriving as a context-free surprise and starts arriving with its lineage, a draw from a region of the field that has been seen before. All teams, regardless of industry regulation or air gap restrictions, get an evaluation they can run where they stand. And the field gains a name for the thing it has been living with, so that the next person to feel the tanglement has a word for it instead of a shrug.

None of this removes the darkness. The instrument should not pretend to. It cannot make a non-deterministic system deterministic, nor should it try; what it can offer is rarer than that. It will not mislead an operator about where the shadow fell, and it will say, out loud, when it is guessing. That should be the whole of its character: an instrument that flags its own uncertainty rather than papering over it and is the more trustworthy for doing so.

And it must be honest about what it cannot see. A change below the substrate resolution leaves no artifact to read, and the instrument that reports that limit, rather than inventing a result to fill it, refines that resolution rather than disguising it. The limit is not a defect to be buried. It is the shape of the field.

It comes down to the nature of light itself. The same light that creates the potentiality field casts its shadows. Shadows are inherent to light. The honest instrument is the one that shows both, calling the darkness by its name instead of noise.

§9. The Artificial Ceiling

The uncertainty this paper describes does more than frustrate measurement. It sets a ceiling on what gets built. A team that cannot tell whether its last change helped or harmed cannot push the system much further with confidence, such that the complexity it is willing to ship is capped by the complexity it can measure.

For most teams that ceiling is not yet in view. The agents in most enterprises today are simple enough that the tanglement stays small: a few tools, a shallow chain, a light source called once or twice per task. At that scale a team can hold the whole system in its collective head, and its judgment that the last change was an improvement is usually sound. Nothing here contradicts that. The harder truth is this: a team meets the ceiling precisely to the degree it succeeds, because every increment of capability it adds is another increment of tanglement beyond what it can measure. There is no doubt teams building at the frontier, with deep chains, complex tool dependencies, and multi-agent orchestration, have already encountered it and been unsettled by what it implies.

But this ceiling is not fixed. It rests upon the floor this paper means to lower, and effort spent lowering that floor raises the ceiling.