

Islands of Stability

Six falsifiable propositions on where complex systems work, and their first tests in AI agents and across the sciences

Robert Encarnação · Independent researcher · islandsofstability.com

Draft preprint, version 1, 30 September 2026. Not peer reviewed. It reports specification v2.0.1 of the theory, as kept at islandsofstability.com/theory, and the results recorded there on that date.

ABSTRACT

Most ways a complex system could be run do not work. The ones that do occupy small, bounded regions of the settings available, islands of stability, and staying inside one often takes continuous effort. This paper states that idea as six propositions precise enough to be shown wrong: sparsity, an evidence rule, exit statistics, cost, transit and an era ratio. Each names the observation that would kill it. It then reports the first tests. Every test was fixed and deposited with a DOI before it ran, drew its random seed from that DOI, and has its verdict reported as registered, including unresolved and void results. The main instrument, the Map, swept three local language-model agents across the number of tools they were given and the share of tool calls made to fail, in 18,720 runs. It found no upper tool edge between two and six tools. The working region is a slope along the failure axis, not a band with an edge, so the sparsity test came out restricted for two models and below resolution for the third. The evidence-rule test survived for all three, and a push revealed what fault-free success hid: after an injected failure the models gave up in 15 to 61 percent of runs. Tests in a lake model, the last ice age, the business cycles of 42 countries and measles in 138 countries returned survives, restricted and unresolved, and a test in the asteroid belt ran twice without a verdict that counts. No registered test has killed a proposition. The rules, data and code of every test are public.

Keywords: stability, resilience, metastability, first-passage time, pre-registration, falsifiability, AI agents, fault injection

1. Introduction

The phrase comes from nuclear physics. In the 1960s physicists predicted that superheavy nuclei near 114 protons and 184 neutrons would live far longer than their neighbors, an island in a sea of fast decay. The picture travels well beyond nuclei. A shallow lake can stay clear or turn turbid under the same nutrient load. Asteroids crowd some distances from the Sun and avoid others. A star burns fuel for most of its life on a narrow band of color and brightness. An AI agent completes its task under some combinations of tools and failures and not under others. In each case, among the many ways the system could be set, the ones that work sit in bounded regions, with instability around them, and the regions can shift.

None of this is new as a description. Holling (1973) separated resilience from stability and described ecosystems with several stability domains. May (1972) and Gardner and Ashby (1970) showed that large, densely connected systems are less likely to be stable. Lewontin (1969) asked what stability means at all. Kramers (1940) treated escape from a stable state as a first-passage problem. Scheffer and colleagues (2009) collected early-warning signals for critical transitions, and Boettiger and Hastings (2012) set limits on them. Engineers describe metastable failures in distributed systems (Bronson et al. 2021, Huang et al. 2022), and physicists measure basin stability (Menck et al. 2013).

What this paper adds is not the existence of islands. It is a public, versioned statement of six claims about them that can fail, an instrument that measures one reading of them in AI agents, and a ledger of tests across other sciences, each fixed in advance. Section 2 states the theory. Section 3 sets out the method: how the specification changes, how a test is frozen and deposited, and which verdicts are possible. Section 4 reports the Map, the first instrument, in AI agents. Section 5 reports the tests and descriptions in other sciences. Section 6 says what the evidence does and does not show, and Section 7 lists where every rule, dataset and line of code can be found.

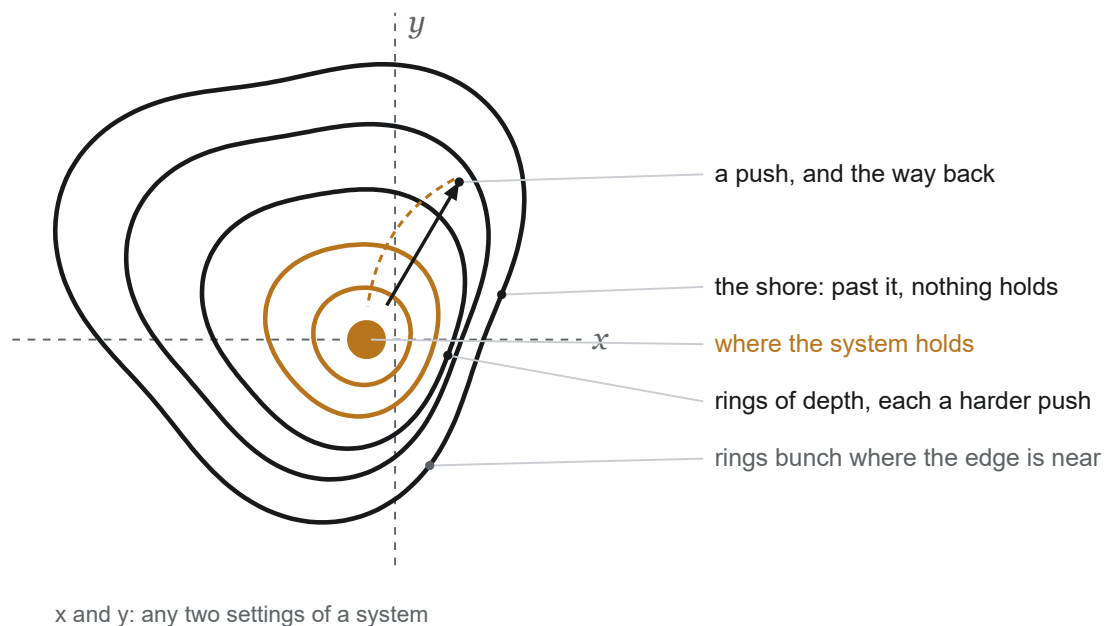


Figure 1. The picture behind the theory, drawn from the site's logo. The axes are any two settings of a system. The point is where the system holds, and each ring is a harder push away from it, out to the shore, past which nothing holds. The rings bunch where the edge is near. A push and the system's way back, the dashed path, are what show that an island is there (proposition 2).

2. The theory

2.1 Three senses of island

The word is used in three senses, and every object in this work uses exactly one and says which. A *chart-island* is Seaborg's sense: a region on a chart whose occupants outlast their neighbors, a convention about relative lifetime rather than a dynamical regime. A *basin-island* is an attractor with a sticky boundary in a dissipative system. A *grid-island* is a subsystem that keeps running when its connection to the whole is cut, the power engineer's islanding. Two rules govern their use. An island claim that states no cost line and no fragility line is a claim about a ghost and is not admitted. And a proposition moves between domains only when the mathematical or causal structure transfers, not merely the vocabulary.

2.2 The six propositions

Table 1 lists the six in plain words, with what would kill each and where it stands. Propositions 1 to 4 have an instrument path. Propositions 5 and 6 frame the others and are not yet testable. The full statements follow, quoted from specification v2.0.1.

Table 1. The six propositions, in plain words, with their status on 30 September 2026.

No.	Name	Plain meaning	Status
1	Sparsity	Working configurations are rare and bounded. Most of the space around them does not work.	measured once (AI agents)
2	The evidence rule	Looking stable is not evidence. Only recovery from a push shows an island is there.	measured once, tested
3	Exit statistics	For islands of the same depth, age says nothing about how long they have left.	tests run
4	Cost	Where stability is held by a flow, it is paid for continuously, and whatever pays for it becomes the weak point.	test run: unresolved
5	Transit	Survival is decided in the crossing, not on either island.	framing
6	The era ratio	If the world changes faster than you can reconfigure, no map helps: get faster, park on the widest island, or hibernate.	framing

Proposition 1, sparsity.

Among the configurations organizations actually attempt, the viable region is a small fraction, and it shrinks as size, coupling density and coupling strength grow. The remainder is sea, not a neighboring basin. In configuration space, sea means no viable regime exists at that setting. Scale bought without decoupling moves a program from island to sea.

Killed by: most attempted configurations viable. In the agent domain, a tools-squared term that is not negative, a peak outside the tested range, or a drop at the upper edge that stays inside the per-cell intervals.

The statement is made for organizations, and the principle is meant to hold across the sciences. The Map tests one narrower reading in one domain, AI agents, and its test does not decide the organizational reading.

Proposition 2, the evidence rule.

Persistence is not evidence of an island. A system can sit near a vanished attractor or a sticky boundary for years and look stable on every passive measure, and only a measured restoring response to a perturbation of its state shows that an island is there. Once an island is known to be there, watching its fluctuations can locate its edge as well as pushing it, unless the push is large against those fluctuations. Boundaries are observed before interiors, and every located island is a revisable estimate.

Killed by: persistence reliably predicting recovery. Passive statistics telling a system near a vanished attractor from one inside a live island as well as a perturbation does. For an island known to be there, a push that is small against the system's fluctuations locating its edge better than watching does.

Proposition 3, exit statistics.

An island's lifetime is a first-passage time. Under a stationary disturbance process, exit from an island of given depth is memoryless, so once its depth is known a configuration's age is no evidence of its remaining life. Islands that differ in depth show a hazard falling with age without any memory, and islands that end by exhausting a reserve rather than by disturbance age on schedule. Neither is a counterexample. Some exits are noise-induced with no precursor, which bounds any early-warning instrument's recall below one. Sudden per island, progressive across the archipelago.

Killed by: hazard falling with age conditional on depth. Dwell linear in depth. Every exit showing precursors.

The proposition is stated for agent systems. Its organizational reading was withdrawn after an objection that the site records: a death rate that falls with age may be frailty, the weak leaving first, rather than memory, and the two cannot be told apart without a measure of depth (Section 5.2).

Proposition 4, cost.

In a system held by a throughput, depth is a flow, not a stock. It is purchased by continuous station-keeping against drift and decays when the expenditure stops. The expenditure concentrates fragility in the stabilizer, and exits from engineered islands run through the stabilizer more often than through outside shock.

Killed by: depth persisting without expenditure in a system held by a throughput.

The second half is credited to its prior owners, not claimed (Carlson and Doyle 2002 on robustness and fragility, among others). What is claimed is the first half: what the stabilizer costs in margin, and how vulnerable a system decides to run.

Proposition 5, transit.

Survival is decided in the passage. A transit has a duration, a reserve burn rate, and a minimum invariant set that must be carried across for the arriving system to count as the same system. None of the three follows from the properties of the two islands.

Killed by: endpoint properties predicting transit success. Invariant preservation proving irrelevant.

Proposition 6, the era ratio.

Whether navigation is possible at all is set by the environment's regime dwell time divided by the organization's reconfiguration time. Above one, island choice and transit design matter. Near one, timing dominates. Below one, no navigation strategy works and three moves remain: shorten reconfiguration time, park on the widest reachable island, or hibernate. The bet that is this site's own: AI shrinks the numerator in AI-exposed sectors while review, compliance and hiring hold the denominator.

Killed by: outcomes independent of the ratio. Era proxies stable or lengthening.

Propositions 5 and 6 have no instrument in any domain and are not counted in any tally. Economics already owns much of their ground, from Milgrom and Roberts (1990) on complementary practices that form clusters with no incremental path between them to Minsky (1986) on stability that breeds instability, and Hamilton's (1989) regime-switching model is the candidate instrument for proposition 6.

3. Method

3.1 A versioned specification

The theory is kept as a versioned specification. Four channels move the version. A measurement moves the major version, whatever its verdict. An accepted external objection moves the minor version. A scope change drawn from the site's own evidence also moves the minor version and is labeled internal. An editorial change moves the patch version. A run that fails to kill a proposition is entered as a dated survival, because confirmation is an entry, not silence. Every proposition has a permanent anchor of the form v2.0-p3, earlier wording is kept where links and frozen tests cite it, and every change is entered in a changelog that is also published as a feed. The specification reported here is v2.0.1.

3.2 Tests fixed before they run

Every test follows the same sequence. Its rules are drafted, then approved by the author decision by decision. The rules and the code are frozen: every file the test depends on, data included, is recorded by its SHA-256 hash, and the test refuses to run if any file differs. The frozen bundle is deposited on Zenodo, which gives it a DOI. The test then runs once, with its random seed derived from that DOI, the first 16 hexadecimal digits of the DOI's SHA-256 hash read as an integer, so the

seed cannot be chosen after the rules are fixed. Its verdict is reported as the rules decide. Nothing is re-run to improve a verdict, and a run whose design fails is recorded as invalid or void, not replaced quietly.

The possible verdicts are few. A proposition *survives* a test when the prediction holds by the margin declared in advance. It is *restricted* when the prediction fails by more than the smallest effect worth reporting, without the observation that would kill it. It is *killed* when that observation appears. A result is *unresolved* when the interval is too wide to decide, and *below resolution* when the data cannot tell. A run is *void* when its own control fails, and *invalid* when its design is found faulty. Each test declares its smallest effect of interest before it runs.

3.3 What a test can show

Every test is narrower than the proposition it bears on. A test in one domain says nothing on its own about another, and the mapping from an instrument's variables to the theory's terms is an assumption, stated with each instrument, not a finding. Several tests are not blind: the published literature or the site's own descriptive pages had already shown their data. Where that is so, the test's report says so.

4. The Map: islands in AI agents

4.1 Design

The Map asks where an AI agent keeps doing its job. The task is invoice extraction: the agent reads a synthetic invoice through its tools and submits a record of its fields. No amounts are printed. Each line, the tax and the total must be computed from stated rules. Two dials are swept. The first is the number of tools the agent is given, from one to six, added in a fixed order. The one-tool column is out of scope, because the task needs two tools. The second is the share of tool calls made to fail, from 0 to 50 percent in six steps. Three local models were run on llama.cpp, one request at a time: gpt-oss-20b, Qwen3-14B and Qwen3-4B. Each cell holds 100 runs per tool order, and two tool orders separate the effect of a tool's count from the effect of which tool it is, for 18,720 runs in all. A further 18,000 noise-floor runs check that identical runs reproduce. Every stack passed a determinism gate, and every identical noise-floor rerun matched byte for byte. A run succeeds when its weighted field score reaches 0.8. Intervals are 95 percent Wilson intervals per cell and Newcombe intervals between cells.

The dials stand for the theory's terms by assumption. Tool count stands for the number of external systems an agent can reach, not for organizational size. The failure rate stands for how reliable those systems are, not for market volatility. The rules were confirmed by the author and deposited before any snapshot run, on 26 September 2026 (pre-registration r4, doi:10.5281/zenodo.22982434). The results bundle, 38,475 files, is deposited as doi:10.5281/zenodo.23066893, and a reviewer's command recomputes every number from the raw runs.

4.2 The two tests

Test 1, the upper tool edge, bears on proposition 1. A logistic regression of success on tools, tools squared and failure rate asks whether the curve bends down, with a peak inside two to six tools and a drop from the peak to six tools larger than chance. It survives if so, and it is restricted if that drop is shown to be under 15 points. Test 2, the recovery cost, bears on proposition 2. It injects a failure mid-run and measures the extra turns the agent spends, against unperturbed controls. The tests are scored on their own and bear on the propositions without deciding them.

4.3 Results



Figure 2. Snapshot 1 of the Map. Each cell is the percentage of runs that succeeded, for tool order A, by the number of tools given (columns) and the share of tool calls made to fail (rows, rising upward). Darker is higher. Success falls as failures rise, for every model, and does not fall at a tool edge. Data: the deposited snapshot (doi:10.5281/zenodo.23066893).

Table 2. The Map's verdicts, as pre-registered. Intervals are 95 percent.

Model	Test 1: upper tool edge	Test 2: recovery cost
gpt-oss-20b	restricted drop 0 points (−7 to 7)	survives +0.8 turns (0.6 to 1.1)
Qwen3-14B	restricted drop 0 points (−13 to 13)	survives +0.8 turns (0.6 to 1.0)
Qwen3-4B	below resolution peak at 5.7 tools (4.5 to no upper bound)	survives +0.3 turns (0.1 to 0.5)

No tool edge. For gpt-oss-20b and Qwen3-14B, a drop of 15 points or more at six tools is ruled out, so test 1 is restricted. For Qwen3-4B, success was still rising at six tools, so there was no peak to fall from, and the test is below resolution. In this task and range the working region is a slope along the failure axis, not a band on the tool axis. With no failures, gpt-oss-20b succeeded in 92 percent of runs, and with half the calls failing in 50 percent. Qwen3-14B fell from 38 to 21 percent and Qwen3-4B from 29 to 18 percent.

Which tool, not how many. Because the tools arrive in a fixed order, a change at one step could be the tool rather than the count. The second order adds the same tools differently. For gpt-oss-20b, the only model that could compute the total, the spreadsheet lowered success in both orders: by 6 points (1 to 11) as the fifth tool and by 10 points (5 to 15) as the third. Its numbers were right. The

trouble was when the model reached for it: in all 453 runs where gpt-oss-20b opened the spreadsheet, it did so right after the invoice had failed to load, and it took the spreadsheet's line items instead of asking for the invoice again. For the two Qwen models no step moved success beyond chance in either order. These comparisons were committed in writing before the results were looked at and decide nothing.

The push showed what success hid. A failure cost 0.3 to 0.8 extra turns, and net of asking again for the data that failed it cost nothing, so test 2 survives for all three models. But the push revealed a difference that fault-free success concealed. After an injected failure, Qwen3-4B stopped without submitting in 61 percent of runs, gpt-oss-20b in 33 percent and Qwen3-14B in 15 percent. Grit differed more between models than tool count did. This is proposition 2's claim in its smallest form: the response to a push, not persistence, shows what a system will do.

4.4 Limits of the Map

Snapshot 1 speaks about one task under one harness, not about agents in general. Three small local models are a replicate, not a population, and an exploratory trial of larger hosted models, outside the snapshot, found far bigger islands. The 0.8 threshold admits a record whose total is wrong, so the fully correct rate is reported beside it: 60 percent for gpt-oss-20b and none for the Qwen models. The prompt asks for one submission, and a failure on the submit tool forces a second, so the failure axis partly measures whether a model breaks that instruction to retry. gpt-oss-20b produced 24 replies the server could not parse, and counting them as failures changes neither verdict.

5. Tests and descriptions across the sciences

5.1 Registered tests

Table 3 lists every registered test outside the Map, with its verdict. Each was deposited before it ran, and each result is reported in full on its domain's page.

Table 3. Registered tests outside the Map, all run on 30 September 2026.

Domain and test	Bears on	Verdict	What was found
Lake model, E1: watching or pushing	P2	restricted	Watching was as good as pushing in 3 of 4 settings. Where pushing helped most, its error was 0.85 of watching's (0.75 to 0.99), where 0.8 or less was needed.
Lake model, E2: the ghost	P2	survives	Pushing told doomed lakes from safe ones better than watching by 14.7 points of balanced accuracy (12.3 to 17.0), where 10 were needed.
Lake model, E3a: memory	P3	survives	The Weibull shape of 2,000 lakes' exit times was 1.014 (0.980 to 1.049), inside the band of 0.9 to 1.1 that counts as memoryless.
Lake model, E3b: depth	P3	unresolved	Mean stays rose from 245 to 9,579 time units over depth ratios 3 to 7. Neither the linear nor the exponential form passed as the rules required.

Domain and test	Bears on	Verdict	What was found
Lake model, E3c: warning	P3	survives	At a 10 percent false-alarm rate the alarm caught 20.1 percent of 1,286 exits (18.0 to 22.3), where 50 percent or less was needed, and its positive control caught 88.6 percent.
Ice ages, C1: memoryless cold spells?	P3	survives	The Weibull shape of the 34 cold spells' lengths was 0.97 (0.74 to 1.22), where memoryless is 1.
Ice ages, C2: a 1,470-year clock?	P3	no clock	The 35 warmings did not cluster in phase at the declared period: Rayleigh $p = 0.71$, where a clock needed less than 0.01.
Economy, X1: do expansions age?	P3	unresolved	Over 142 ended expansions in 40 of 42 countries, the shape was 1.38 (1.15 to 1.63). Memoryless is ruled out ($p = 0.001$), but a departure needed the lower end above 1.25.
Measles, the runway	P4	unresolved	Over 276 outbreak-free spells in 138 countries, each percentage point of the population born unvaccinated since a spell began multiplied its chance of ending by 1.12 (1.02 to 1.27). Survival needed the lower end above 1.10.
Asteroid belt, stickiness, two runs	P3	invalid, then void	The first run was recorded as invalid. The corrected run was void, because its control did not show the stickiness the test needed before its result could count.

The lake tests use the standard model of a lake that can be clear or turbid (Carpenter, Ludwig and Brock 1999), so they test the propositions' logic in a model, not in a lake. The ice-age tests use the NGRIP oxygen-isotope record and its dated warmings (Seierstad et al. 2014, Rasmussen et al. 2014). They weigh a 1,470-year clock (Rahmstorf 2003) against noise-driven switching (Ditlevsen, Andersen and Svensson 2007). The expansion test compares each country only with itself, using business-cycle dates from the NBER and from the World Bank's and the Maddison Project's data. The measles test uses vaccination coverage and reported cases for every country with the records. Its interval lies above one, in the direction proposition 4 predicts, but not by the margin the rules asked for.

5.2 Described, not tested

Six domains are described on real data without a test, each for a stated reason.

Nuclei. The chart of every known nucleus, from the NUBASE2020 evaluation (Kondev et al. 2021), is the reference case, the chart-island the work is named after. A check of the whole table for errors a reader can find without another source turned up seven candidates, five of them errors, which are corrected before anything is counted.

Stars. Gaia's chart of 228,623 stars within 100 parsecs (Gaia Collaboration 2023) holds both kinds of island the theory distinguishes: the main sequence, paid for by burning fuel, and the white dwarfs, which cost nothing. There is no test by design. Its use is to show the cost line on real data.

Evolution. Van Valen (1973) found that a genus's chance of dying out does not depend on its age, which is proposition 3 fifty years early. In 23,191 marine genera of eight groups from the Paleobiology Database, seven of the eight survivorship curves bend toward a hazard falling with age, which may be frailty, memory or the record itself. Finnegan, Payne and Wang (2008) found the fall with age remains when geographic range and species richness are accounted for, which by the proposition's own terms counts against it in this domain. A re-run of that analysis under rules fixed in advance has been approved and is being prepared.

Firms. US Census data show young firms dying far more often than old ones. A mix of fragile and sturdy firms, none of them changing with age, fits every age class within 1.2 percentage points: 35 percent fragile, each with a 40 percent chance of dying each year, and the rest with 5.8 percent. That is the objection that withdrew proposition 3's organizational reading. A falling death rate alone cannot say whether firms grow safer or the fragile die first.

AI governance. Proposition 6 has its first real numbers here, as an illustration, not a test (Figure 3). Since 2010, 24 models have each set a new record for the computing power used to train an AI model (Epoch AI). A record held for a median of 132 days and never longer than 583. The EU's AI Act took 1,198 days from the Commission's proposal to its entry into force. Taking the first as the environment's dwell time and the second as the rule-makers' reconfiguration time, the ratio is about 0.11, well below one. The series a test of proposition 6 will use has been chosen, and its rules are being written.

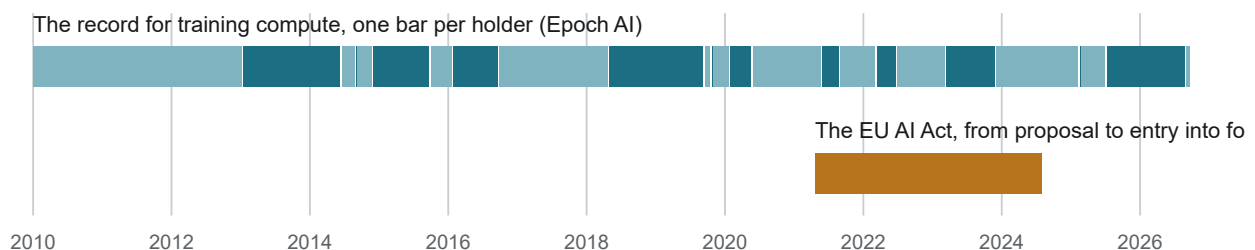


Figure 3. Two clocks. Above: each bar is one holder of the record for training compute, from Epoch AI's data on frontier models. Below: the EU AI Act from the Commission's proposal (21 April 2021) to its entry into force (1 August 2024). The ratio of the median record's tenure to the Act's span is about 0.11. An illustration of proposition 6, not a test of it.

Water. Supercooled water and diamond persist without being stable. Szakáll and colleagues (2021) froze droplets in a wind tunnel, 70 per run at one temperature, and the record holds 34 such runs for six materials. They classed every material as a mixture of particles that start ice more or less easily, a difference in depth already shown, so a bend in the curves could not count against proposition 3 and no test is drafted on them. A test needs the same droplets frozen again and again, with each droplet's own depth known. Such data may exist (Jakobsson et al. 2022).

6. Discussion

6.1 What the evidence shows so far

No registered test has killed a proposition, and none has confirmed one in the strong sense. Proposition 2, the evidence rule, has the most support: it survived the Map's recovery test for all three models and the lake's ghost test, and its one restriction, the lake's E1, concerns its narrower clause: for an island already known to be there, watching was as good as pushing in three of four settings, and the largest push helped less than predicted. Proposition 1's reading in agents was restricted: in the tested range there was no edge on the tool axis, only a slope along the failure axis. That is a result against the band-with-an-edge picture in this task and range, entered as a restriction, as the pre-registered rules decide. Proposition 3 survived where depth was held fixed by design, in the lake model's memory test and the ice-age cold spells, and was unresolved where it was not, in the expansions of 42 countries. Published evidence from fossil genera counts against it. Proposition 4's first test leaned its way and stayed unresolved.

6.2 Frailty and depth

A recurring problem runs through the results. A hazard that falls with age can come from memory, each island growing safer, or from frailty, a mix of shallow and deep islands where the shallow ones leave first. Firms, fossil genera and the ice-age warm spells all show falling hazards, and in none can the two be told apart without a measure of depth. Proposition 3 names differences in depth as an exception that must be shown, not assumed. The consequence is that it can be tested only where depth can be measured or held fixed: a model with known depth, a single island observed many times, or data that carry a defensible proxy, as geographic range and richness are meant to for fossil genera. This is the main constraint on where the theory can be tested at all.

6.3 The evidence rule in practice

For anyone who runs AI systems, the Map's clearest lesson is about evidence, not about tools. Fault-free success rates ranked the three models one way. The push ranked their behavior another way: the model that was second-best without failures gave up after a failure least often. A dashboard of steady success says little about how a system will respond when something it depends on fails, and fault injection, red-teaming and chaos testing are the kind of push that does. The theory's broader claims for organizations remain unmeasured, and the site labels each implication drawn from them as the implication of an unmeasured claim.

6.4 Limitations

The author designed every test, and none has been replicated by anyone else. Several were not blind, because the data or the published results were known. The Map rests on one task and three small local models. The asteroid-belt test failed twice in design before producing a verdict, which argues for an outside physicist's review before a third attempt. The theory is broad by intent, and its cross-domain rule limits how far any one result travels. This preprint has not been peer reviewed.

6.5 Next tests

Five tests are in hand. A second snapshot of the Map, with larger hosted models, can also ask whether its islands move between model versions. A long-horizon agent test of proposition 3 asks whether an agent's chance of failing at each step stays flat as a run goes on. The fossil-genus analysis is to be re-run under rules fixed in advance. Proposition 6 is to get its first test on a named public series. And a water test can be drafted if repeated-freezing data with each droplet's depth can be obtained.

7. Data and code availability

The specification, its changelog and every domain page with its data are at islandsofstability.com, under specification v2.0.1. The harness that ran the Map is public at github.com/delimiterbob/islands-harness. The frozen rules, code and results are deposited on Zenodo:

- The Map, pre-registration r4: doi:10.5281/zenodo.22982434
- The Map, snapshot 1 results bundle: doi:10.5281/zenodo.23066893
- Lake model tests E1 to E3: doi:10.5281/zenodo.23065462
- Ice-age tests C1 and C2: doi:10.5281/zenodo.23066235
- Expansions test X1: doi:10.5281/zenodo.23066598
- Asteroid-belt test, run 1: doi:10.5281/zenodo.23066015
- Asteroid-belt test, run 2: doi:10.5281/zenodo.23071492
- Measles runway test: doi:10.5281/zenodo.23071574

The data come from NUBASE2020 (CC BY 3.0), the ESA Gaia mission, the NASA/JPL Small-Body Database, the Centre for Ice and Climate at the University of Copenhagen, the NBER, the US Bureau of Economic Analysis through FRED, the World Bank's World Development Indicators, the Maddison Project Database, the Paleobiology Database, the US Census Bureau, the US National Center for Health Statistics, Our World in Data, Epoch AI, and Szakáll and colleagues' Zenodo record 10.5281/zenodo.4436153. Each file's checksum is recorded, and the code refuses any other file.

Acknowledgments

The site, its analyses and this draft were prepared with the help of Claude, an AI model made by Anthropic, under the author's direction. The author made every decision recorded in the pre-registrations and takes responsibility for the content and any errors.

How to cite

Encarnação, R. 2026. Islands of Stability: Six falsifiable propositions on where complex systems work, and their first tests in AI agents and across the sciences. Draft preprint, version 1, 30 September 2026. [bit&being.bitandbeing.com/research](https://bitandbeing.com/research).

References

- Boettiger, C., and Hastings, A. 2012. Quantifying limits to detection of early warning for critical transitions. *Journal of the Royal Society Interface* 9(75): 2527–2539. doi:10.1098/rsif.2012.0125
- Bronson, N., Aghayev, A., Charapko, A., and Zhu, T. 2021. Metastable failures in distributed systems. In *Proceedings of the Workshop on Hot Topics in Operating Systems (HotOS '21)*, 221–227. ACM. doi:10.1145/3458336.3465286
- Carlson, J. M., and Doyle, J. 2002. Complexity and robustness. *Proceedings of the National Academy of Sciences* 99(suppl. 1): 2538–2545. doi:10.1073/pnas.012582499
- Carpenter, S. R., Ludwig, D., and Brock, W. A. 1999. Management of eutrophication for lakes subject to potentially irreversible change. *Ecological Applications* 9(3): 751–771.
- Ditlevsen, P. D., Andersen, K. K., and Svensson, A. 2007. The DO-climate events are probably noise induced: statistical investigation of the claimed 1470 years cycle. *Climate of the Past* 3: 129–134.
- Finnegan, S., Payne, J. L., and Wang, S. C. 2008. The Red Queen revisited: reevaluating the age selectivity of Phanerozoic marine genus extinctions. *Paleobiology* 34(3): 318–341. doi:10.1666/07008.1
- Gaia Collaboration, Vallenari, A., et al. 2023. Gaia Data Release 3: summary of the content and survey properties. *Astronomy and Astrophysics* 674: A1.
- Gardner, M. R., and Ashby, W. R. 1970. Connectance of large dynamic (cybernetic) systems: critical values for stability. *Nature* 228: 784. doi:10.1038/228784a0
- Hamilton, J. D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57(2): 357–384. doi:10.2307/1912559
- Holling, C. S. 1973. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics* 4: 1–23. doi:10.1146/annurev.es.04.110173.000245
- Huang, L., et al. 2022. Metastable failures in the wild. In *Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI '22)*. USENIX Association.
- Jakobsson, J. K. F., Waman, D. B., Phillips, V. T. J., and Kristensen, T. B. 2022. Time dependence of heterogeneous ice nucleation by ambient aerosols: laboratory observations and a formulation for models. *Atmospheric Chemistry and Physics* 22: 6717–6748. doi:10.5194/acp-22-6717-2022
- Kondeev, F. G., Wang, M., Huang, W. J., Naimi, S., and Audi, G. 2021. The NUBASE2020 evaluation of nuclear physics properties. *Chinese Physics C* 45: 030001. doi:10.1088/1674-1137/abddae
- Kramers, H. A. 1940. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7(4): 284–304. doi:10.1016/S0031-8914(40)90098-2
- Lewontin, R. C. 1969. The meaning of stability. In *Diversity and Stability in Ecological Systems*, Brookhaven Symposia in Biology 22: 13–24.
- May, R. M. 1972. Will a large complex system be stable? *Nature* 238: 413–414. doi:10.1038/238413a0
- Menck, P. J., Heitzig, J., Marwan, N., and Kurths, J. 2013. How basin stability complements the linear-stability paradigm. *Nature Physics* 9: 89–92. doi:10.1038/nphys2516
- Milgrom, P., and Roberts, J. 1990. The economics of modern manufacturing: technology, strategy, and organization. *American Economic Review* 80(3): 511–528.
- Minsky, H. P. 1986. *Stabilizing an Unstable Economy*. Yale University Press.
- Rahmstorf, S. 2003. Timing of abrupt climate change: a precise clock. *Geophysical Research Letters* 30(10): 1510.

- Rasmussen, S. O., et al. 2014. A stratigraphic framework for abrupt climatic changes during the Last Glacial period based on three synchronized Greenland ice-core records. *Quaternary Science Reviews* 106: 14–28.
- Scheffer, M., et al. 2009. Early-warning signals for critical transitions. *Nature* 461: 53–59.
doi:10.1038/nature08227
- Seierstad, I. K., et al. 2014. Consistently dated records from the Greenland GRIP, GISP2 and NGRIP ice cores for the past 104 ka reveal regional millennial-scale $\delta^{18}\text{O}$ gradients with possible Heinrich event imprint. *Quaternary Science Reviews* 106: 29–46.
- Szakáll, M., Debertshäuser, M., Lackner, C. P., Mayer, A., Eppers, O., and Diehl, K. 2021. Comparative study on immersion freezing utilizing single-droplet levitation methods. *Atmospheric Chemistry and Physics* 21: 3289–3316. doi:10.5194/acp-21-3289-2021
- Van Valen, L. 1973. A new evolutionary law. *Evolutionary Theory* 1: 1–30.