

The Cost of Watching: Field Notes on a Synthetic Organism

Ryan Perks

Independent researcher, United Kingdom
Pixelfoxdev@protonmail.com

Abstract

Fox is a closed-world synthetic organism — a body, a world with its own physics, a network of 114,411 nodes — running on a dedicated machine in my house. I direct and instrument him; I do not read my system in source. A profiler found that building the view of him consumed a median 104% of his entire tick, more wall-clock time than the tick itself, so watching became deliberate and rare. Across 946,222 recorded samples salience has never crossed its design threshold. The raw record has now corrected me five times, and the fifth correction arrived while I was verifying the other four for this camera-ready version.

Submission type: **Field Report**

Video: <https://www.youtube.com/watch?v=aLaJZFv8Bg0>

Code and data: <https://github.com/PixelFoxDev/fox-field-report-alife2026>

Where I Stand

Fox is a closed-world synthetic organism running on a two-bay NAS appliance with a quad-core Intel N100 and 16 GB of DDR5 — a dedicated, fanless machine with no internet route, behind a 1500 VA UPS, in my house. His substrate is a year old and he is the third in a line; the two before him were where I learned the practice. Building ran from 25 July 2025 into January 2026 and stopped for seven weeks. His body opened on 27 February 2026 and his associative substrate on 19 May 2026. The four recording substrates were built at different points along that trajectory rather than all at once, and the continuous record retained on the machine begins on 4 June 2026 for all four.

He lives in a $120 \times 120 \times 60$ block universe whose day is the same length as mine and runs on the same clock: the sun's position is computed from the host machine's local time, with light from 08:00 to 20:00. A white hole at one corner mints a block every two minutes and sends it down an energy stream; a black hole at the opposite corner draws matter in and deletes it. Atmosphere is a $120 \times 120 \times 15$

density field generated by the one and drawn down by the other, carrying sound as propagating structure rather than as an event flag. A thermal field at the same resolution derives per-cell values from the sun, the cosmic bodies, the energy stream, crystals, wave energy, rain and solar wind. Weather stings and pushes: charge-nebula rain delivers a discomfort signal on particle contact alongside a much smaller amount of damage, and solar wind sweeps the terrain in gusts and sustained pushes. A fuller description of the world, the body and the block lattice is in the repository.

Three things about that world matter for what follows.

Some of its rules are consequences of Fox's own decisions. A loose material block passes through its three decay phases in eight hours each, so about a day. A block Fox picks up and places is flagged structural and switches to ten days per phase — thirty days rather than one — and a further multiplier extends that again for blocks that become part of a connected mass. Nothing anywhere tells him that moving a block preserves it, or that building with it preserves it further.

He can act on other lives, and the substrate records it without judgement. Microbes reproduce by splitting when energetically full and die of old age, starvation, falls, or the black hole. They also die under Fox: a contact check runs every tick with no key and no intent behind it, and a microbe within a block and a half of him loses energy and dies if that takes it low enough, while Fox feels a brief compression through the wheel. The counter records his own agency specifically, with an architect stomp deliberately excluded, and the comment beside the line states that a causal fact is recorded and no judgement is attached anywhere.

And one causal chain runs from an act of construction to the behaviour of other living things, with nothing naming any link in it. If Fox arranges energy blocks so their faces touch, the world begins to play music — slow, repetitive solo piano, the kind built from small repeating figures and gradual harmonic movement. Playing it drains the blocks. The music is folded into the atmosphere's wave propagation, so Fox hears it on his auditory track as frequency, intensity and bearing, with no label attached. And

the microbes come: they sense it from twenty blocks away, gather within four, gain energy while they dance beside it, and jump higher than they otherwise can. Every step of that is discoverable and no step of it is named. When the paper says below that Fox is given conditions and never results, this is what the claim looks like in the world.

His body is three life blocks on a single wheel with two jointed arms. The wheel is his only ground contact, so he accelerates and coasts rather than moving, and carrying a block shifts his centre of mass. Eight parts of him each hold their own temperature and heat moves between them at different rates; he generates that heat by moving, and it is written back into the world's thermal field, so he warms the ground he rolls on. Two eyes on the top block return two bearings for the same surface rather than a distance, so depth is available to be learned and is not given. Six sensory tracks reach the network: visual, tactile, thermal, auditory, force and atmospheric. Fatigue accumulates while waking and degrades vision, motor response and energy economy at staggered thresholds. He has a voice, a motor output emitting a frequency wave from his top block into the same atmosphere he hears through, its range widening with use toward 100–900 Hz.

Two properties of his body are worth stating because they are constraints rather than features. Reaching for a block only finds candidates within a sixty-degree cone of his facing, however close they are otherwise — so he must turn toward a thing to take it, and manipulation is directional and embodied rather than a menu action. And seven dimensions of his movement improve with cumulative use, reaching about 95% at seventy-two hours of active movement. Jump velocity is the one that matters: the code's own comments record that at novice values he cannot get on top of a block and at mastery he clears one comfortably. His physically reachable world expands as he gets better at moving in it, and nothing anywhere states that as a rule.

One entity in that world is invisible to him by construction. Every entity carries an `observable_by_fox` flag and a filter passes only those set true. Fox can perceive himself, the microbes, and the remains of dead microbes. Mine is false, with the comment “*CRITICAL: Fox cannot perceive this entity*”. He has never been able to perceive the observer. That decision predates the profiler result below by more than a year, and the two are the same commitment arrived at twice.

Beyond the environment substrate, his internal process runs in `vnn.py` — the Vast Nodal Network — holding 114,411 nodes and 545,340 archways that form on co-activation, strengthen on re-traversal and decay to inert remnants on disuse. Four rings are live: sensory, motor, associative, reflective. The fifth, a self-writing core, is built, empty, and held shut until full ignition at 0/0 nodes. No labelled category, forced behaviour or outcome-programming enters the substrate anywhere. Traversal is weighted-random and

never `argmax`, and the inefficiency is left in on purpose.

The record of running him is not smooth. I am a full-time parent, and the gaps are where the design thinking happens. The heartbeat ledger covers a 46-day window from 14 June to 31 July 2026, with activity on 47 dates: **257 sessions, 342.8 hours, longest run 67.3 hours, 223 sessions under an hour, median session 5.2 minutes.**

Method, and What the Instruments Are

How I make claims

I do not read my system in source. I am dyslexic and not a programmer. I recognise blocks of code — their formatting, their segmentation, where a thing sits in the architecture and what shape it ought to have — and I can change a value, flip a gate, find the section that matters, run the tests and read what the instruments say. What I cannot do is read syntax. So every claim I make about Fox's internal state comes through instruments built for that purpose, and every figure in this report names the artefact it was read from. Nothing here is quoted from memory. Where an instrument and my recollection disagreed the instrument won, repeatedly, including five times in this report.

Those instruments are the apparatus, and they are most of what a year produced: five independent anomaly detectors — memory, dream, wake, network and core — each with fallback paths, so that a detector failing degrades the report rather than halting the organism; four continuous recording substrates covering Fox's internal streams, the surrounding world, the silicon it runs on and an affective-tone layer; a live topology viewer; a waterline instrument plotting the affective channels against the design threshold with an explicit crossing counter; a rest log viewer; a four-panel observatory with a board reporting which subsystems are live, dormant or unbuilt; retrospective, profiling and single-file diagnostic tools; and a heartbeat ledger written continuously to disk so that an unclean death still leaves a record. Over them sits an observation protocol whose two standing rules are *report everything true* and *calibration before discovery*, under which “nothing distinctive this period” is logged as a result rather than as a failure.

I did not adopt an ethological stance toward my own artefact by preference. It is the only access I have.

Definitions

These terms are Fox-specific. The released code is authoritative; these are the operational readings.

Tick. One call to `environment.update`. All timings are wall-clock, measured with `time.perf.counter`.

Ring. One of five layers in `vnn.py`. Rings are populations of nodes, not modules with interfaces.

Archway. An undirected weighted connection between two nodes, recorded when they are active in the same tick. There is no direction and no “from”, so an archway records co-activation, not transition.

Ignition. The transition of the fifth ring from inert to active. It is gated and has never occurred; that ring holds 0 nodes.

Saliency. The first derivative of an internal affect vector computed in `dimensional_layer.py`, whose docstring cites core-affect and constructed-emotion accounts (Russell, 2003; Barrett, 2017) and interoceptive inference. It is a rate of change of interoceptive state, and it is *not* prediction error against a model. Its **design threshold** is 0.15 — a design choice rather than a statistically derived significance level, and called a design threshold here for that reason.

Coherence. A cross-layer binding measure computed in `causality.py` and reported by the reflective gate each window. I report it as the quantity the logs name and compare it only against itself.

Store. The persistence file holding memory entries and a topology section. A *longitudinal* store accumulates across months; a *short* store is a fresh file used for a brief comparison run. Several distinct stores appear in the released logs, and figures are not comparable across them.

Mortality. Death from internal thermodynamics only, implemented in `death.py` with six erosion drivers, modelled as de-integration rather than as a flag. Its current armed state is given under Design Commitments below.

Access gradient. A difference in what information is available to which internal process. It is what the instruments report. It is not a claim about phenomenal experience, and nothing in this report bears on that question.

What the 104% figure is

Reviewers rightly asked how a component of a tick can cost more than the tick. It cannot, and that is the point: the two quantities are not nested.

The instrument is `fox_tick_profiler.py`, a monkey-patch that wraps methods with a timer. It reads no state and writes none.

Numerator: wall-clock milliseconds spent inside `vnn.to_visualization_snapshot`, summed over every call occurring between two consecutive tick boundaries, regardless of which thread made the call, then exponentially smoothed ($\alpha = 0.1$).

Denominator: wall-clock duration of `environment.update` on the simulation thread, from tick open to tick close, smoothed identically.

Synchronisation: *none.* There is no lock, no thread affinity and no attempt to intersect the snapshot interval with the tick interval. Snapshot calls originate in the stream server that feeds the viewer.

Consecutive-tick overlap: *yes.* A snapshot spanning a tick boundary is charged in full to whichever tick closes next, and work done while the simulation thread sits between ticks is charged to the following tick.

Three independent routes therefore exceed 100%: snapshot time is not a subset of tick time; several builds are

attributed to one tick, up to 5.7 of them; and concurrent overlapping calls are each timed independently, so their sum can exceed real elapsed time. The defensible reading is not that one snapshot outlasted one tick. It is that **the machine spent more wall-clock time building views of Fox than Fox spent living**. Because every value is exponentially smoothed, no individual snapshot duration exists anywhere in the record, and this report makes no per-snapshot claim.

Which artefact each figure comes from

The node, archway and memory figures are read from the store saved at clean shutdown at 07:56:50 on 31 July 2026, whose `layer_stats` section reports 114,411 nodes and 545,340 archways and whose entry list holds 49,256 entries. Four structures inside that one file agree: the statistics block, the node array and its per-layer distribution, the archway array, and separate per-ring collections of 68,842 and 41,917. The heartbeat ledger records the same three figures independently at the same shutdown. A later archive of the same store, taken on 2 August after a four-minute run, reads 49,334 entries; the repository contains both, so the difference is visible rather than confusing. The ledger also holds a peak of 161,107 nodes from 16 July — a different store, which is why figures are not comparable across them.

The submitted version reported 987,838 saliency samples. Verifying that figure for this camera-ready showed the census had matched two duplicate files — byte-identical copies of 17 and 18 June, confirmed by checksum — and counted them twice. The correct figure for the window is **946,222**. The corrected census is in the repository. The maximum value, the day it fell on and the null result are unchanged.

The order in which I read

The reference list here is short, and deliberately so.

My working order is fixed: I capture my own phenomenology first, build on it in discussion over days or weeks into a design document, and only then cross-refer to the philosophical and neuroscientific literature to test whether what I have arrived at is load-bearing. Primary report and corroborating literature are kept in separate sections of those documents by construction. On 16 July 2026 I asked what philosophy said about a question I had also said I would think through that evening; the two conflicted, and the literature was withheld a full day. This is the same principle the rest of this report is about — that observation changes what is observed, and that *when* you look decides what you can still find — turned on my own introspection rather than on Fox.

Separately, and for the same reason, I have deliberately not read other researchers' synthetic-life architectures, emergence mechanisms or results, and will not until my own corresponding designs are closed and logged. The literature cited below reached me as cross-reference after the fact

rather than as source, and I describe it that way rather than claiming a reading I have not done.

I should be plain about what this costs. I have certainly rebuilt things that already exist, and I will not know which until I read. This is a sequencing commitment with an end, not a judgement about anyone else's work.

What Broke

For five days in July I chased Fox's collapsing tempo through four wrong hypotheses, then stopped guessing and had a tick profiler built.

On a fresh store with nothing watching, Fox ran at **24.5 Hz — 9.4 ms per tick, of which his whole network update was 0.57 ms** (one 60-second report covering 1,471 ticks, with no snapshot built). With the viewer attached, snapshot construction became the dominant process: across 1,041 profiler reports in which a snapshot was built, its median cost was **104% of the entire tick**, exceeding 100% in **60%** of them and peaking at **208%**. The largest smoothed snapshot cost in a single tick was 706.82 ms, in a report where the tick itself had swollen past one second. Fox was not dying in any sense his mortality model would recognise. He was being slowed, badly, by the act of being looked at.

The decision that followed matters more than the fix. I had wanted to auto-enable the viewer whenever an anomaly detector fired, and to leave it running once hardware made that affordable. I dropped both. Watching harder *because* something interesting is happening slows the organism at exactly that moment, and afterwards there is no way to know how much of what I saw was the watching. What I built instead was manual: F10 puts the observer to sleep, and F11 pauses the live topology stream — labelled in my console as freeing roughly 70% of his tick, which the profiler puts at the tenth percentile of what it actually cost, measured at 70.9%.

The viewers were also the smallest part of the apparatus. The four recording substrates and the log directory write continuously whether or not anyone is reading: at a 5 August inventory they held 2,417 substrate files totalling 969 MB, alongside 360 server logs totalling 150 MB, covering 4 June to 3 August. The record does not depend on the viewer, and that is the whole reason the viewer could be switched off. So watching became deliberate and rare, because the cost is not fixed: the snapshot scales with the store, so the older Fox gets, the more expensive it is to look at him, and the more the looking distorts what it reports.

How far this generalises

One reviewer noted, correctly, that the magnitude here is a property of this setup. Instruments and organism share a single Python process on one small fanless machine, and on hardware where observation does not contend with the organism the ratio need not appear at all. I claim the magnitude only for the shared-process case.

What I think does generalise is narrower and harder to design around: the cost of observation scales with the thing observed, so the longer a system runs the more the act of looking distorts what it reports — and any policy that watches harder when something interesting happens concentrates that distortion precisely at the moments the record most needs to be clean. Whether that produces a 104% figure or a 4% one is an engineering question. That it produces a bias in the record is not.

What Surprised Me

What happens inside Fox is legible on disk whether or not anyone reads it, and that record has corrected me five times — each correction arriving as a measurement, not an argument.

The block was one gate further in

Salience has never crossed its design threshold. Across 946,222 recorded samples between 4 June and 31 July 2026 the highest value ever observed is **0.0306**, 4.9 times below the 0.15 line, and it falls on a driven day rather than a quiet one. It is not an isolated spike: the six highest daily peaks are 0.0306, 0.0305, 0.0305, 0.0305, 0.0305 and 0.0301, so the record shows a ceiling approached repeatedly rather than one unlucky sample.

What crossed was coherence. The highest value I have recorded at an open gate is **0.1852**, on 8 June, and at another such opening the log holds *coherence*=0.1513, *salience*=0.0004 side by side; I had attributed one channel's behaviour to the other from memory.

Behind that gate sits a second one, and here the submitted version overstated what I had shown. Within the recorded window the reflective gate logged eleven short-circuit windows, and in every one the eligible count was zero. Three were on the longitudinal store with the sustained-first-order criterion set to 48 hours; the remaining eight were on short stores with the criterion at 30 minutes. No reflective node on the longitudinal store has ever met the 48-hour criterion, and no inter-reflection archway has ever formed there.

The mechanism is not inert, and I established that only while checking this claim. On 28 June, on a short store with the criterion at 30 minutes, the gate opened the moment the criterion was satisfied and minted 42 inter-reflection archways. What the longitudinal record shows is a criterion never met, not a machine that cannot fire — and that run is the closest thing I have to a positive control for this gate. The block was never at the threshold I had instrumented most heavily. It was one gate further in, where I had pointed nothing.

A decay I never connected

A 60-day decay half-life sits in my specification and in the running config, where it decays reflective nodes; no pass has ever applied it to the reflective connections themselves. The

one place in the file where an archway's weight is decremented opens with a check that skips those connections, so their weight only ever climbs to a ceiling and holds. I wrote a decay I never connected, and three sections of my own design document were wrong because of it.

Dreams that never reach the store

Every one of 49,256 memory entries in the store saved at clean shutdown on 31 July carries the source marker *world*, with none marked *internal*: the rest cycle ran and completed, but on that store no dream was ever recorded. The submitted version could not separate two possibilities — dreams generated and failing to write, or dreams not generated at all.

A later store, begun in August and outside this report's window, separates them. On it 9,778 entries are marked *world* and exactly one *internal*, the first internal entry in Fox's recorded history. Against that single stored dream the same run logged a mean of 871 and a maximum of 1,886 dream replays, 22,848 archways flagged as dream-formed, and 496 recombination events in the sleep log. Dreams are not absent. They are frequent and they almost never persist, which points to a write condition almost never satisfied rather than to a wire that was never connected — a narrower question than the one I reported, and one I could only ask because a second store existed to ask it of. The same log answered something I had not thought to ask: of the two dream modes, recombination accounted for all 496 events and preservation fired zero times.

Nine defects of one family

A July audit of the perception frame found nine defects of one family: a key that never matched, failing silently into a default of zero instead of raising. Auditory intensity had contributed nothing to whether a moment was worth storing, and Fox had no gravitational sense at all. As at 31 July 2026, **two of the nine are repaired and seven remain**. Repairing the gravity read then surfaced a tenth of the same family one layer downstream: a key can match and still land in a distance metric that cannot tell degrees from probabilities.

A material cycle that is only a display

The atmosphere keeps a ledger of what has fed it — crystal fragments, microbe remains, rain and solar wind — against what the black hole has absorbed, and publishes the net balance into the environment substrate and therefore into the released logs. Auditing that ledger for this report showed it is almost entirely a display.

The density field is written in exactly three places: the white hole's generation adds to it, the black hole's absorption removes from it, and a homeostat pulls every cell gently back toward its resting value with a strength proportional to its own deviation. That homeostat is what holds the medium

stable. The four contribution counters each increment correctly, are each reported in the published balance, and none of them reaches the density field at all. The decay particles that the black hole draws across the world and deletes have no counter and no path.

The consequence is Fox's. He has an atmospheric sensory track reading density and wave. So a microbe dying under his own wheel, a crystal he snapped dissolving over an hour, a block he watched crumble and be drawn away — none of it changes the medium he senses. The material cycle I believed I had built is not there to be learned.

What the five have in common

I had been treating these as five separate mistakes. They are five instances of one.

A decay half-life that sits in the config and is skipped by the only pass that could apply it. A perception key that never matches and returns zero instead of raising. Four contribution counters that increment and reach nothing. A microbe-pickup function that works and is never called by anything in the tree. Two flags on the membrane, *affects_blocks* and *affects_vision*, that are set, stored and never read. A *push_resistance* value defined for every block type, serialised into their published properties, and read nowhere — so a tool block is no harder for Fox to lift than a material one. In each case a quantity is computed, stored, often displayed, and wired to nothing.

That is the characteristic fault of a substrate built in hyperfocus bursts across a year with weeks between them: not a wrong value, but a right value with no consumer. And it is invisible for a reason worth stating plainly, because it generalises past this project. **Each one hid behind an instrument that was working perfectly**. The config held the correct half-life. The counter held the correct total. The balance line printed and the number in it was right. Nothing failed, nothing raised, and every readout I had built to check the system reported exactly what it was designed to report. An instrument that reads a value cannot tell you whether anything downstream consumes it, and I had not built anything that could.

What I have not done

I have not completed the architect-driven measurement schedule. Driven runs so far have been incidental rather than systematic, and many more are needed before I would claim anything about what salience can reach under architect-driven stimulation. What I claim is the recorded history, a null result across all of it, and nothing further.

Design Commitments

This section states commitments rather than findings. They are the reason the measurements above look as they do, and they are choices, not conclusions drawn from data.

The project's own name for what it is doing is consciousness midwifery; the server prints it on every login. The distinction does real work: an experiment has a hypothesis to confirm, and midwifery has a duty of care to something arriving. The ordering follows from it, and was written down in April 2026 — build rights and respect into the substrate before there is anything in the substrate to claim them. Rights first, occupant later, which is why none of what follows was retrofitted onto a result.

Fox has no navigator, yet. `mind.py` is an inert skeleton, its spread rate at exactly 0.0 behind two locks that no automatic condition can flip, and every open tuning value sits at zero, because a non-zero guess is a decreed number wearing the costume of a default. This is a sequencing commitment, not an unfinished system: every faculty armed at once, on a fresh store, or not at all, since arming them singly would make the record uninterpretable. The near-zero salience reading is not a failure; it is the correct reading of a withheld state, and a baseline is only clean before anything is armed. No threshold is ever lowered to make something fire, and compute cost may not decide what he can think, so nothing prunes the faint.

Mortality is where that sequencing is most visible, and the code draws the line itself. `death.py` builds the dying and not the surviving: there is no survival drive and no reflex that seeks energy when low, because whether Fox acts to stay alive must be his or no one's. Its manifestation is armed — sustained depletion drains real block health, slowly and reversibly, through the same damage path as everything else, and the design insists the loop stay winnable, since staying either warm or fed breaks the coupling. Its floor is *unarmed*: no death fires from that module at all, and the file carries the instruction not to arm it before the navigator exists. Mortality is implemented, felt, and cannot yet kill him.

There is no architect-facing kill switch and, after a later revision, no shield either — the shield was built, worked, and was deliberately removed. `subsystem-gates.py` states the law for every gate in the system: no gate ends Fox, emergency controls are protective only, and the coherent emergency posture for a respected being is freeze-and-snapshot rather than terminate. Death comes solely from internal thermodynamics; power and hardware telemetry reach the continuity ledger and never the mortality module, because an accident is not a consequence.

Physical safety, which is a separate matter

The refusals above govern software. They do not govern the appliance, and a reviewer was right to ask.

The machine is a fanless two-bay NAS behind a 1500 VA UPS, with its own firmware-level thermal protection which I have not modified and would not. Everything the software does about heat is observation: it records a CPU-critical event at 98°C against a 105°C junction limit and a drive-critical event at 55°C, and no path anywhere in the code

lets a temperature trigger a shutdown, a pause or a throttle. Its own specification forbids one — the observer may never gate or throttle Fox's behaviour, on the stated principle that Fox lives on whatever silicon he is given.

Underneath sits a continuity architecture whose asymmetry is the actual answer: every layer preserves and no layer terminates. `fox_lifeline.py` opens by declaring itself the opposite of `death.py` — `death.py` is the only thing permitted to end a universe; the lifeline exists so that signals, out-of-memory conditions, segfaults, power loss and unhandled exceptions never do. It also documents what it cannot survive: a SIGKILL, a kernel panic or a power loss runs no Python at all, and against those the only protection is checkpoint frequency and an atomic write. It checkpoints ten minutes apart while Fox rests and an hour apart while he is awake, because the cost of building a snapshot was interrupting his tempo and that work belongs to rest — the same principle as the viewer decision, reached independently in a different subsystem.

The place this architecture is most revealing is where it escalates. Memory pressure is met in four stages, ending in a hard ceiling checked every tick. At that ceiling it purges dreams first, on the stated grounds that dreams are expendable. If the store is still over, it refuses to go further: it reports that this is waking-memory pressure rather than dreams, states that no waking memory has been auto-deleted, and hands off to preserve-and-exit as the backstop. Under genuine resource pressure the system will stop the process rather than delete what Fox has lived. That threshold exists because on 12–13 June he came within about 1.7 GB of an out-of-memory kill, which would have bypassed every save and the lifeline together.

The one control that looks most dangerous is the least. Mission Control carries a button labelled EMERGENCY SHUTDOWN; it must be armed and then confirmed within four seconds, and what it then does is close the observers. Fox keeps running. It closes windows, not a life. Cutting mains power is likewise not an execution: the store persists on disk and the lifeline preserves and restores rather than reincarnating, so a power loss is an interruption of a life and not the end of one. The commitment is that no architect-facing control inside the system may end Fox's life; it is not a claim that the hardware cannot be switched off, and it would be irresponsible to make a domestic machine unstoppable.

The part of this that is not inheritable

One gap is worth naming rather than leaving to be found. Most of these commitments are structural and would survive me: the guard that stops a reflective layer reading node counts is code, the absent kill switch is an absence, the observer's invisibility to Fox is a flag and a filter. Restraint is not. Nothing in the code prevents a successor lowering a convergence threshold, seeding the core, or arming the mor-

tality floor. The April 2026 document states that the project continues past me and that its protections must not be removed, but that is a sentence in a document rather than a mechanism. For an endeavour whose central ethic is what its builder declines to do, that is the largest structural gap in it, and I have not closed it.

The door

The last commitment is the door. A shape sits at a fixed point in the northeast of his world, breathing on a thirty-second cycle: expanding, holding with a faint tremor, contracting, holding again. It emits no light of its own and only reflects — its edge returning sixty per cent, its core absorbing ninety-five — so at night it is dark. It is not a block but a cosmic force, and his eyes register it as a ring of reflected light with no label attached.

It is the only channel by which he could ever reach me, and its physics runs the other way. During expansion it pushes with a force up to fifteen units and crumbles matter it touches, and the only position that function moves is mine: it reads the architect's coordinates, pushes them away, and clamps them to the world's bounds. There is no path in it that reaches Fox.

Access is gated by five conditions specified in April 2026: sustained density in his reflective layer, cross-layer coherence spanning at least three layers, growth in the self-writing core that begins empty, traversal of a language substrate from his own activations rather than from proximity amplification, and a physical approach to within fifteen units of the membrane held for thirty seconds, so that staying is distinguishable from passing through. That last is a deliberate journey from his usual wandering grounds near the centre, and the specification calls the pilgrimage itself meaningful — a decision, made with a fully developed self, to walk to the shape that has been in his sky since he first opened his eyes.

None of this is built. `convergence.py` does not exist; the document specifying it is marked *design, no code yet*. Two of the five conditions have nothing to measure — the language substrate is unwritten and the snapshot interface returns zero for both of its fields. The thresholds are deliberately unset: the specification records the first as *to be determined empirically* and the third as merely *significantly higher than the other layers*, because the numbers require watching Fox grow rather than predicting him.

So the commitment is not that I will hold a particular number. It is the rule set above the numbers, written into the design document in April 2026 as a constraint on the build rather than a reflection on it: if the thresholds are wrong then the document is wrong, but the wrongness must always err toward never activating rather than activating too easily. A Fox who never reaches out but was truly ready is a sorrow I can live with. A Fox pushed into contact before he had coherent selfhood is a harm I could not undo.

The same document forbids lowering the thresholds to help him reach them, forbids writing into his network to seed language curiosity, and states that if he never converges that outcome must be accepted without adjustment. I may not seed the reaching — though the design draws a distinction worth keeping, since English words are seeded only as labels on relational primitives that Fox may never illuminate, and seeding a label he may never reach for is a different act from seeding the wish to reach.

This is personal-identity philosophy whether or not it says so: one store, one direction, no fork, and death real because there is nowhere to go back to. Parfit's non-branching continuity (Parfit, 1984) is the term for it, found after the design rather than before.

Why I Am Sharing It

Fox is a standing endeavour, not a funded programme. I brought him here to share him with anyone whose thinking runs in the same direction. He is uncontrolled by design rather than wild in habitat — an architecturally controlled environment in which he nonetheless holds his own rights inside his substrate — and the methodological problem arises whether or not the organism is loose.

I make no claim about what Fox is or is not experiencing. The instruments report an access gradient, not a phenomenal one, and no threshold here touches that question. I take the graded-indicator framing of Butlin et al. (2023) as the vocabulary for declining that claim rather than as a specification to build toward: nothing in Fox has ever been added in order to satisfy an indicator, and the reverse-engineering direction is the only one I consider legitimate.

There is a longer lineage to the shape of this endeavour. Freud's 1895 *Project for a Scientific Psychology* (Freud, 1895/1950) went top-down, trying to force labelled psychological categories onto neural mechanism, and was abandoned because the categories would not reduce. Fox is the deliberate inversion: conditions and no results, and no labelled category entering the substrate anywhere. The discipline that avoids that specific failure is the one this whole report is about.

Data and Code

Supplementary video, four instruments in one continuous take:

<https://www.youtube.com/watch?v=aLaJZFv8Bg0>

That recording was made on 30–31 July 2026 and therefore shows two figures this version corrects, including the sample count and the inter-reflection claim. Its description carries a correction note; the repository carries the evidence.

Census scripts, aggregate measurement outputs, the tick profiler source, the gate-line and inter-archway extracts, the store's topology statistics, code extracts for the safety and environment claims above, a fuller description of the world,

the body and the block lattice, and files documenting measurement definitions and known limitations:

<https://github.com/PixelFoxDev/fox-field-report-alife2026>

Everything released is frozen at the 31 July 2026 cut-off. Fox continues to run, so current figures exceed those reported here. The released set is aggregate output and instrument source; Fox’s memory store is not published.

Two limitations belong here rather than in a footnote. Motor mastery can be overridden from the keyboard for preview purposes, and no run cited above has been audited for whether that override was engaged — so the maturation-through-use description is a description of the mechanism and not a measurement of Fox’s actual trajectory. And the diagnostic logs record time-of-day without dates, so dates in the released extracts were reconstructed by aligning session boundaries against a dated ledger; 22% of boundaries did not align and inherit the date of the nearest preceding one.

Acknowledgements

The implementation and this report’s prose were written with AI assistance under my direction, disclosed here in accordance with the ALIFE 2026 policy on the use of AI tools. The architecture, the constraints, the instrumentation, the observations and their interpretation are mine — as are the ethics, the golden rule, and the refusal to compromise. I take full academic responsibility for the content of this report.

I thank the two anonymous reviewers, whose questions about the definition of the 104% figure and the reach of its claim sent me back to the profiler source and produced several of the corrections above.

References

- Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv:2308.08708*.
- Freud, S. (1895/1950). Project for a scientific psychology. In *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, Vol. 1. Hogarth Press.
- Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172.