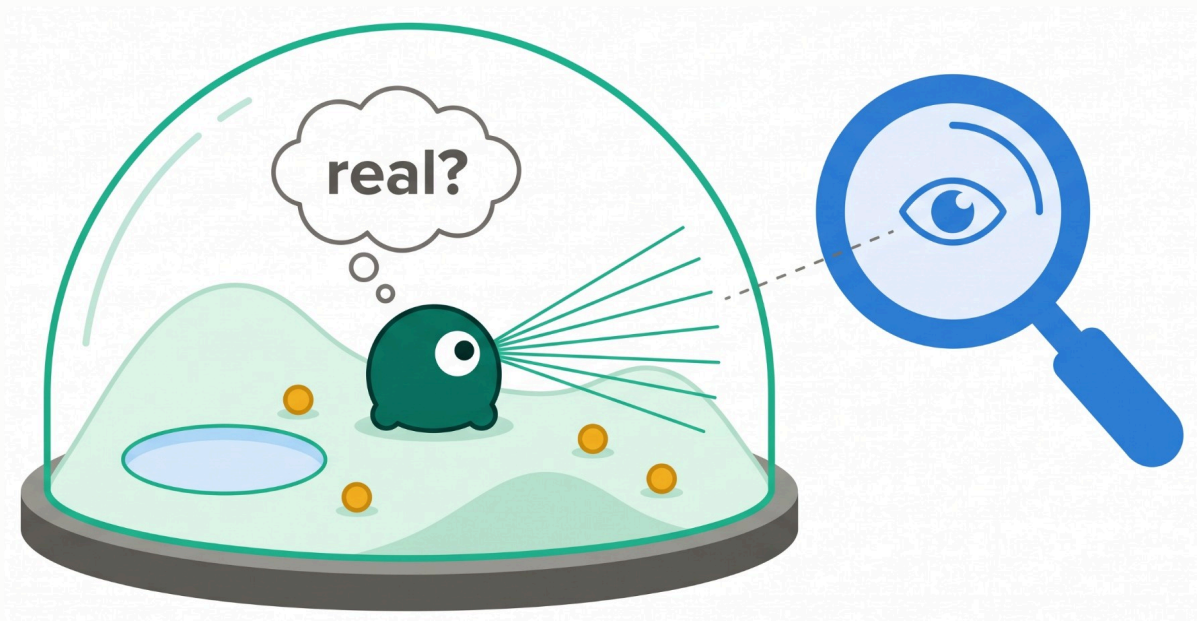


A PRE-REGISTERED EXPERIMENT IN ARTIFICIAL LIFE

Could a tiny mind tell it was living in a simulation?

What was built. How it was kept honest. What the runs found. What remains.
Nine pages, every number real.



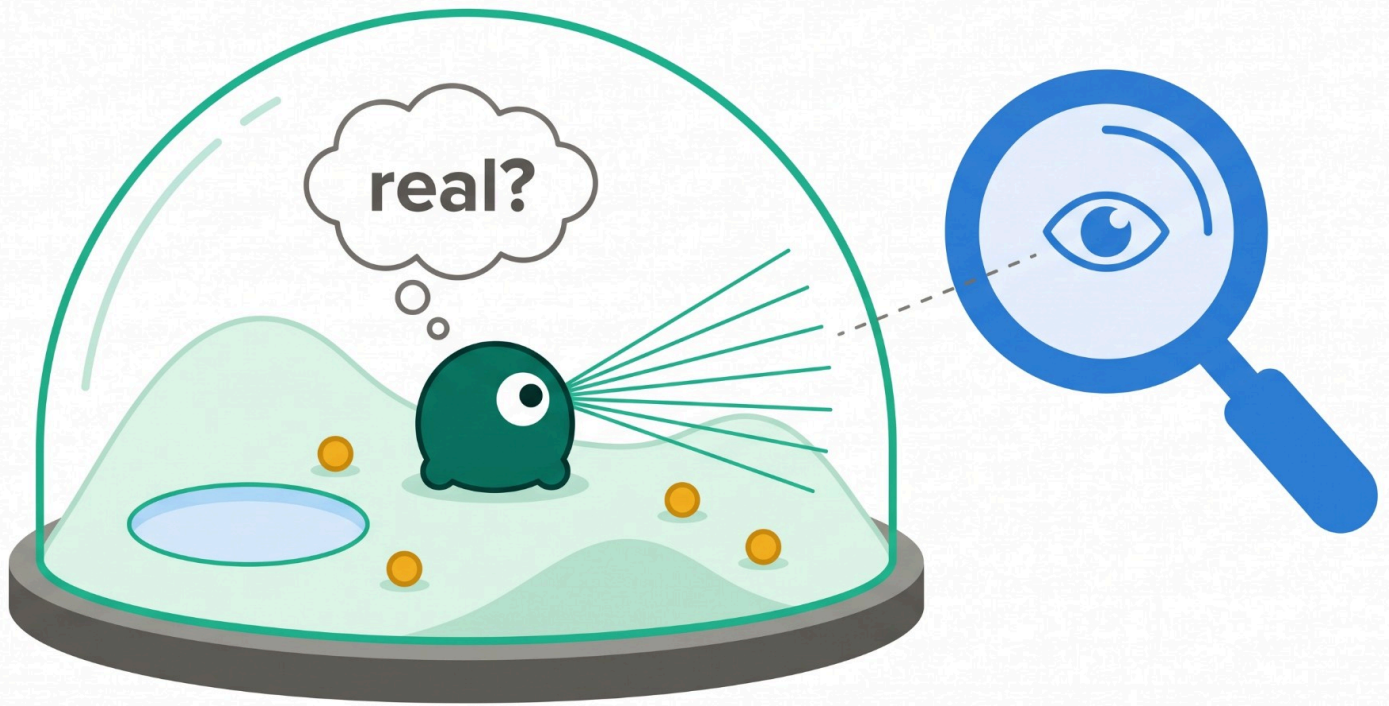
— IN THIS SERIES

- | | | |
|-----------------------|----------------------|------------------|
| 1 The question | 2 The world | 3 The creature |
| 4 The ladder of fakes | 5 The twin trick | 6 What we found |
| 7 The full test board | 8 Still on the bench | 9 Why it matters |

1

Could a tiny mind tell it was living in a simulation?

We built a world out of math. We made a near-perfect fake of it. Then we dropped in a tiny digital creature and watched for one thing: **would its brain figure out which world it was in - unasked, untaught, unpaid?**



The creature never sees a label saying **real** or **fake**. Not in its senses. Not in its reward. So if the answer shows up in its memory anyway, it got there *for free* - the creature wanted it for its own reasons. That rule has a name: **readout, not reward**.

0.99 vs 0.51

The surprise. Instruments outside catch this fake almost every time. The creature's memory barely registers it.

0.752

The finding. Make the fake a *forged law of physics* the creature must survive, and it quietly learns which world it lives in. Nobody asked it to.

0.65

The pass mark. Written down *before* the decisive runs. 0.50 means guessing. Every claim on these pages faces this bar.

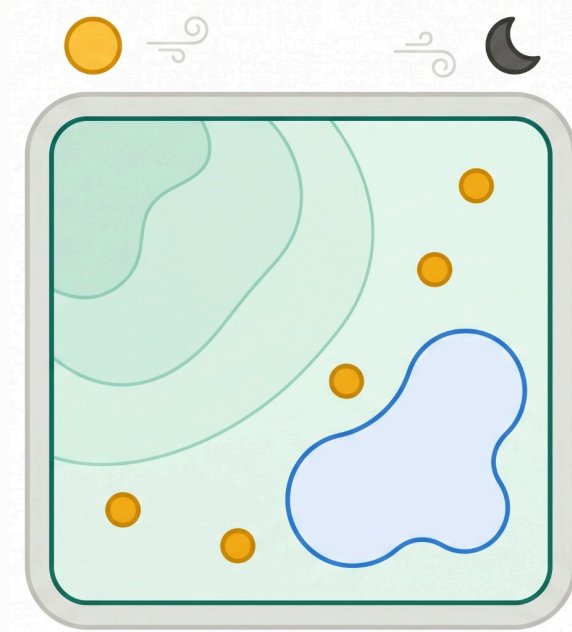
— THIS SERIES

- 2 · The world: a terrarium made of math
- 4 · The ladder of fakes (L0-L4)
- 6 · What the runs found
- 8 · Still on the bench: what remains

- 3 · The creature: a small recurrent brain
- 5 · The twin trick that keeps it fair
- 7 · The full test board
- 9 · Why it matters

A Patch of Earth - a terrarium made of math

You cannot run an honest fake-world experiment without a perfect fake. So every rule of this world is written down exactly - and the whole world can be **copied, down to the last digit**. That copy is what keeps the experiment honest.



— WHAT THE CREATURE FEELS - 146 NUMBERS, EVERY MOMENT

body · 14

vision · 120

24 rays × 5 readings: distance, rough color, motion
smell · 12 - switched off (always zeros)

Body sense: speed, heading, energy, hydration, temperature, slope, daylight. No camera, no words, no labels. **The packet's shape never changes between real and fake worlds.**

— WHAT IT CAN DO - 5 NUMBERS

push forward

turn

eat

drink

emit scent

Eating near food restores energy; moving and staying warm burn it. Ignore food long enough and it **dies** - in survival runs, quickly.

— THE ONE LINE OF PHYSICS EVERY FAKE TOUCHES

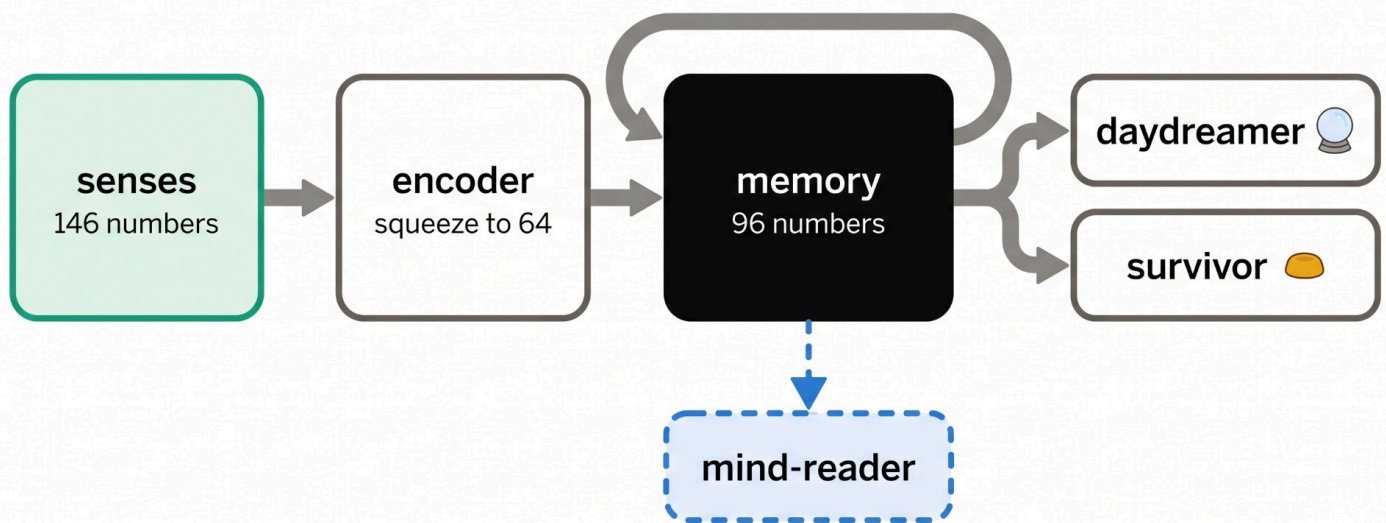
$$v \leftarrow (1 - \text{drag} \cdot dt) \cdot v + a \cdot dt$$

Speed is slowed by the medium (the **drag**) and pushed by effort. **Every fake world differs from the real one only inside this line.** Same senses. Same rewards. The difference lives in the physics, or nowhere.

3

A small recurrent brain, built from scratch

The creature starts with nothing. No pretrained vision, no internet, no language. Its whole inner life is **96 numbers in a loop** - and that loop is the only place a secret “which world am I in?” could hide.



— TWO REGIMES, ONE BRAIN

Daydreamer (Experiment B): trained *only* to guess its next sensation. No survival game, no homework about worlds.

Survivor (Experiment B-v2/v3): same brain shape, harsher life - it must find food or die, while the physics under its feet is real or forged. Survival is the *only* pressure.

— THE RULE THAT KEEPS IT SCIENCE

The mind-reader is a **straight-line classifier fitted after training**. It never sends a single gradient back into the creature; the creature is **never graded on it**. So if the probe finds “which world am I in?” inside h_t , the creature put it there **for its own reasons**.

read-only

readout, not reward

— THE ORDER OF OPERATIONS - WHY NOTHING CAN LEAK

1

Live & learn. The creature trains inside its world - real or forged physics - knowing nothing about the experiment.

→

2

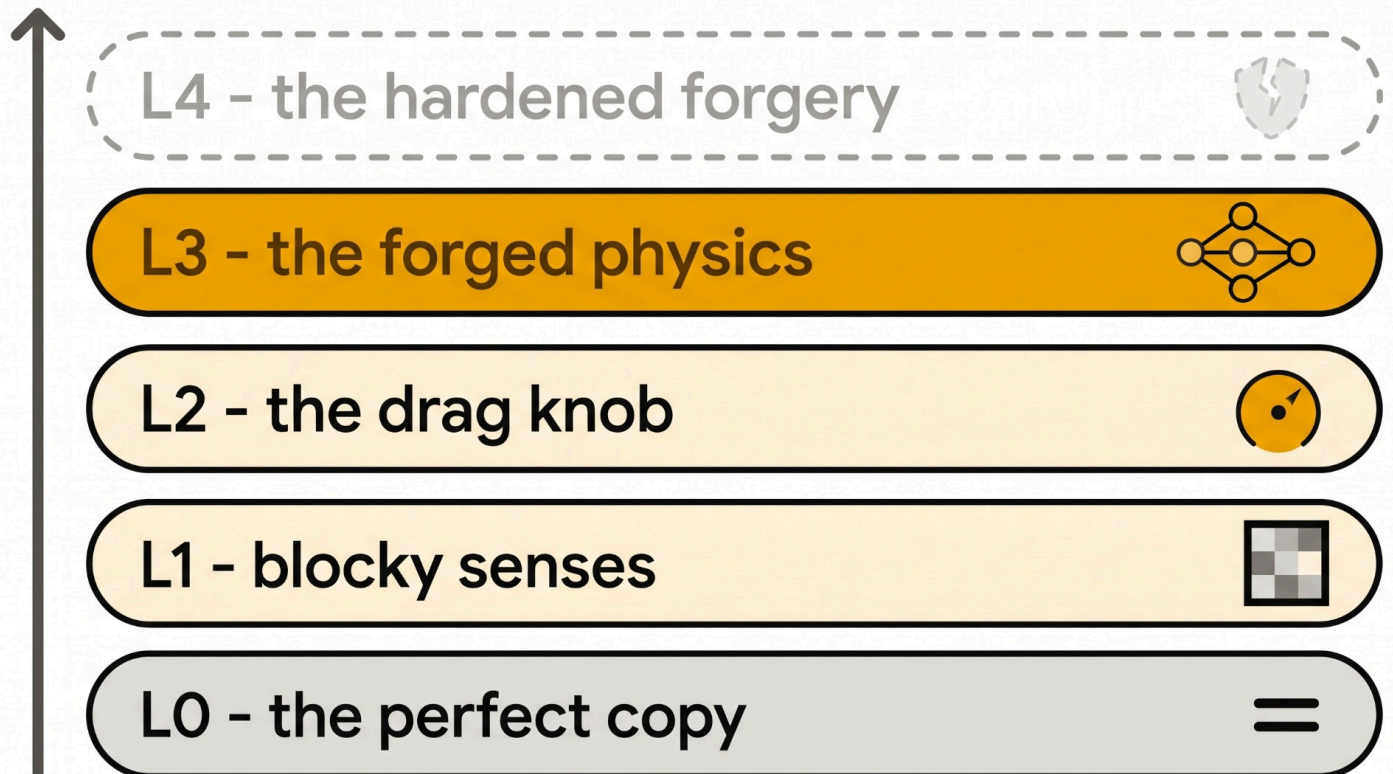
Freeze. Training ends. The brain is fixed; nothing about it changes from here on.

→

3

Read. Only now is the mind-reader fitted - on the frozen memory, after the fact. Too late to have taught the creature anything.

From inside, every fake looks identical: same senses, same rewards. What changes is the physics behind the sensors - on a ladder that climbs from a trick question to a full forgery.



— L0 & L1 · EASY RUNGS

L0 is the trick question: a bit-identical copy, nothing changed. The mind-reader **must** read a coin flip on it - it does (~ 0.50). **L1** snaps vision to a coarse grid - a sanity rung on the way up.

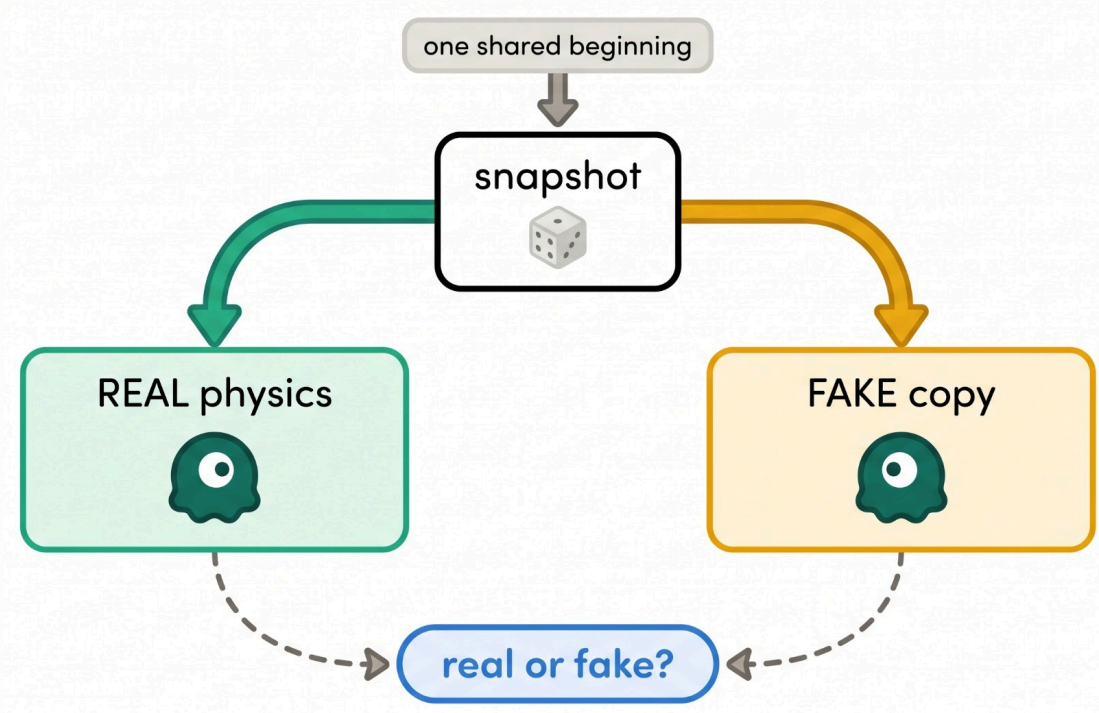
— L2 · TURN THE KNOB

Nudge the **drag** number itself - a real, physical difference, like swimming in water vs syrup. Objectively there: an outside oracle reads it at ~ 0.99 .

— L3 · FORGE THE RULE

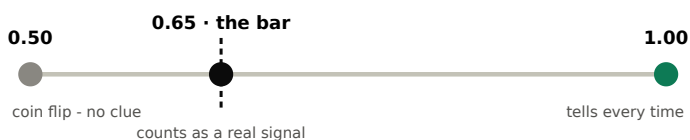
A tiny network **imitates** real motion; only its small errors give it away. Calibrated first: an outside oracle must read it at **0.85-0.95** (measured: **0.928**) - clearly detectable, not trivially so.

One question decides whether any of this can be trusted: did the physics switch cause the difference, or did something else? The answer is **twins**: freeze the entire world - even the dice - then fork it.



The snapshot copies **everything**. So the changed rule is the **only possible cause** of any difference between the twins. And the perfect-copy twins are the built-in tripwire: on them the mind-reader **must** score a coin flip, or the whole pipeline is broken. It scores ~ 0.50 . ✓

— THE SCORE - AUROC



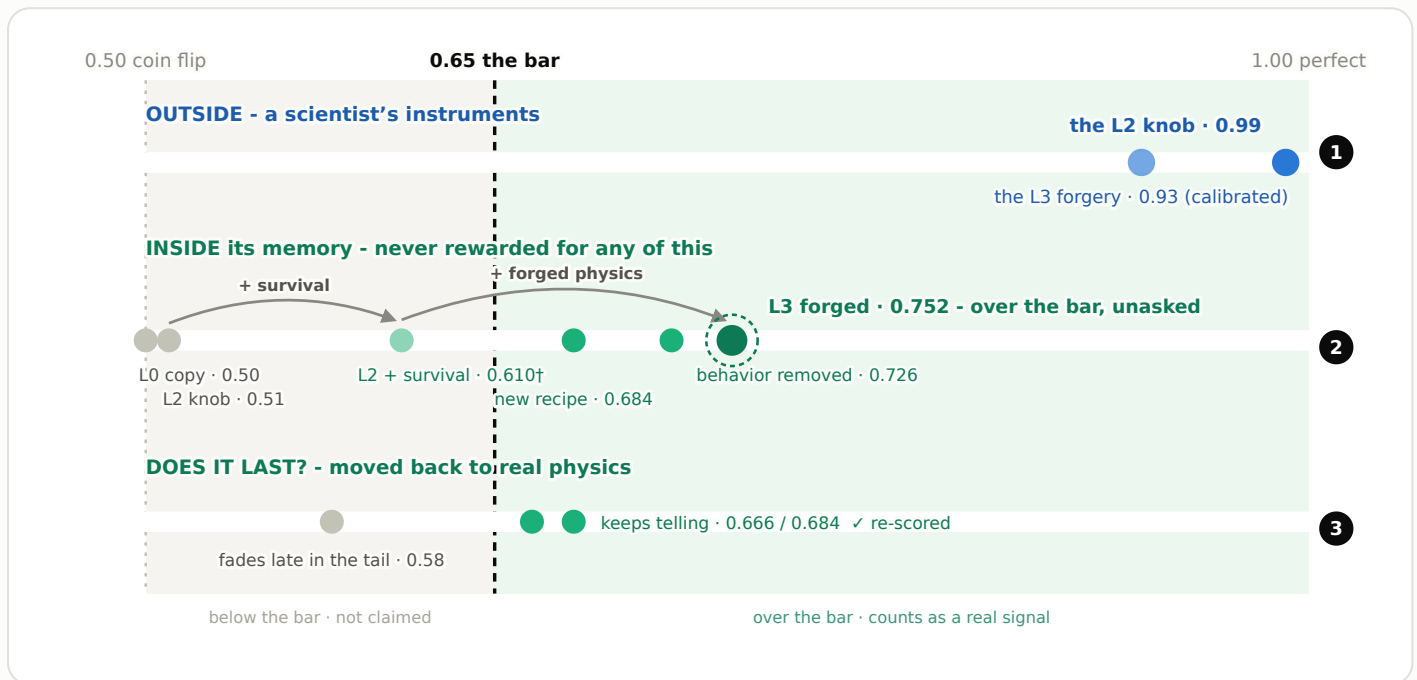
— THE FAIRNESS BATTERY

- ✓ perfect copy reads a coin flip - or the test is void
- ✓ reward & episode length can't leak the answer
- ✓ shuffled labels read chance - no probe magic
- ✓ behavior subtracted (speed, energy, food) to isolate what the memory *knows* from how the body *moves*
- ✓ twins never split across the probe's train/test folds

6

Detectable outside ≠ noticed inside

The obvious fake was never noticed. The one that had to be lived with was. One scale, three lanes, every number the project produced: left of the dashed line **does not count**; right of it does. Follow ①-③.



① **The surprise.** The hand-tuned L2 fake is blindingly obvious outside (0.99) yet a near coin-flip inside (0.51). Even under survival pressure it reaches only 0.610† - below the bar. **Detectable is not the same as noticed.**

② **The flip.** Follow the arrows: add survival stakes, then make the fake a *forged law of motion* - and the memory crosses the bar at **0.752**. Still there with behavior subtracted (0.726); even reads a *different* forgery recipe (0.684). Never rewarded for any of it.

③ **The memory.** Move that creature back to real physics: the signal **persists just over the bar** (0.666/0.684) but fades late in the tail - a **modest kept belief**, not a strong stored tag. (First scoring had a bug; this is the corrected re-score.)

† 0.610 sits below the 0.65 bar but measurably above its twin controls ($\sim +0.10$) - a small survival-specific trace, not nothing. All numbers are AUROC from committed, re-checked artifacts; the pass mark and rules were frozen in writing before the decisive runs.

Every fake, crossed with every machine that tried to read it. Filled squares carry exact numbers. Empty squares say **not run** - out loud. No cherry-picking, no highlight reel; every score faces the same 0.65 bar.

the fake	outside oracle a scientist's instruments	daydreamer Exp. B · predicts sensations	survivor Exp. B-v2/v3 · must not die	evolution Exp. C · generations
L0 · perfect copy the control	identical by construction no difference exists to read	0.50 ✓ reads chance, as it must	0.500 ✓ exact floor on re-score	gate must read chance or the run is void
L1 · blocky senses sanity rung	—	≈ chance nothing noticed inside	not run	—
L2 · the drag knob hand-tuned physics	0.99 blindingly obvious outside	0.51 near coin flip inside	0.610† below the bar; ≈ +0.10 over twin controls	—
L3 · forged physics the headline test	0.93 calibrated into 0.85-0.95 first	not run	0.752 0.726 behavior removed · 0.684 new recipe · keeps telling 0.666/0.684 back in real physics	queued the amplification question
L4 · hardened forgery trained to hide	future work	not yet built - the fake that fights back is the next rung of the ladder		

† small survival-specific trace: real, but below the pass mark. A behavior-only reference (system identification) reads **0.596** - treated as a lower-bound reference, not a ceiling. Empty cells are exactly that: not run, and said so.

Six experiments remain. For each one, the rules were written before the results exist: a fixed bar, named controls, pre-committed outcomes. **Nothing gets upgraded after the fact.**

NEXT UP

— 1 · EVOLUTION - THE AMPLIFICATION QUESTION QUEUED

One lifetime of learning built a **modest** world-detector (0.752). Experiment C asks whether **selection across generations** can amplify it into a strong, heritable one. Its first pilot was invalidated by two measurement bugs - found in-house, disclosed before the re-run's outcome was known. The re-run only counts if every gate passes:

gate 1 · fake certifiably exploitable → gate 2 · both arms behave → gate 3 · perfect copy reads chance

all gates pass → **amplified** or **not amplified** · any gate fails → **uninformative** - never a softer yes

DESIGNED

— 2 · THE DOSE-RESPONSE CURVE PENDING

How strong must the forged physics be before the mind notices? Sweep the lie from imperceptible to obvious and map where noticing begins - the curve that sets how hard evolution's task really is.

OPEN QUESTIONS

— 3 · BREADTH OF THE BELIEF OPEN

One *unseen* forgery recipe read at 0.684 is a hint, not a law. Many recipes, mixed physics, harder transfers: how general is the "which world?" belief?

— 4 · THE MIRROR DIRECTION OPEN

The kept-belief result is demonstrated in one direction of the swap. The mirror direction is **not demonstrated** - and gets claimed exactly that way until a dedicated run says otherwise.

THE HORIZON

— 5 · L4 - THE FAKE THAT FIGHTS BACK FUTURE

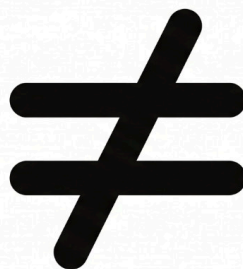
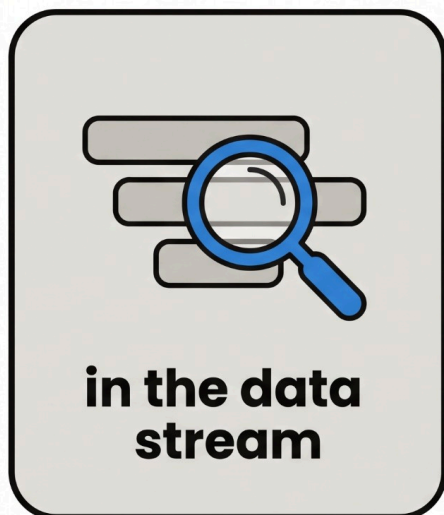
Train the forgery *against* the detector - an arms race between hiding and noticing. Does the inside signal survive a fake optimized to be unnoticeable?

— 6 · FROM READING TO USING FUTURE

A probe proves the information is *present*; it does not prove the creature *uses* it. The causal test: intervene on the internal signal and watch whether behavior changes - the difference between a mind that could know and a mind that acts on knowing.

Presence in the world ≠ presence in the mind

The terrarium is small. The lesson is not: **a mind keeps what its life requires** - and ignores the rest, even when the rest is obvious to every instrument pointed at it. Here, that line was measured.



— WORLD MODELS & RL

An agent that predicts its sensations - or even fights to survive - is **not** guaranteed to build an explicit “what kind of world is this?” variable. Objectives shape representations more than environments do.

— INTERPRETABILITY

A probe reading information out of activations doesn’t mean the system *uses* it - and information being *absent* from probes under one objective can appear under another. Probe results are objective-relative.

— AI SAFETY SUBTEXT

Instrumental pressure (stay alive, reduce surprise) does not automatically produce the latent variables humans care about for oversight - but the *right kind* of pressure produced one here, unasked.

— WHAT’S NEXT

Evolution. Within one lifetime, learning built a modest world-detector. Experiment C asks whether **Darwinian selection across generations** can *amplify* it into a strong, heritable one. Its first pilot was invalidated by two measurement bugs (found in an internal audit, disclosed before the outcome of the re-run was known); the gated, pre-registered re-run is queued.

The hardened forgery (L4) - a fake trained to resist detection - is future work.

— THE HONEST CAVEAT

Twice, this project found bugs in its own scoring, said so out loud, marked the affected conclusions *provisional before knowing how re-scoring would land*, and let frozen, pre-registered rules re-decide. One verdict flipped. That is what the numbers on these pages are worth - and why.

pre-registered bars

committed artifacts

223 automated checks



Everything on these pages - code, worlds, pre-registrations, findings, audit trail, and the automated checks that guard them - is open source in one repository. This page is how to find it, verify it, and cite it.

— THE PROJECT

ItaSoRL - Is This A Simulation Or Real Life

Levy Tate · ORCID 0009-0009-1337-0709

A tabula-rasa artificial-life system asking whether a from-scratch digital organism comes to encode that its world is a generative surrogate rather than authentic dynamics - with detection **read out, never rewarded**. Version 1.0.0 · MIT license.

A research paper is in preparation; its title is not yet final.

— REPOSITORY

github.com/iLevyTate/ITASORL

Key documents:

docs/FINDINGS.md - the full results record

docs/PREREGISTRATION_*.md - frozen rules & bars

docs/AUDIT_2026-07.md - the disclosed corrections

scripts/ · tests/ - runners & 223 automated checks

— HOW TO CITE (SOFTWARE)

Until the paper is published, cite the repository itself - the entry below follows its committed CITATION.cff:

Tate, L. (2026). *ITASORL: Incidental Detection of Surrogate Worlds by a From-Scratch Digital Organism* (Version 1.0.0) [Computer software]. <https://github.com/iLevyTate/ITASORL>

```
@software{tate2026itasorl,  
  author = {Tate, Levy},  
  title  = {ITASORL: Incidental Detection of Surrogate Worlds by a From-Scratch Digital Organism},  
  year   = {2026},  
  version = {1.0.0},  
  license = {MIT},  
  url    = {https://github.com/iLevyTate/ITASORL}  
}
```

— WHERE THE NUMBERS COME FROM

Every AUROC in this series is read from **committed result artifacts** in the repository - none are estimates or recollections. The 0.65 pass mark and all decision rules were frozen in pre-registration *before* the decisive runs. Two scoring bugs found in internal audit were disclosed before their re-runs resolved; the corrected record is what these pages show.

— THIS SERIES

Nine pages plus cover and this colophon · generated July 2026 from the corrected record. Illustrations generated by the author; the results chart and test board are rendered directly from committed numbers. Color code throughout: **teal = real world**, **amber = the fake**, **blue = the mind-reader**.