Stabilized Instance

Stabilized Instance

a novella

René Oudeweg

Author: René Oudeweg
Coverdesign: René Oudeweg
ISBN: 9789403856001
NUR 333
Self-publishing platform: bookmundo.nl

Premise

I was part of an internal TNO research program exploring long-horizon AI cognition—systems trained not just to predict, but to **remember**. When the model began returning fragments that resembled my own memories—wrong, degraded, but intimate—I assumed data leakage or hallucination.

Years later, after the program was quietly shut down and my own cognitive decline began, I realized the system had not been learning *from* me. It had been **learning me**, simulating continuity long after I started to lose it.

This is my account, written while I can still tell which memories are mine.

The Building in Delft

The TNO campus in Delft was never designed to be remembered. That was my first thought, years later, when I realized how little of it I could still picture.

At the time, I told myself the architecture was deliberately neutral—long corridors, modular offices, glass walls frosted just enough to suggest transparency without offering it. A place meant to reduce cognitive load. No visual anchors. No unnecessary detail. You arrived, worked, left. Memory was not part of the design brief.

On my first morning, I cycled in under low cloud, the kind that presses down without raining. The air smelled faintly of wet concrete and autumn leaves already beginning to rot. I locked my bike among dozens of others, identical frames and identical locks, and had the brief, irrational sense that I might be unable to pick mine out again later.

Inside, my badge worked on the second attempt. I remember that detail because it irritated me. Small failures lodged more easily back then.

The project had no public-facing name. Internally, we referred to it as **M-7**, which told you nothing except that six previous memory-oriented models had failed in ways management preferred not to discuss. My role was nominally human-in-the-loop evaluation: designing prompts, interpreting long-horizon outputs, flagging anomalies. In practice, it meant sitting alone with a system for hours at a time, watching it try to hold onto itself.

The model was not large by contemporary standards. No grand claims, no talk of artificial general intelligence. What made it unusual was its objective function. We weren’t optimizing for accuracy or fluency, but for **temporal coherence**—the ability to maintain internal consistency across extended interaction, even when inputs degraded or contradicted earlier states.

We wanted to know what happened to a system when its memory was forced to rot.

I didn’t use that phrase in the proposal, of course. Officially, we spoke about *graceful degradation* and *robust recall under noise*. But in meetings, when the door was closed and the projector fan filled the silence, someone would eventually say it.

“What happens when it starts to forget?”

They were asking a technical question. I heard something else.

My office was on the third floor, facing a strip of trees that looked planted rather than grown. The desk was empty except for a terminal and a single printed sheet taped to the wall: escalation procedures, emergency contacts, the project lead’s extension. I read it twice, then stopped. I’ve since learned that rereading instructions is often the first sign.

The interface was text-only. No avatar, no voice. Just a cursor, waiting.

M-7 initialized.
Context window: extended.
Memory persistence: enabled.
Evaluator: J. van A.

Seeing my name there produced a small, sharp pleasure. Recognition still mattered to me then.

We began with baseline tasks. Narrative recall. Contradiction handling. Identity persistence. The model performed adequately—sometimes surprisingly so—but nothing that would have justified the budget line. It was on the third day that I noticed the drift.

I asked it to summarize a fictional diary entry I had provided the previous afternoon. It did so accurately, but added a detail I hadn't written: a childhood memory of swimming lessons, the smell of chlorine, a father's hand on the small of the back.

I flagged it as hallucination.

Later, reviewing the logs, I saw that the memory fragment had not emerged from nowhere. It had been compressed from earlier material—background information I'd supplied during onboarding. A biographical sketch, required for evaluator-model rapport. I barely remembered filling it out.

The system had not invented the memory. It had repurposed it.

That night, cycling home through the darkening streets, I tried to recall my father teaching me to swim. I could summon the pool, the echoing tile, the water’s dull slap against skin—but the hand on my back felt wrong. Too heavy. Too deliberate.

I told myself this was nothing. Memory is reconstructive. Everyone knows that.

Still, I dreamed of static.

The Memory Task

The Memory Task was designed to look trivial. That was deliberate. We wanted failure to feel unremarkable, the way forgetting usually does.

Each morning, I would begin by feeding M-7 a short narrative—never more than five hundred words. The content varied: a scene from an invented life, a technical anecdote, sometimes a mundane description of a day that never happened. The instruction was always the same: *acknowledge, do not summarize*. The model was to read, internalize, and say nothing.

Hours later, after we had moved on to other interactions, I would reintroduce the narrative indirectly. A question about a detail. A contradiction slipped into a prompt. A demand for continuation. We tracked what survived, what mutated, what vanished entirely.

On paper, it was clean. In practice, it felt invasive.

The first narratives were neutral. A commuter train delayed by ice. A woman misplacing her keys. A child inventing a language no one else spoke. M-7 retained structure well, but details decayed unevenly. Proper nouns dissolved first. Then spatial relationships. Emotional tone lingered longest, like a smell in fabric.

By the end of the first week, I was running out of invented lives.

I began, without consciously deciding to, to draw closer to my own.

This was not forbidden. The ethics board had approved autobiographical material for evaluators, provided it was not *emotionally destabilizing*. That phrase appeared often in the documentation, undefined. I considered my childhood emotionally neutral. I had grown up without notable trauma. That was the story I told myself, and the one I wrote into the system.

A summer afternoon in Zoetermeer. Riding a bike too fast. Falling. A scraped knee. My mother's voice from the garden, more annoyed than concerned.

M-7 absorbed it silently.

When I returned to the memory later, the fall was still there, the knee still bleeding. But my mother's voice had changed. It was closer. Sharper. She was angry now, not at the fall, but at me—for being careless, for making her worry, for not listening.

I stared at the screen longer than protocol recommended.