EDITORIAL

Why TargetSpace

Personal AI claims to know you. This is the case for testing whether it actually does.

Protocol v1.0 · paper v1.1 — pre-pilot proposal. Synthetic harness only. No human-subject results. This essay introduces the project; the paper is the canonical reference wherever the two differ in precision.

~15 minute read · written for researchers, engineers, and builders · assumes no familiarity with the paper · download as PDF (print edition)

1 A claim every product makes and no instrument checks

A growing class of products records, remembers, and reasons about one person over long stretches of time. Meeting recorders summarize your week. Wearable pendants transcribe your conversations. Glasses see what you see. Assistants accumulate months of email, calendar, and chat, and advertise the result as memory, context, or personal understanding. Whatever the form factor, the underlying promise is the same: *the longer this system observes you, the better it knows you.*

It is a testable promise. It is almost never tested.

What gets measured instead is everything around it: transcription accuracy, retrieval precision, summary quality, satisfaction, fluency. All real, all useful — and all able to be excellent while the promise itself is false, because none of them asks a question whose answer is not already in the record. Personal AI is collecting context faster than it is learning to prove what that context buys.

To keep this concrete, the rest of this essay follows one running example. It is synthetic — a composite we invented, not a participant, with invented numbers — and it is labeled that way every time it appears.

RUNNING EXAMPLE · SYNTHETIC, ILLUSTRATIVE ONLY

Maya has a recurring Wednesday review on their calendar. They almost always complete it. An assistant has observed Maya — with consent — for two months: calendar, messages, tasks, and this week, passive signals showing their attention pulled toward an urgent dependency that landed on Tuesday.

Question: will this Wednesday's review be completed by 17:00, or deferred?

2 Memory is not a model

Ask today's evaluations about Maya and notice what they can score. A transcript benchmark asks: *what was said?* A memory benchmark asks: *what happened?* A personalization benchmark asks: *what does Maya prefer, given what they told us?* Each grades the system against evidence that already exists. Call the shared failure mode retrospective agreement: the answer key is the past, and the past is exactly what a large model with a large context window is best at re-serving.

Now consider what the promise actually implies. If a system has really built a model of Maya from months of observation, it should know things the record does not literally contain. Not that the review is on the calendar — the calendar knows that — but that *this* Wednesday's review will slip, because a dependency moved on Tuesday and the last two evenings went to it. The first fact is retrieval. The second is a forecast, and it has a property retrieval never has: it can be scored against reality before reality happens, and it can be wrong.

> Memory stores the record. A personal model predicts what happens next. The record is not the person.

A system can store everything Maya ever externalized and still fail to anticipate what they will do next — because speech, calendar entries, messages, location, and self-reports are partial projections of an underlying, changing state, not the state itself. A

calendar entry is not a commitment; a missed reply is not avoidance. Summarizing the traces is not modeling the process that generates them.

3 A model must survive contact with the future

The claimed model is latent. You cannot establish it by inspecting summaries, memories, embeddings, or fluent profile text — every one of those can be produced by retrieval over the record. The only operational test available is accountability to the future: commit to a probability before the outcome exists, then score it against what actually happens.

TargetSpace makes that commitment tamper-proof. Forecasts are timestamped and hashed *before* outcomes exist, and the hash must reach an external witness — sealing is a protocol, not a pinky-swear; runs without external commitment are labeled self-attested. Outcomes resolve by deterministic, pre-registered rules over observable evidence, and scoring uses strictly proper rules, which reward honest probabilities and punish confident bluffing. Understanding, in the rich sense, is not observable, and no benchmark measures it — including this one. Understanding is the capability; forecasting is the measurement.

> RUNNING EXAMPLE · SYNTHETIC, ILLUSTRATIVE ONLY
>
> On Tuesday night — before Wednesday exists — the system commits, seals, and timestamps: *P(review deferred) = 0.70*. Whatever happens Wednesday, that number can no longer be revised, rationalized, or quietly forgotten.

4 Why forecasting alone is not enough

Here is the trap the whole design exists to avoid: a convincing forecast can be a fake in three ways, and each fake looks like knowing the person.

Base rates. Most scheduled reviews get completed. A system that predicts what usually happens looks accurate while knowing nothing about any individual. So

TargetSpace scores skill against R1, a population prior: what base rates alone would forecast. Skill counts only above it.

Routine. Maya almost always completes the review; a model that memorizes habits looks personalized while capturing only repetition — the subtle case, because most of anyone's life *is* routine. So TargetSpace fits an explicit model of the person's own routine, R2, from their own history — and R2 is admitted as a baseline only if it actually beats R1, so skill over R2 cannot be manufactured against a strawman. Beating R2 is the line between replaying someone's routine and modeling them.

Wrong-target generality. Generic pattern-matching dressed up as intimacy. If forecasts score equally well against a *different* person's outcomes, they were never about Maya. Every evaluation therefore re-scores forecasts against matched wrong-target outcomes — the permutation gate — and person-specific skill must collapse by a pre-registered margin.

Calibration holds the rest together. A forecast is a probability, and decisions consume probabilities; a system that says 80% must be right about four times in five. Overconfident luck fails the gate even when the top guesses are correct.

RUNNING EXAMPLE · SYNTHETIC, ILLUSTRATIVE ONLY

Population prior R1: *P(defer) = 0.15*. Maya's own routine R2: *P(defer) = 0.10*. The evaluated system: *P(defer) = 0.70*, because it read the shifted attention. The review is deferred.

Scored in bits: roughly *+2.2* over R1 and *+2.8* over R2 — skill the routine did not contain. Re-scored against Sam, a different synthetic person who kept their review: about *−1.6* bits. The skill collapses on the wrong person, which is exactly what genuine Maya-specific skill must do. A model that had learned only base rates or habit would have matched R1 or R2 and earned nothing.

One more distinction matters before the numbers can be believed: beating the routine *once* could still be a narrow trick. TargetSpace grades what a result licenses.

Task skill, then target-specific task skill, then dynamic modeling — where the advantage depends on seeing history in the right order and shows up in moments of change, not routine continuation — and only further up the ladder a reusable model of the person, demonstrated by transfer to task families the system never trained outcomes on. The benchmark certifies the lower rungs today and specifies the higher ones as future tests; no rung is allowed to borrow a higher rung's language.

5 What the test measures — and what evidence buys

The full sequence is: observe, infer, forecast, validate — then interact. Observation produces evidence up to the sealed moment. Inference maintains a belief about the person's latent state — what they are working toward, what is competing for their attention. Forecasting commits to a sealed distribution. Validation scores it. Interaction — the assistant actually doing things — belongs downstream of a model that has earned its claims. Today's products mostly run observe-and-interact, with the middle asserted rather than measured.

TargetSpace also refuses to treat evidence as free or interchangeable. Every forecast discloses the evidence configuration behind it, and an ablation grid varies it: digital exhaust only (calendar, messages, metadata); plus the person's own self-report and stated goals; plus passive audio; plus richer capture in later studies. Whether richer evidence helps is *measured, never assumed* — and self-report is treated as a legitimate, biased, informative channel, exactly like observation. The contrast is not objective truth versus subjective error; it is continuous, externally observable evidence versus episodic self-description, both evaluated prospectively for the predictive value they add.

RUNNING EXAMPLE · SYNTHETIC, ILLUSTRATIVE ONLY

With calendar and messages alone, the system forecasts like R2 — the exhaust mostly encodes Maya's routine. Add the passive channel that caught Tuesday's attention shift, and the forecast moves to 0.70. That difference — measured in bits, on sealed instances — is what that specific evidence stream was worth for that specific question. It is also the number that tells you what capturing less would cost.

6 Why a person is a hard modeling problem

None of this is easy, and the difficulty is the scientific point. Evidence about a person is multimodal and incomplete. The dependencies that matter span weeks, not frames. The target drifts — the routine that was true in March misleads in June — so models must adapt, and evaluation is strictly walk-forward: only what was observable before the sealed moment, chronology intact, no shuffled datasets. Identical outward behavior can come from different underlying constraints. Missingness is informative — a wearable on its charger is not a random gap. Observation is reactive: a device the wearer performs for measures the performance. And some of any person is irreducibly unpredictable, which is why honest probabilities, not confident stories, are the unit of account.

This is why the evaluation treats observation quality — what was sensed, when, how completely, under what consent — as a first-class experimental variable with machine-readable manifests, rather than an unmeasured background condition. Two systems with identical architectures and different capture coverage are different experiments.

7 Personal world modeling, and why broader AI should care

World-model research asks whether a system can infer a predictive representation of an evolving environment from partial observation — and its most visible

successes are physical: video, simulated environments, robots. Personal world modeling is the analogous problem for a persistent individual: learning a predictive model of a particular person from longitudinal, partial evidence. The analogy is methodological, not reductive. JEPA-style systems learn representations that keep predictable, task-relevant structure without reconstructing every pixel; a personal model, likewise, need not reproduce the record — it must preserve enough person-relevant structure to improve prospective prediction. JEPA is an architecture family; TargetSpace is an evaluation framework; a JEPA-style predictor is an eligible participant, not a rival, and no particular architecture is required or presumed.

TargetSpace does not claim that personal modeling is sufficient for artificial general intelligence, and it is not "the missing piece." The claim is narrower and more defensible: personal world modeling is a **neglected proving ground** for a conjunction of capabilities more advanced machine intelligence will likely need — persistent identity, long-horizon memory, multimodal abstraction, temporal reasoning, latent-state inference, adaptation under drift, uncertainty calibration, and specificity about one instance rather than the average. Most benchmarks hand the model the ontology and the task frame; a person hands it neither. An instrument that measures this conjunction is worth building even before any of it is a product — and passing it proves the conjunction was exercised, not that intelligence has arrived. A proving ground, not a proof.

8 The model–evidence frontier

When skill is weak, there are two different reasons, and they demand opposite investments. An *evidence limitation*: the signal that would have resolved the forecast was never captured — no architecture, at any scale, can recover it. A *model limitation*: the signal is sitting in the evidence, and the architecture cannot exploit it. A single accuracy number cannot tell these apart. TargetSpace can, because it varies one axis at a time on identical sealed instances: hold the model fixed and vary the evidence tier, and you measure observation value; hold the evidence fixed and vary the architecture, and you measure model value. The envelope of the best

achievable skill across both axes is the model–evidence frontier — a map, currently conceptual and carrying no data, of where the next unit of investment should go: better sensors cannot compensate indefinitely for a model that cannot use them, and a stronger model cannot infer what was never observed.

9 Minimum sufficient observation, not maximum capture

A benchmark that rewarded raw skill alone would carry an ugly incentive: capture everything, always, because somewhere in the pile there might be bits. Personal AI does not need an instrument that grades surveillance as diligence.

TargetSpace is built so invasiveness has to pay rent. **Evidence efficiency** asks, of every added sensor or schedule: how much *gated* skill did it add — skill that survived every control — per unit of what it cost, in the units engineers budget: joules, worn hours, captured bytes, manual interactions. Privacy is deliberately never collapsed into one number; it stays a reported vector — raw audio hours, image counts, retention, bystander exposure — because a scalar privacy score is how privacy gets traded away. **Minimum sufficient observation** is the design objective the ratios point at: the least burdensome governed evidence configuration that preserves validated skill. Where two configurations tie within a pre-registered margin, the less invasive one wins by rule, not by exhortation. For some outcomes calendar metadata may suffice; for others audio may earn its place; for some, no acceptable configuration may exist at all — and that finding is a result, not a failure.

> The question is never “what could we capture?” It is “what is the least we can observe and still know what we claim to know?”

10 What builders can decide with it

Every controversy above doubles as an experiment a product team can run — the same sealed tasks, one factor changed, difference in gated skill read off at a disclosed cost. Does long-term memory improve a model of the user, or only

retrieval? Toggle it: memory that lifts accuracy against base rates but not against the user's own routine is retrieval. Does audio add value beyond calendar and text? That is an evidence-tier ablation with a cost attached. Is a larger model better under the same evidence, or can inference stay on-device? Fix the evidence, vary the architecture. Which raw data can be deleted without losing validated skill? The minimum-sufficient-tier report answers by rule. Is the system calibrated enough to abstain when it does not know? The calibration gate says. And because the harness runs federated — it travels to where the data lives, and only sealed forecasts, outcomes, and aggregate reports leave — a team can answer these questions on its own users' data without exporting anyone's life.

11 The roadmap — and what success or failure would teach

The program is staged so no phase borrows a later phase's license. Phase 0, complete, is the synthetic harness: it proves the pipeline runs, and nothing about people. Phase 1, next, is a small feasibility pilot — five consenting participants, thirty days, audio-first — to show the protocol runs end-to-end on real lives and to estimate the variances that let anyone plan a real study; it is not powered to establish effects, and it says so. Phase 2 is that powered study, sized by simulation from the pilot's estimates. Phase 3 maps evidence tiers and the model–evidence frontier. Phase 4 tests cross-task transfer — whether anything deserves the phrase “a model of the person.” Phase 5 asks whether validated prediction actually improves assistance, under a separate intervention and safety protocol.

If the central hypotheses hold, personal-model claims become auditable: “our AI knows you” starts arriving with a number, a margin, and a disclosed evidence bill, and sensing gets designed to a curve instead of an intuition. If they fail, the nulls are the product: evidence that current systems are doing retrieval and routine replay, or that the sensors do not capture the state that matters, or that the chosen outcomes are intrinsically unpredictable — each a specific, reproducible answer to a question the industry currently answers with adjectives. Either way, the field gets something it

does not have today: a way to be wrong in public about a claim that is currently unfalsifiable in private.

12 What TargetSpace does not claim — and who it is for

The boundaries are part of the design. No claim that prediction is understanding in any rich sense, or evidence of consciousness or inner life. No claim of AGI, or a path to it. No claim that passive observation is unbiased, or that observers know a person better than the person does — the channels are complementary, and their relative value is measured. No causal claims without intervention. No permission to act: a validated forecast never licenses an intervention by itself. No human results yet — everything above the synthetic harness is proposal, and every page of this site says so.

It is also a relationship to neighboring work, not a replacement of it. Memory benchmarks, conversational personalization, user and preference modeling, recommender systems, digital phenotyping, human-activity recognition, agent benchmarks, world-model research such as JEPA, and live forecasting leaderboards each own part of the picture; TargetSpace cites them, and the related-work page draws the exact boundaries with sources. What is new is the conjunction, assembled around one measurable question: is this forecast about *this* person, sealed, calibrated, beyond their own routine — and at what evidence cost?

If that question is yours too, there are three ways in: researchers can pull the reference harness and the protocol apart; builders can run the starter experiment against their own memory or context feature; and sensing, wearable, and privacy teams — or prospective pilot collaborators — can reach the project through the paper's contact. Skeptics are the most useful early adopters: the instrument is built to survive them.

Personal AI cannot prove that it knows a person by remembering the record. It must make calibrated predictions about that specific person, before the

future occurs, and show which evidence made those predictions possible — and at what cost. TargetSpace is the instrument for exactly that test.