

THE ONE FACE YOU CANNOT READ

On Self-Opacity, and What It Would Take to Build the Instrument

AUGUST BROOKS

The system described in the second half of this essay was conceived on 17 March 2025. Whether it ever gets built is a separate question from whether the reasoning behind it is sound, and I would rather the reasoning exist in public than wait on the resources to ship the product.

I. The Asymmetry

You can read a face. You do it constantly, involuntarily, and faster than thought — hunger, cold, grief, the particular flat exhaustion of a person who has stopped expecting help. Nobody taught you. There is no manual and you have never needed one.

Now turn it around. Ask why you said the thing you said in that argument last Tuesday. Ask why you trust the person you trust, or why one topic makes your jaw tighten and another does not. You will produce an answer immediately, fluently, with confidence.

The answer will very often be invented.

This is not a moral failing and it is not unusual. It is the best-documented finding in the psychology of introspection. Nisbett and Wilson laid it out in 1977: when people are asked to explain their own mental processes, they do not consult those processes, because the processes are not available for consulting. They construct a plausible story from what they know about themselves and about situations like this one, and they experience that story as a memory rather than as a construction. In their shopping-display experiment, position on the table drove the choice; every shopper explained it by the quality of the goods, and none of them accepted the real explanation when it was offered.

It gets worse, and more precise. In 2005 Johansson and colleagues ran a study where people chose which of two faces they found more attractive, and were then handed back — by sleight of hand — the face they had *rejected* and asked to explain their choice. Most did not notice the switch. They explained, at length and with feeling, a preference that was not theirs. They gave reasons for a decision they had never made.

Gazzaniga found the same machinery in split-brain patients: the verbal hemisphere, cut off from the information that actually drove a behavior, generates an explanation anyway, instantly and without any sense of straining. He called it the interpreter. It does not know when it is guessing. It has no signal for that.

So the instrument that reads other people so effortlessly has one blind spot, and it is the only spot that matters for the person operating it. You are inside the one system you cannot get a look at from outside.

II. Why Nothing We Have Fixes This

Every existing self-knowledge tool has the same defect: it asks the broken instrument to report on itself.

A personality inventory is self-report. It asks how organized you are, and the part of you that answers is the part with the interest in the answer. Type systems are worse, because they hand back a compact identity that then works as an explanation for anything you do afterward — you are no longer observing yourself, you are performing a category.

Self-help asks you to identify your limiting beliefs. But a belief that is genuinely load-bearing does not present itself as a belief. It presents itself as the way things are. The ones you can name are not the ones costing you anything.

Therapy actually works, and I do not want to be read as saying otherwise. But it operates on what you bring into the room, and what you bring is already filtered — selected, edited, arranged into a narrative for another person. It also happens for one hour at intervals, which is a very sparse sampling rate for a process running continuously.

Journals are written by the part of you that wants to be a certain kind of person. Social media is that same impulse with an audience and a scoreboard.

Then there is the trap underneath all of it. In 1949 Bertram Forer handed his students a personality profile he said was individually prepared for each of them. It was one generic text, identical for everyone, assembled from an astrology column. They rated it, on average, above four out of five for accuracy. That is the Barnum effect, and it is not a curiosity — it is the entire economics of this category. **People rate a description of themselves as accurate in proportion to how much they like it.** Any product that measures success by user satisfaction will be pushed, by ordinary market feedback and with no bad intent from anyone, toward telling people flattering things. The horoscope is not a failure of self-knowledge tools. It is their equilibrium.

III. What Would Actually Have To Be Built

If the reports are unreliable, stop scoring the reports. Score the behavior of producing them.

Measure the process, not the content. What someone says about themselves is a claim. How they say it is evidence. Where the sentence stalls. What gets restarted, hedged, joked away. Which questions get answered with a story about someone else. Which subject reliably produces a change of topic. None of that is available to the speaker in the moment, and none of it depends on the accuracy of self-report, because it is not self-report at all — it is behavior, recorded while it happens.

Speech, not typing. Typing is editing. The keyboard gives you a revision buffer between the thought and the record, and everything interesting gets revised out. Spoken answers preserve the hesitation, the false start, the sentence that changes direction halfway through. That single interface decision is most of the difference between this and a questionnaire.

Longitudinal, not a snapshot. One session gives you a profile, and a profile is a horoscope with extra steps. The value is in the second derivative: what changed. You said in March that you did not care what your brother thought. It has come up unprompted nine times since. That is not a personality reading. It is a fact about a record, and it is checkable by the person it is about.

Report deltas, not diagnoses. The output should be closer to a bank statement than a fortune. *Here is what you said. Here is when. Here is what you say now.* No labels, no types, no scores that become identities. The user does the interpreting; the system supplies what they could not have retained on their own, which is the whole of what a memory-limited creature actually needs from a mirror.

IV. It Has To Be Built To Disappoint You

This is the entire design problem, and it is why I am not sure anyone will build it properly.

Every incentive in the product pushes toward validation. Flattering readings get better ratings, better ratings get retention, retention is the business. Left alone, the system converges on a machine that tells you that you are perceptive, misunderstood, and on the verge of a breakthrough — and you will rate it five stars, because Forer's students did.

So the constraint has to be structural rather than aspirational. The system's job is to surface the thing you have been avoiding, which means a well-functioning session is one you do not enjoy. That is a deliberate retention penalty, accepted at the design stage, and it is also the

only defensible moat in the category: any competitor optimizing for delight is building a different and lesser product, and cannot follow you without destroying their own numbers.

It also implies a revenue model that does not depend on how good the output makes you feel. The moment the reports are the upsell, the reports get warmer. That is not cynicism about anyone's motives; it is just what happens to a metric when money is attached to it.

V. What I Concede

An argument that admits no cost is an advertisement, so here are the costs.

The machine confabulates too. A language model reading your transcripts is a pattern-matcher producing a plausible story about you — structurally the same failure the essay opened with, moved outside your skull. The mitigation is to keep the system as close to the record as possible: quote the person back to themselves, timestamp everything, prefer *you said this, then this* over *you are the kind of person who*. The further the output drifts from quotation toward interpretation, the more it becomes the same horoscope with better tooling.

It could install a new authority. A person who outsources their self-assessment to a system has not gained self-knowledge; they have swapped one unexamined source for another. Anything that produces confident verdicts about who someone is risks becoming exactly the kind of external ratification I have spent other essays arguing against. The output has to stay evidentiary, and the interpretation has to stay with the user, or the tool defeats its own purpose.

The clinical line is real and it is not a footnote. Sit anyone down for months of unstructured spoken reflection and some of them will disclose genuine crisis. A system that surfaces that and responds with a monthly insight report is worse than one that never asked — it obtained a disclosure and did nothing with it. Any honest version of this needs detection, an immediate handoff to actual crisis resources and human beings, and hard limits stated to the user before they start: this is not treatment, it does not diagnose, and here is precisely what it will do if it hears something serious. I would rather ship a narrower product than a version of this that quietly positions itself as a substitute for care.

And the data is the most sensitive that exists. A continuously updated record of how a named person actually thinks is worse to lose than their financial records. It should live on the user's device, or under encryption the operator genuinely cannot read, and if that makes the analytics harder then the analytics get harder. Anything else is a breach waiting to happen to people who trusted you with the inside of their head.

VI. Why It Is Worth Doing Anyway

The moral instrument I have argued for elsewhere works by reading a face. It is fast, it is untaught, and it is largely reliable. It has exactly one gap: it cannot read the face it is mounted behind, which is the one whose conduct you are actually responsible for.

Every method we have built to close that gap routes through self-report, which is precisely the channel the research says is unreliable. That was an unavoidable limitation for the whole of human history, because closing it would require something that could hold a complete, unedited, timestamped record of how a person actually reasons across months and hand it back without flattery. No person can do that for another. No person can do it for themselves.

A machine can. That is a recent development, it is not a small one, and almost everyone building in this space is using it to generate compliments.

I do not claim the tool would fix anyone. People are not fixed by information; they are changed by seeing something for themselves, and most who see it clearly will still choose the comfortable thing. But a small number would not. And the difference between a person who can name the pattern running their life and a person who cannot is, in my experience, most of the difference between the two lives.

That is the case. Whether it gets built is a question of resources and will. Whether it *should* be built the way the market wants it built is not a close call at all.

THE CANONICAL VERSION OF THIS ESSAY IS AT [AUGUSTBROOKS.COM/ESSAYS/THE-ONE-FACE-YOU-CANNOT-READ](https://augustbrooks.com/essays/the-one-face-you-cannot-read)