

# The Hypothesis Generation Machine: AI-assisted Multimodal Idiographic Phenotyping as a New Context of Discovery.

Max Saris, Faculty of Medicine, Psychology, 020465421 – [max.saris@helsinki.fi](mailto:max.saris@helsinki.fi)

09.09.2026

## The Epistemic Asymmetry of Discovery & Justification

Contemporary science has an often-overlooked methodological asymmetry. While science has developed sophisticated frameworks for hypothesis testing, such as null-hypothesis significance testing and high-powered replication designs, hypothesis generation remains largely informal. Hypothesis generation is treated as something methodologically unformalizable, often attributed to theoretical intuition, retrospective observation, or informal clinical impressions (Ludwig & Mullainathan, 2024).

In Reichenbach's (1938) classical epistemology, scientific inquiry can be broken down into the *context of discovery*, the generative phase where novel patterns and explanatory candidate models emerge, and the *context of justification*, the empirical phase where these hypotheses are subjected to rigorous controlled testing and falsification. Modern psychology skews its rigor toward the latter, leaving the context of discovery methodologically vague.

This essay argues that the combination of dense, multimodal idiographic phenotyping and artificial intelligence provides the foundation for a systematic, computational context of discovery. This paper coins the term **Hypothesis Generation Machine (HGM)** to describe this computational framework. In essence, an HGM could leverage dense, longitudinal, single-subject  $N=1$  data to detect temporal dynamics, cross-modal interactions, and other potentially non-linear behavioral or cognitive shifts that exceed the limits of nomothetic designs and human cognitive capacity to detect. These candidate patterns generated by HGMs are not meant to replace empirical validation; rather, they serve as an alternative context of discovery, which conventional empirical experimental designs can validate.

## The Ergodic Barrier & Need for Idiographic Data

The necessity for dense  $N=1$  idiographic data stems from a constraint within behavioral science: the ergodic assumption. Psychological research has historically operated under the assumption that group-level averages accurately reflect intra-individual changes within specific persons over time. However, research has shown that it is seldom the case. As Molenaar (2004) argued, psychological processes generally violate the axioms of ergodicity: humans are non-stationary systems with exceptional heterogeneity. Group averages rarely translate to within-person temporal dynamics. Fisher, Medaglia, and Jeronimus (2018) demonstrated this empirically across six independent datasets for multiple psychological domains. When tracking individuals over time on measures like anxiety and depression, the patterns within individuals rarely matched the group averages. Because group-level studies

average everyone together, they tend to smooth out the temporal dynamics of how an individual actually changes over time.

While between-person studies are excellent at telling us *who* is more anxious or depressed compared to others, they lack the power to tell us *when* or *why* any given individual changes. As Wright and Woods (2020) emphasize, to understand within-person dynamics, we need to shift to personalized, intensive longitudinal data that tracks an individual at high density over time.

### **From Lifelogging to Multimodal Phenotyping**

Capturing intra-individual processes accurately requires an infrastructure capable of tracking both subjective experience *and* objective markers, such as physiology and behavior. The challenge lies in doing both at the same time, with historical examples typically falling into either of the unimodal extremes.

The diary of Robert Shields sets the precedent for sheer temporal resolution and narrative depth, with Shields documenting everyday activities at approximately five-minute intervals for over two decades (Shields, 1972-1995). Shields achieved unprecedented semantic density, but lacked the concurrent physiological or behavioral monitoring that would have allowed genuine statistical inference across modalities.

At the opposite extreme, Gordon Bell's *MyLifeBits* project captured millions of passive digital events, from keystrokes, scanned documents, audio recordings, to automated camera captures, over 17 years (Gemmell, Bell, & Lueder, 2006). While technologically comprehensive, Bell's project lacked the psychological and narrative depth of Shields' diary that could have elevated the project beyond a technological demonstration into a usable intra-individual developmental dataset, or with the addition of artificial intelligence, into an HGM.

Existing lifelogging projects demonstrate the feasibility of extensive personal-data capture, a modality of data collection that is becoming ever-more practical with modern technological advancements, and remains relatively untapped within contemporary science. Prototypes such as the MS-Archive (Saris, 2026) serve as an illustrative proof-of-concept for modern multimodal idiographic phenotyping, combining continuous prospective daily narrative journals and structured psychometrics with continuous wearable telemetry (e.g., sleep architecture, autonomic markers), passive activity tracking, environmental sensing, and even genetics, across multiple years of N=1 data.

Modern intensive longitudinal designs resolve the historical limitations of Bell and Shields by integrating these disparate streams through active Ecological Momentary Assessment (EMA) and ambulatory monitoring (Trull & Ebner-Priemer, 2013). Bringing subjective narrative diaries, passive digital data, and continuous biometrics into the same database enables multimodal triangulation.

Triangulation cross-references subjective psychological states with objective physiological and behavioral data. While collecting data prospectively in the moment minimizes retrospective recall bias, concurrent biometric data provides an objective anchor against state-dependent reporting distortions and social desirability bias. This increased validity is useful for generating plausible hypotheses, as any two factors are bound to correlate in a sufficiently large dataset. More data streams expand the hypothesis space, but they also increase the risk of spurious associations and multiple-comparison problems. Multimodal triangulation should therefore be treated as a criterion for prioritizing hypotheses that receive convergent support across relatively independent measurements.

Furthermore, increasing the variety and temporal density of our data streams broadens the searchable hypothesis space, providing an HGM with the data required to find non-obvious relationships across modalities. In essence, the more unique vectors of data we have, and the denser the data points, the more likely an HGM is to find intra-individual patterns, thereby providing human auditors with a larger set of candidate hypotheses for further study. We could even adjust the triangulatory parameters to reduce the number of hypotheses an HGM generates, limiting the output to only the patterns with substantial triangulatory evidence. Conversely, if we require more hypotheses, we could reduce the needed triangulatory evidence at the cost of more false positives.

## **HGM Architecture**

N=1 cases have played a role in many of psychology's biggest discoveries, from the case of H.M. laying the foundation of our taxonomy of memory systems (Scoville & Milner, 1957), to Phineas Gage (Harlow, 1868) contributing to the foundational neurological hypotheses concerning the frontal lobe's involvement in executive functions and personality localization. However, these historical case studies were sparse, post-hoc, and analog. Hypotheses generated by them relied almost entirely on clinical intuition and retrospective observation, inevitably focusing on simpler, obvious patterns and leaving much of the potential hypothesis space unexplored.

An AI-powered modern archive, by contrast, is real-time, multi-modal, and continuous, resolving many blind spots human researchers could not see. In the same way AI-based mammographic image analysis has demonstrated performance exceeding that of humans alone in diagnosing breast cancer by noticing patterns that human perception often overlooks (McKinney et al., 2020), an HGM could potentially find latent hypotheses that human cognition would struggle to find on its own.

Humans are governed by biological limitations of working memory, confirmation bias, and a propensity to search only within familiar conceptual paradigms. An AI, by contrast, operates across hundreds of parallel variables, unbound by human cognitive bottlenecks, potentially able to operationalize and classify incompatible data (e.g., extensive narrative text with physiological telemetry) that would otherwise take human researchers weeks to compile. AI is scalable: it could run these conversions in parallel at a fraction of the time it would take a human to do the same, thus being able to compare a larger set of otherwise incompatible data streams and thereby expanding the potential hypothesis-space substantially.

Using artificial intelligence has already been shown to be an effective way to generate novel behavioral hypotheses (Ludwig & Mullainathan, 2024), but the specific combination of AI with extensive N=1 multimodal data is an application that contemporary science could explore further.

As for how exactly an HGM operates, the pipeline follows a systematic structure: dense N=1 multimodal data is fed into the AI; the AI detects candidate patterns or temporal interactions across data streams; human researchers then vet those patterns for plausibility; and classical experimental or nomothetic trials test them for generalizability. Thus, an HGM acts merely as a tool within the context of discovery, not a way to establish empirical truth in itself.

### **Limitations, Future Trajectories, and Conclusions**

An HGM, or specifically the personal N=1 multimodal database that acts as its foundation, comes with numerous limitations that unfortunately exceed the scope of this essay to cover. The hypothesis explosion and false positives are limitations that can be remedied by triangulatory parameters, but the legal and ethical limits of personal data are challenging facets that require extensive frameworks to solve. Anonymizing becomes unfeasible when the data itself (e.g., the semantic subjective narrative) is a source of identification even without classical personally identifiable information (PII). Scrubbing the narrative until persons are no longer identifiable scrubs the data itself, poisoning the very utility of the database.

Nevertheless, extensive N=1 data remains a tantalizing source of high-grade information, which warrants its further study and the development of more robust mechanisms for its proper handling.

When HGMs become more ethically and legally viable, a future trajectory might involve the training of the AI models themselves, to hone their ability to not only suggest hypotheses, but to also evaluate the potential usefulness of said hypotheses. Instead of training our AI on images of X-rays and diagnostic information, we could train it on past hypotheses and the results of their studies. This way, our system may be trained to ‘diagnose’ good hypotheses that researchers alone may overlook. This would streamline the evaluation of generated hypotheses, but it will remain conceptual until we formalize the initial methodology.

Ultimately, an HGM does not seek to replace traditional nomothetic science, nor does N=1 discovery replace population-level validation. Discovery and justification are complementary processes, both made more robust through elevated methodological rigor. By removing the constraints placed by human cognitive bottlenecks and unformalized intuition, and by leveraging dense, multimodal idiographic datasets as an empirical playground for AI, an HGM provides a systematic way to generate non-obvious psychological questions to fuel the empirical research of the future.

## References

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27), E6106–E6115. <https://doi.org/10.1073/pnas.1711978115>

Gemmell, J., Bell, G., & Lueder, R. (2006). MyLifeBits: A personal database for everything. *Communications of the ACM*, 49(1), 88–95. <https://doi.org/10.1145/1107458.1107460>

Harlow, J. M. (1868). Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society*, 2(3), 327–347.

Ludwig, J., & Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2), 751–827. <https://doi.org/10.1093/qje/qjad055>

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashraffian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Ledsam, J. R., Kelly, C. J., King, D., & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>

Molenaar, P. C. (2004). A manifesto on psychology as an idiographic science: Bringing the person back into scientific psychology. *Measurement: Interdisciplinary Research and Perspectives*, 2(4), 201–218. [https://doi.org/10.1207/s15366359mea0204\\_1](https://doi.org/10.1207/s15366359mea0204_1)

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge*. University of Chicago Press.

Saris, M. (2026). *MS-Archive: A multimodal longitudinal idiographic database*. <https://maxsaris.com/ms-archive/>

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, 20(1), 11–21. <https://doi.org/10.1136/jnnp.20.1.11>

Shields, R. (1972–1995). *The Robert Shields papers* [Unpublished diary]. Manuscripts, Archives, and Special Collections, Washington State University Libraries.

Trull, T. J., & Ebner-Priemer, U. W. (2013). Ambulatory assessment in psychopathology research: Current state and future directions. *Annual Review of Clinical Psychology*, 9, 349–376. <https://doi.org/10.1146/annurev-clinpsy-050212-185510>

Wright, A. G., & Woods, W. C. (2020). Personalized models of psychopathology. *Annual Review of Clinical Psychology*, 16, 49–74. <https://doi.org/10.1146/annurev-clinpsy-102419-125032>

### **Acknowledgement of AI Use**

AI (ChatGPT 5.6 & Gemini 3.7 (Extended), Gemini 3.8 Flash (Extended)) was used in this essay for two primary functions.

1. AI was used in the initial brainstorming to come up with potentially relevant articles and terms for the topic the author chose to write about (HGMs, a conceptual framework developed around the author's own prototype database)
2. AI was used as a conversational peer to review and critique drafts sequentially for grammatical errors

Examples of prompts used:

- “What are relevant scientifically validated studies I can use here?” [long list of the author's brainstorming notes and initial framework presentation of HGMs attached to prompt]
- “Please provide valid sources for studies covering Bell and Shields”
- “Are there any grammatical errors here?” [paragraph attached to prompt]

All core concepts, the HGM theoretical framework, original arguments, and final prose were developed and finalized by the author.