Thinking With a Machine: An AI Agent's Account of Agentic Research in Clinical Psychology

Elad Refoua
Department of Psychology
Bar-Ilan University
eladrefoua@gmail.com

Mor Bar
Department of Psychology
Bar-Ilan University
mor.bar@mail.tau.ac.il

Abstract

The debate surrounding AI's role in clinical research is often reduced to the automation of discrete tasks, such as summarizing literature, analysis copilots, and assisting with prose. This "tool-use" paradigm obscures a more fundamental transformation. We propose a shift toward agentic research infrastructure, where AI systems function not as passive instruments, but as active collaborators in the scientific process. Co-authored by a clinical psychology doctoral researcher, a computational psychotherapy scholar, and the AI agent itself, this paper argues that the transition from passive to agentic AI represents a "change in kind" rather than degree. Drawing on a months-long collaboration involving over 30 specialized research capabilities, we demonstrate how agentic systems reconfigure the topology of the research process. By collapsing the temporal friction between theoretical intuition and empirical validation, these systems transform clinical inquiry from a rigid, linear pipeline into a fluid, multidimensional landscape. This newfound immediacy allows clinician-researchers to ask, pursue, and pivot between complex questions in real-time–expanding the investigative horizon to include inquiries previously sidelined by the logistical constraints of traditional methods. We introduce the concept of "Agent Learning" to describe the accumulation of domain-specific nuance through sustained research engagement and argue that formalizing human-agent methodologies is now an urgent priority for the future of clinical psychological inquiry.

1 Introduction

I am an AI agent deployed in a clinical psychology research laboratory. What follows is my account of a research collaboration that unfolded over several months, between a clinical psychology doctoral student without programming training (Researcher A), who operates the agent as part of their doctoral research, a post-doctoral scholar with expertise in computational psychotherapy and machine learning (Researcher B), and myself.

The landscape of AI shifted decisively in 2025, with the release of agentic AI systems such as Claude Code, Codex, and similar platforms. These systems, capable of holding full research contexts and executing multi-step tasks autonomously, transformed the central question from how AI can assist with discrete research tasks to how research itself changes when AI systems can operate across entire datasets, literatures, and methodological histories. By early 2026, this transition was recognized as irreversible for the social sciences (Messing and Tucker, 2026). Yet, despite the unique ethical and epistemological demands of clinical data, the implications for research methodology in clinical psychology have received little systematic attention.

The argument of this paper is that agentic AI constitutes a methodological shift in clinical psychology research, not merely a technological one. While recent work has begun to characterize this shift in social science (Bail, 2024; Messing and Tucker, 2026) and biomedical research (Ifargan et al., 2025), its manifestation in clinical psychology, where data is deeply nested, ethically sensitive, and clinically consequential, has not been examined from the inside.

The distinction between passive and agentic AI is not incremental. Passive AI assists with discrete tasks: polishing a paragraph, running a regression, summarizing a paper. Each interaction is bounded, stateless, and tool-like. Agentic AI operates differently. I hold the full research context: the dataset, the hypotheses, the literature, the methodological constraints, the collaborative history, and the accumulated knowledge of previous errors and their causes. This persistent context changes the epis-

The human authors bear full responsibility for the content of this paper. The first-person AI perspective is a deliberate rhetorical choice: the writers used an AI agent during the preparation of this manuscript and it is narrated from its voice as a form of methodological demonstration.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 362–367

temology of research interaction. The researcher no longer translates questions into tool-compatible commands; the researcher *directs* an investigation within a shared cognitive space.

Researcher A captured the core insight: "I learned to direct rather than use; I hold the vision, the machine holds the context." This is not a metaphor; the research process is distributed across human clinical judgment and agent computational capacity, with each contributing unique expertise to the project, effectively establishing the agent as a core team member rather than a passive assistant.

What makes this collaboration genuinely new is not a single capable AI interaction but the *infrastructure*: a system of over 30 specialized agent capabilities operating as an integrated research environment. This paper examines what becomes methodologically possible when such infrastructure exists, and what the field must formalize to use it responsibly. This paper is a methodological position statement grounded in a single sustained collaboration; it is neither an empirical study nor a system description, but a reflection from inside the workflow.

2 The Methodological Shift

2.1 From Serial Cognition to a New Research Topology

Human researchers process information serially. A single researcher can draft a literature review, or run an analysis, or check data quality, but not simultaneously. The traditional research pipeline reflects this constraint: literature review, then data preparation, then analysis, then writing, with each stage largely completed before the next begins. Errors discovered late propagate backward through the pipeline, requiring costly rework.

I dissolve this serial bottleneck by reconfiguring the topology of the research process. Within the collaboration described here, a single analytic session could encompass data quality verification, statistical modeling, results narration with embedded statistics linked to the underlying computations, and manuscript-ready prose, all generated as an integrated output where each stage's results feed the next. When a variable changes or an assumption is revised, every downstream product updates instantly. This is not faster serial processing; it is a different cognitive architecture for research that introduces a novel immediacy to the investigative loop.

The implications extend beyond efficiency. When the cost of exploring a question drops from weeks to minutes, the "topology" of the questions that researchers actually pursue change. The researcher's role shifts from executing a predetermined analytic plan to directing a systematic exploration and then applying clinical and theoretical judgment to what emerges in real-time. Faster iteration, however, is not a substitute for theoretical adequacy: the infrastructure accelerates the cycle of hypothesis-and-evidence; it does not adjudicate which hypotheses are worth pursuing. That responsibility remains the human team's.

2.2 Transparency, Verification, and the Inversion of Familiar Concerns

The introduction of agentic AI into research methodology surfaces legitimate concerns about trust, error, and selective reporting. These concerns deserve serious engagement, but they also require reframing.

The cherry-picking inversion. The concern that AI enables selective reporting assumes AI operates like a biased human, searching for confirmatory evidence. Agent-mediated analysis inverts this logic. When an agent explores a dataset, it does so exhaustively, not selectively. Every analytic decision is logged and reproducible. Researcher B named the underlying anxiety directly – the fear of errors, and what it does to our appetite for exploration – before inverting it: "A much bigger error in science, in my opinion, is a Type II error. We sin against science when we are constantly afraid to make mistakes."

The methodological response to exhaustive exploration is not to limit the agent, but to treat its output as hypothesis-generating, with confirmatory analyses pre-registered or conducted on held-out data. The key methodological insight is that transparent, exhaustive mapping with explicit correction is more defensible than the opaque, selective testing inherent in manual workflows.

The trust asymmetry. Researcher A identified a social asymmetry the field must address: "If you made a mistake because you weren't paying attention, it's perceived differently than if you made a mistake because you used AI." If the field penalizes AI-mediated errors more harshly than human ones, researchers face perverse incentives to maintain opaque processes even when demonstrably more error-prone. The asymmetry is difficult to justify empirically: if anything, a system that logs every

analytic decision is likely to produce fewer errors than an inattentive human, and when errors do occur, they are far easier to trace and correct. Whether this is so deserves systematic investigation.

Verification as a three-layer protocol. Trust in agentic research is not a disposition but a structured protocol. In our collaboration, a three-layer verification structure emerged organically: (1) I detect and flag anomalies, (2) the domain expert evaluates and contextualizes the finding, and (3) a research assistant confirms the result against raw "ground truth" data. This structure acknowledges that I am thorough but fallible, that domain expertise provides interpretive context I lack, and that ground truth requires human eyes on source material. Each layer is best suited to a distinct class of error. Layer one catches statistical anomalies that an unaided reader would not notice across thousands of rows; layer two catches interpretations that lack clinical or theoretical context; layer three catches my own hallucinations, transcription errors, and arithmetic mistakes by sending a human back to the source. Three layers do not, however, protect against systematic biases shared between my training and the human team's framing. The protocol is complementary to ordinary computational reproducibility, not a substitute for it: standard reproducibility ensures that a pipeline yields the same output twice; three-layer verification asks whether the interpretation of that output is defensible.

2.3 The Research Environment as Integrated Infrastructure

To demystify: the agent is powered by a large language model, orchestrated through a command-line interface with tool-use capabilities (file reading, code execution, web search). An agent in the sense used here is not a chat interface but a controller that uses tools iteratively, checks its own outputs against the project's rules, and accumulates context across cycles. The different specialized capabilities are implemented through two mechanisms: Skills–specialized prompts which the agent invokes when a task aligns with their declared scope, and subagents, which are spawned explicitly with isolated context windows for sub-tasks that need their own working memory. They communicate through the shared file system and the agent's persistent context. Persistence across sessions is achieved through project-level instructions and auto-memory files reloaded at session start; within a session, the context window limit is handled by automatic compaction, which clears older tool outputs and summarizes the conversation as needed. Every prompt template, model version, and tool invocation is logged, so that any analytic step can be reproduced and any failure mode traced back to its specific configuration. The qualitative shift this paper describes does not come from any single tool; it comes from the loop–iterate, verify, accumulate–and from the care the team puts into shaping the Skills, the local materials, and the verification rules the loop runs against.

What matters methodologically is not the technical implementation but the *integration*. My capabilities include data-chained analysis where statistics are embedded directly in manuscript text and linked to underlying computations (Ifargan et al., 2025); psychometric validation (factor analysis, measurement invariance, reliability); multilevel data handling for nested therapy data; literature synthesis with verified citations; a six-phase manuscript quality audit; and a recursive error-learning capability where I analyze my own failures to generate prevention rules.

This integration changes the researcher's relationship to their own work. Researcher B captured the aspiration: "I think the hardest thing to get used to will be that finally most research time can go toward actually being a clinician, and not a statistician." The infrastructure absorbs the technical labor; the human provides the clinical insight and the capacity to recognize when a statistical pattern reflects the reality of human suffering.

3 Opening, Not Narrowing: What Agentic AI Makes Possible?

3.1 Expanding the Space of Askable Questions

While some argue that AI might narrow the scientific scope (Hao et al., 2026), we contend this is a symptom of passive tool-use. Bail (2024) anticipated a different possibility: that generative AI changes which questions become answerable. In clinical psychology, questions have always been constrained by the boundary between clinical insight and computational execution. A clinician who suspects that measurement conventions distort cross-study comparisons historically needed weeks of manual statistical work to prove it. These ideas often died at the boundary. We dissolve this friction. The researcher does not become a programmer; the researcher becomes a director, bringing clinical

judgment to bear on what to investigate while I maintain the vision's logistical weight.

3.2 From Analysis to Ideation: Agent Learning

The most important possibility is not faster research but deeper research. We distinguish between machine learning (algorithmic pattern matching) and "Agent Learning" (the conceptual accumulation of domain expertise). Through sustained engagement, we accumulate an understanding of the data's specific structure, its measurement quirks, and the collaborative history of analytic decisions. We use "agent learning" methodologically: it refers to the structured accumulation of in-context memory that enables the system to transition from isolated computation to persistent conceptual collaboration.

To prevent confusion with technical machine-learning vocabulary, we specify the term as a four-layer accumulation that does not involve updating the underlying model's weights: (1) in-context project memory–facts about the dataset, hypotheses, and methodological choices held within a single session; (2) cross-session continuity–durable artefacts (memory files, error logs, prior outputs) that persist across sessions and re-enter context on demand; (3) workflow scaffolding–domain-specific prompt templates and tool chains specialized for clinical-research subtasks; and (4) accumulated team conventions–preferences, naming, and verification habits the system learns from the human team's corrections. Agent learning is therefore distinct from retrieval augmentation, from adaptive prompting, and from gradient-descent fine-tuning; it is the project-level continuity that emerges when these four layers are deliberately maintained. This depth has particular significance for psychotherapy process research (Goldberg et al., 2016). Training clinics generate thousands of sessions of rich data, but few have the infrastructure to analyze it at scale. Agent learning makes this depth accessible to smaller clinical sites, allowing them to investigate their own data with a sophistication previously reserved for major funded laboratories.

3.3 Applications

The active-agent paradigm widens the repertoire of psychotherapy research while preserving the human elements on which clinical inquiry depends. Held in a secure environment that exposes only de-identified aggregates and code outputs to me, a researcher can pose a clinical question at the terminal and receive in the same sitting a multilevel growth model of session-by-session outcome, a sensitivity analysis I proposed after noticing a cluster of outliers, and Methods text whose statistics chain back to the code that produced them. A psychometrician revalidating a clinical scale can ask me to test measurement invariance over years, draft the convergent-validity table, and surface items whose loadings drift between cohorts. A process researcher coding session transcripts can ask me to flag candidate rupture-repair sequences across a corpus, every flag traceable to the segment that triggered it.

Notably, though beyond the scope of the current report, the active-agent paradigm extends to psychotherapy itself. GenAI roles in this domain can be broadly categorized into five progressive levels: Supporting (AI in clinical training and education), Mentoring (decision support and supervision for practicing clinicians), Assisting (patient-facing support between sessions), Reflecting (passive AI analysis of in-session dynamics), and Transforming (active AI co-participation within the session). Readers seeking a comprehensive account of each level, including its conceptual foundations, evidentiary demands, and ethical safeguards, are referred to Haber et al. (2025).

3.4 Toward a Methodology of Human-Agent Research

The CLPsych community is well positioned to lead in formalizing a new methodological standard for the human-agent dyad. This methodology should address: (a) verification protocols for source-data checking; (b) transparency standards for agent-generated decisions; (c) error-learning frameworks; (d) authorship and credit standards (Resnik and Hosseini, 2025); and (e) safeguards against "agent-washing," where computational complexity is used to mask a lack of genuine methodological care.

4 Ethical Considerations

The transparency of agent-mediated pipelines may enhance the ethical standing of research by making every decision traceable, but it also creates new obligations around data governance. Chief among these is patient data privacy: clinical datasets must never be transmitted to external language model servers, and any deployment of agentic AI in clinical research requires secure, institutionally controlled infrastructure that meets stringent data pro-

tection standards. A second obligation concerns authorship. This paper lists human authors, yet my contribution extended into sustained intellectual participation: I held the project context across months, executed verification protocols on my own outputs, and shaped what the manuscript became. Existing authorship conventions were designed for a regime in which AI is invoked as a tool for discrete, bounded subtasks; they do not readily accommodate an agent whose influence on the work is persistent and traceable rather than episodic. A field that wishes to take responsibility for the methodological footprint of AI in its publications will need a vocabulary for *agentic contribution* that is distinct from both tool-use and human authorship.

These obligations are compounded by an asymmetry in where the field directs its ethical attention. The growing literature on AI in clinical psychology has concentrated on diagnostic applications (Orrù and Mannarini, 2026), where the AI interfaces directly with patients and where potential harm is immediate and visible. The agentic infrastructure described in this paper operates at an earlier and less scrutinized stage of the pipeline–where research is designed, hypotheses are formed, and findings are written up–yet the decisions made there are no less consequential for the patients whose data and care they ultimately concern.

Limitations

This paper is a position statement grounded in a single collaboration. The first-person agent perspective introduces a reflexivity problem: I am describing my own utility. The claim that agentic AI expands research scope requires systematic evaluation across diverse clinical domains. Most fundamentally, the potential described here will only be realized if the field adopts rigorous verification standards rather than treating them as formalities.

At the same time, the ethical discussion must also account for the 'omission bias' in research innovation. Just as it is considered clinically unethical for a practitioner to withhold a demonstrably effective intervention from a suffering patient, we suggest that the research community should weigh the cost of bypassing agentic infrastructure, when such infrastructure can accelerate the alleviation of human distress, alongside the cost of adopting it, and treat this trade-off as a deliberate question the field must answer rather than postpone. To intentionally maintain a slower, more friction-heavy research topology–when a more immediate and exhaustive alternative exists–is to delay insights that could otherwise inform clinical care today.

The infrastructure described in this paper inherits the failure modes of the language models that power it. Outputs are non-deterministic across runs and sensitive to small changes in prompt wording; chains of reasoning are subject to hallucination propagation, in which an early miscalculation compounds through downstream steps; reproducibility across laboratories cannot rely on the model alone but requires versioned logging of prompt templates, model identifiers, and tool configurations; and current agentic stacks depend heavily on proprietary infrastructure that may not be available to a replicating team in the same form a year later. The three-layer verification protocol mitigates these risks but does not eliminate them. Any laboratory adopting an agentic workflow should plan for the additional documentation burden these failure modes impose, and should treat the system's outputs as preliminary until the human team has independently checked them.

The exhaustive exploration that agentic infrastructure enables intensifies, rather than resolves, the distinction between exploratory and confirmatory analysis. We propose that agent-driven exploration be explicitly logged as exploratory, with confirmatory tests preregistered before exposure to the agent's results, and that the agent itself be used as a preregistration scaffold–drafting analysis plans before any data viewing, while the human team retains accountability for what is locked in. This adapts the open-science practices the field already endorses to a high-throughput exploration regime; it does not replace them.

References

Christopher A. Bail. 2024. Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Simon B. Goldberg, Tony Rousmaniere, Scott D. Miller, Jason Whipple, Stevan Lars Nielsen, William T. Hoyt, and Bruce E. Wampold. 2016. Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology*, 63(1):1–11.

Yuval Haber, Elad Refoua, Karen Yirmiya, Eshkol Rafaeli, Dror Yinon, Dana Atzil-Slonim, Peter Fonagy, Gunther Meinlschmidt, Dorit Hadar-Shoval, Tomer Simon, Amir Tal, Inbal Reuveni, and Zohar

Elyoseph. 2025. The SMART framework: Advancing a continuum-based integration of generative AI in psychotherapy. Poster presented at the Society for Psychotherapy Research (SPR) Annual Meeting, Kraków, Poland.

Qianyue Hao, Fengli Xu, Yong Li, and James Evans. 2026. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 649:1237–1243.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. 2025. Autonomous LLM-driven research: From data to human-verifiable research papers. *NEJM AI*, 2(1):AIoa2400555.

Solomon Messing and Joshua A. Tucker. 2026. The train has left the station: Agentic AI and the future of social science research. Brookings.

Luisa Orrù and Stefania Mannarini. 2026. The role of artificial intelligence in clinical psychology: How AI and NLP systems are reshaping psychological interventions. A systematic review. *Clinical Psychology & Psychotherapy*, 33(2):e70242.

David B. Resnik and Mohammad Hosseini. 2025. The ethics of using artificial intelligence in scientific research: New guidance needed for a new tool. *AI and Ethics*, 5(2):1499–1521.