

More Than "Means to an End": Supporting Reasoning with Transparently Designed AI Data Science Processes

Venkatesh Sivaraman *University of California, San Francisco*
Patrick Vossler *San Francisco*

Julian Hong *University of California, San Francisco*
Jean Feng *San Francisco*

Adam Perer *Carnegie Mellon University*

Main theme(s): *design strategy*

Target domain(s): *data science, healthcare*

Cognitive 'target(s)': *reasoning, critical thinking, problem formulation*

Type of contribution & main idea

We propose that AI data science agents should be designed around **transparent, steerable intermediate artifacts** rather than optimizing for the final answer regardless of process [4]. Drawing from two medical AI systems, HACHI [1] and Tempo [2], we show how well-designed intermediates can surface important analytical choices, empower non-experts to contribute their domain expertise, and potentially improve the quality of the final result.

The AI tool's key characteristics

Both HACHI and Tempo explicitly surface **intermediate artifacts** at critical decision points.

HACHI helps clinicians build predictive models by automatically identifying and extracting concepts from clinical notes, then presenting these concepts in an interactive interface where users can inspect, critique, and refine them throughout model training.

Tempo translates natural language queries into a readable, precise query language for extracting temporal event data from electronic health records, allowing domain experts to verify and edit data extraction procedures without requiring SQL expertise.

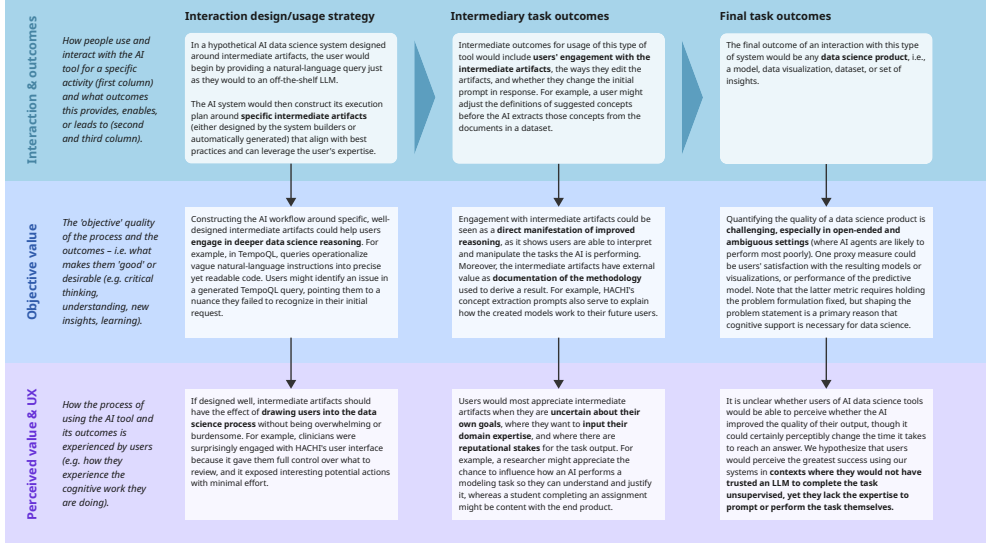
These tools share three design principles:

- They use concise representations that don't require technical expertise to interpret
- They reveal underlying analytical judgments that domain experts are uniquely positioned to evaluate
- They enable direct user control over key decisions.

Rather than generating hundreds of lines of opaque code, they produce **precise, steerable artifacts** like structured concept lists or domain-specific query languages that make the AI's reasoning transparent and modifiable.

What you would like to discuss

1. **When in an AI workflow** should we surface intermediate artifacts?
2. **How should we convey intermediate outputs** to best facilitate user understanding and control?
3. **How can we evaluate** whether intermediate artifacts effectively support reasoning?



What would you like to take away from the workshop?

Our research group is currently embarking on a project to **build a new agentic AI data science system** to help non-data scientist clinical experts create datasets and models on a complex health care dataset. We would like to ideate design characteristics that we can follow in our system so that it has the best chance of supporting our users. We would also like to learn about possible ways to validate this system's effectiveness.

Key references (e.g. of main theories, empirical evidence, measurement methods etc.)

- [1] Feng, J., Kothari, A., Vossler, P., Singh, C. (2026). Human-AI Co-design for Clinical Prediction Models (arXiv:2601.09072). arXiv: <https://doi.org/10.48550/arXiv.2601.09072>
- [2] Ma, Z., Boyce, R. D., Perez, A., & Sivaraman, V. (2025). TempoQL: A Readable, Precise, and Portable Query System for Electronic Health Record Data (arXiv:2511.09337). arXiv: <https://doi.org/10.48550/arXiv.2511.09337>
- [3] O'Brien, G. (2025). How Scientists Use Large Language Models to Program. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, 1-16. <https://doi.org/10.1145/3706598.3713668>
- [4] Shi, W., Xu, R., Zhuang, Y., Wang, M. D. (2024). EHAAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records (arXiv:2401.07126). arXiv: <https://doi.org/10.48550/arXiv.2401.07126>
- [5] Subramonyam, H., Pea, R., Pondoc, C., Agrawal, M., & Seifert, C. (2024). Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. CHI '24, 1-19.

How do people's goals 'interact with' the AI tool's goals? Is there a tension?

Users want to solve domain-specific problems (e.g., predicting traumatic brain injury) while maintaining accountability and understanding of their analytical choices. The latter goals are particularly important when the **problem statement is poorly defined or the validity of the results depends on the methods used to derive them**, such as in experimental research.

Meanwhile, the goal of AI agents by design is often to automate complex data science tasks end-to-end, which can create tension when automation obscures decisions that require domain expertise or value judgments.

Our approach resolves this by designing the AI's workflow around intermediate goals, each associated with interpretable artifacts, that align with users' needs for transparency and control rather than pursuing only the final answer.

How do you expect people to continue using this AI tool?

We envision iterative workflows where users and AI agents collaborate through multiple rounds of refinement. The artifacts they collaboratively produce would become part of the documentation and methodology, making analyses more reproducible and easier to communicate to stakeholders.

How to proceed with this work/idea?

We propose **developing design principles** for selecting which workflow stages should surface intermediate artifacts, balancing transparency needs with user time constraints. We also call for **new evaluation frameworks** that can assess AI workflow designs in high-stakes, open-ended problem contexts.