Qeios

Peer Review

Review of: "Mini Me: Personal AI That Was There Continuum, a Context Substrate for Measurable Presence and Safe Action"

Rajeew Vishvakarma[1]

1. Infosys (India), Bengaluru, India

Overall Assessment

This is an ambitious and thoughtful paper addressing an important emerging question: what infrastructure would be required for a personal AI to maintain meaningful continuity across time, devices, applications, and physical environments while preserving privacy and keeping consequential actions under appropriate control?

The proposed **Continuum** architecture is particularly interesting because it treats persistent personal AI as more than a memory problem. The manuscript brings together typed and provenance-aware memory, purpose-scoped disclosure, cumulative exposure accounting, cross-device trust, bounded autonomy, and reversible action. The concept of **presence**, expressed through continuity, comprehension, anticipation, and authority, also provides a useful way of thinking about personal AI beyond conventional context-window or retrieval-based approaches.

A further strength is that the authors are appropriately explicit that this is a **reference architecture and position paper rather than an implemented system**. The manuscript also demonstrates considerable awareness of privacy, prompt injection, bystander interests, compromised devices, and the risks of excessive autonomy.

I found the paper stimulating and potentially valuable. My comments below are intended mainly to help strengthen the transition from an extensive conceptual architecture toward a more testable research framework.

doi.org/10.32388/BWSVZZ

Major Comments:

1. A focused empirical demonstration would substantially strengthen the paper.

The manuscript provides an unusually detailed architecture, evaluation framework, and set of falsifiable hypotheses, but the principal mechanisms remain conceptual. For a position paper, this is understandable, although even a small prototype would significantly increase confidence in the proposal.

The authors do not necessarily need to implement the entire Continuum architecture. A focused experiment around one key mechanism could be sufficient—for example, comparing purpose-scoped context disclosure with full-context disclosure, evaluating typed memory against conventional RAG, or testing the proposed action controls under indirect prompt-injection scenarios.

The evaluation section already provides a strong foundation for such work, so this seems a natural next step rather than a change in the paper's direction.

2. The novelty claim could be positioned more precisely relative to recent persistent-memory and personal-context systems.

The paper already reviews a broad range of relevant work, which is a strength. However, because persistent agent memory and portable personal context are developing rapidly, it would help to state even more explicitly where Continuum's contribution differs from recent memory-centric approaches.

My reading is that the strongest contribution is not any single component in isolation, but the **integration of persistent memory, privacy-aware disclosure, cross-device context, action authority, and reversibility within one policy substrate**.

A short comparison table showing how Continuum differs from recent personal-memory, agent-memory, and context-portability approaches would make this contribution immediately clearer to readers.

3. Cumulative exposure accounting is promising and deserves somewhat stronger specification.

The idea that privacy risk should be assessed cumulatively rather than disclosure-by-disclosure is one of the most interesting aspects of the paper. Individually harmless disclosures may become sensitive when combined over time, and the proposed exposure ledger appropriately recognizes this.

However, some questions remain around how exposure would be quantified in practice: how sensitivity weights are assigned, how correlations between attributes are handled, and how changing external knowledge affects what can be inferred.

The authors already acknowledge that the exposure function requires empirical calibration. I would encourage them to add a small set of formal properties or design requirements that any implementation of the exposure function should satisfy. This would make the concept easier for future researchers to operationalize and evaluate.

4. The security discussion could distinguish more explicitly between authorization and decision integrity.

The manuscript makes a strong design choice in preventing environmental content—emails, webpages, speech, documents, QR codes, etc.—from directly creating authority. This is an important safeguard against prompt injection.

A remaining issue is that malicious content may not need to obtain additional authority. It might instead manipulate the AI into selecting an incorrect action that already falls within an existing authorization envelope.

It may therefore be useful to distinguish between:

- **authorization safety** — whether the system is permitted to perform an action; and
- **decision integrity** — whether the evidence and reasoning used to select that action are trustworthy.

This distinction could strengthen the Four Gates model and would fit naturally with the paper's existing use of provenance.

Relatedly, persistent memory deserves continued attention as a security surface. A false or malicious observation that is promoted into long-term memory could influence multiple future decisions. The manuscript's typed memory, provenance, expiry, and correction mechanisms already provide a good foundation; explicitly incorporating memory-poisoning tests into the proposed evaluation would make the security argument even stronger.

Minor Comments:

The manuscript is comprehensive, but its length may make the central contribution harder to identify. Some repetition around bounded capture, reversibility, remote disclosure, and personal control could be reduced without losing substance.

I would retain the detailed evaluation framework, as it is one of the paper's strongest sections. The scenarios are also useful, particularly because they demonstrate that the architecture is intended to support both routine assistance and higher-risk situations.

The discussion of bystander privacy is thoughtful. In future versions, the authors may wish to develop this further into a more explicit multi-party policy model, since persistent personal AI operating in shared physical environments inevitably involves people other than the primary user.

Recommendation:

Revision recommended.

This is a well-conceived and forward-looking systems proposal. Its principal value lies in treating persistent personal AI as a combined problem of **memory, privacy, provenance, authority, and safe action**, rather than addressing these dimensions independently.

The manuscript would be strengthened most by a focused empirical demonstration, sharper comparison with recent persistent-memory architectures, modest formalization of cumulative exposure, and further development of decision-integrity and persistent-memory security.

These comments do not require changing the central thesis of the paper. Rather, they would help make an already interesting architecture more testable, easier to compare with related work, and more useful as a foundation for subsequent research.

Declarations

Potential competing interests: No potential competing interests to declare.