



GuardianAgent: A Trustworthy Personal Agentic AI Platform



Ruiyi Yang¹, Rahat Masood¹, Flora D. Salim¹, Salil Kanhere¹, Michael Davis²,
Sacha Molitorisz², Elise van den Hoven²
¹UNSW Sydney, ²UTS Sydney

Background

Problem:

- ❑ Continuous data collection by apps, websites, and devices, often without genuine awareness or consent.
- ❑ Dark patterns and excessive data collection erode user trust and meaningful consent.
- ❑ Existing products rely on rigid rules, lack transparency, and provide limited real-time user guidance.

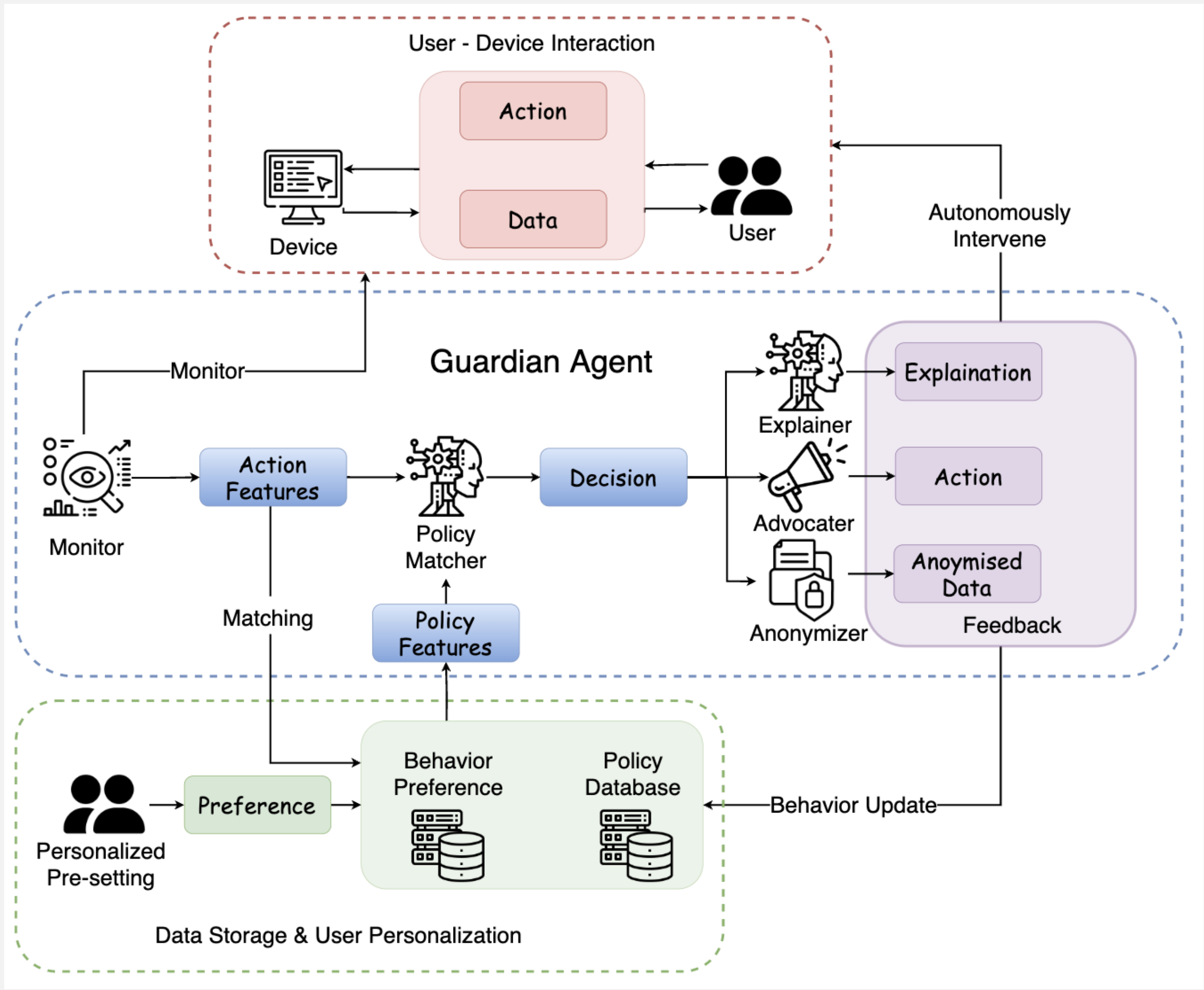
What We Need:

A user-centric, privacy-preserving agentic system that detects risk before submission, minimises data, and explains actions clearly.

Challenges:

- ❑ **C1:** Context-aware risk detection & accurate policy matching.
- ❑ **C2:** Data minimisation & de-identification without breaking the task.
- ❑ **C3:** Real-time, on-device enforcement with transparent user control.

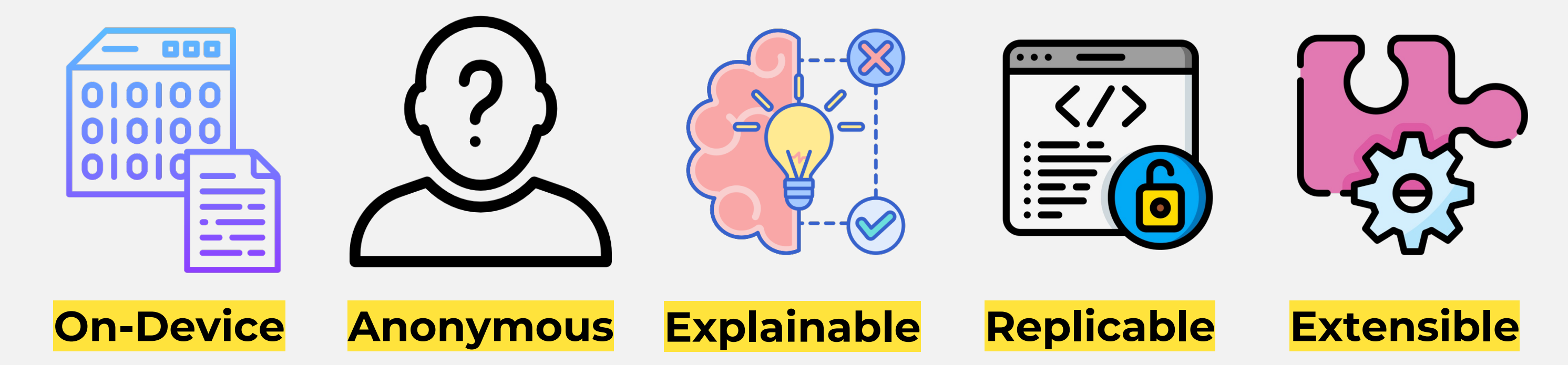
Framework



- ❑ **Monitor** (Monitoring user-device interactions) – browser extension, software, apps
- ❑ **Policy Matcher** (Connect to policy related database, Maps signals to user policy & legal norms; decides allow/deny/transform) – LLM + RAG usage
- ❑ **Advocater** (Execute allow/deny/transform)
- ❑ **Anonymizer** (Anonymize users' information)
- ❑ **Explainer** (Explain decisions to users) – LLM usage

Able to **autonomously intervene** user's interaction, while iteratively **update background knowledge**.

Contributions & Features



Milestones

- ❑ **Q3 2025:** Competitive Analysis, Framework Design
- ❑ **Q4 2025:** Prototype Development + baseline metrics.
- ❑ **Q1 2026:** Submission of Application for User Studies to the Ethics Committee
- ❑ **Q2 2026:** Prototype Development Completion and in-lab Testing.
- ❑ **Q3 2026:** User Studies and Analysis; Paper writing

Deliverables

Software Artifacts:

- ❑ Monitoring SDKs (tools-only)
- ❑ Policy Agent (local RAG) package
- ❑ Advocate Module (enforcement API)
- ❑ Anonymizer Module
- ❑ Explainer Agent (real-time UI kit)

Knowledge & Data

- ❑ **Policy KB:** Normalised policy/purpose/permission spans for common apps/sites; provenance + versioning.
- ❑ **Evaluation Sets:** Labeled cases for policy matching, risk detection, and field-level minimisation (pre/post payloads).

Research Paper

A paper on the framework, user studies and analysis and submit to a CORE A/A* conference

Acknowledgment:

This work is supported by the UNSW-UTS Trustworthy Digital Society initiative and the project “GuardianAgent” team: Salil Kanhere, Flora Salim, and Rahat Masood from UNSW Sydney, and Elise van den Hoven, Michael Davis, and Sacha Molitorisz from UTS Sydney.