ScrollSense: AI Risk Analysis and Mitigation Plan

Sri Harsha P.
Divya Sri Bandaru

Part 1: Risk Identification and Analysis

Risk 1: Malicious Use – Targeted Social Engineering

Risk Category: Malicious Use & Digital Security
Description: ScrollSense collects detailed behavioral data (goals, scrolling habits, vulnerable time periods, message-response patterns). If the system or API connection were compromised, attackers could use this data to craft hyper-personalized phishing messages disguised as legitimate productivity nudges. Because users trust these prompts, malicious actors could exploit moments of low willpower to deliver harmful links or credential-harvesting attempts.
Worst-Case Scenario: A compromised extension sends highly personalized phishing nudges at the exact times users are most distracted, enabling scaled social engineering attacks across all users simultaneously.
Affected Populations: Graduate students, young professionals, and international students with sensitive academic or employment data; users experiencing stress or mental health struggles who are more vulnerable to persuasive prompts.

Risk 2: Privacy – Exposure of Behavioral Vulnerability Profiles

Risk Category: Privacy & Digital Security
Description: Although ScrollSense stores data locally, the extension still holds highly sensitive information daily routines, mental health-related goals, scrolling vulnerability windows, tone/blur preferences, and persuasion responses. Malware, malicious extensions, sync compromises, or data-poisoned extension updates could extract this data.
Worst-Case Scenario: A breach leaks complete “persuasion profiles” that reveal psychological weaknesses, daily behavior patterns, and personal goals. This data could be sold, used for targeted scams, or combined with other datasets to enable harassment or surveillance.
Affected Populations: Users in high-risk roles (researchers, tech workers, politically sensitive individuals), people with mental health disclosures, marginalized users at higher risk of targeted harassment, and individuals in restrictive political environments.

Risk 3: Fairness and Bias – Reinforcing Productivity Culture & Ableism

Risk Category: Fairness & Bias
Description: ScrollSense assumes that reducing screen time is universally beneficial, which can marginalize users with different cognitive needs, cultural norms, or legitimate reasons for extended social media use. Standardized nudges and blur effects may shame users with ADHD,

chronic illness, mental health conditions, or those relying on social media for community support.
Worst-Case Scenario: The tool functions “as designed” but harms vulnerable groups by intensifying guilt cycles, pathologizing healthy usage, or discouraging critical coping mechanisms reflecting biased assumptions rather than user wellbeing.
Affected Populations: Neurodivergent users, disabled users, people with chronic illness, international students or users in collectivist cultures, LGBTQ+ users relying on online communities, and low-income users using social media for work or essential communication.

Risk 4: Security – Data Poisoning & Adversarial Attacks on AI Features

Risk Category: Security & Resilience
Description: ScrollSense’s adaptive timing and Claude-generated nudges create multiple attack surfaces. Attackers could poison early behavioral data to create harmful defaults, manipulate aggregated data if used, or perform prompt injection to force Claude to generate unsafe messages.
Worst-Case Scenario: Poisoned behavior models set harmful scrolling norms while adversarial prompts cause the AI to produce malicious or misleading nudges, tricking users into unsafe actions under the guise of legitimate advice.
Affected Populations: Users heavily dependent on ScrollSense for self-regulation, users with low AI literacy, early adopters whose data influences defaults, and organizations deploying the tool at scale.

Risk 5: Reliability – AI Hallucinations & Cultural Misinterpretation

Risk Category: Reliability & Safety
Description: Claude may hallucinate deadlines, misinterpret user goals, or generate culturally inappropriate messages, especially for non-native English speakers or users with indirect communication styles. Hallucinated urgency or judgmental tones could trigger stress or confusion.
Worst-Case Scenario: The AI produces false or harmful nudges during high-pressure moments e.g., inventing deadlines, misinterpreting mental-health-related goals, or delivering linguistically garbled prompts leading to panic, demotivation, or unsafe behavior.
Affected Populations: Users with anxiety, OCD, impostor syndrome, international students, multilingual users, people from high-context cultures, and anyone with low AI literacy who may assume hallucinated messages are accurate.

Part 2: Mitigation Plan

Risk 1 Malicious Use (Social Engineering)

Design

- Mark all AI messages with an “AI-generated” badge and distinct UI.
- Make AI nudges text-only; only two hard-coded actions allowed: **Done for now** / **Continue 10 min**.
- Session intent is client-side and sanitized into templates (e.g., “work on [category]”) before any API call.
- Optional “view raw prompt/response” mode for power users.

Development

- Sign/validate API requests and validate responses against strict JSON schemas.
- Enforce a strict Content Security Policy; disable external scripts and inline JS.
- Rate-limit + anomaly detection (e.g., >50 requests in 5 min -> disable AI).
- Output sanitization: length limit, URL/keyword blocklist, urgency flags; failing responses replaced with a safe fallback.
- Certificate pinning + cached known-good message hashes.

Deployment

- Public security docs, responsible disclosure, bug-bounty program.
- “Report suspicious message” button that logs full context for triage.
- Remote kill-switch to disable AI globally and fallback to pre-written messages.
- Security checklist before each store update.

Risk 2 Privacy (Exposure of Vulnerability Profiles)

Design

- Collect minimal data: only abstracted goal categories and aggregated metrics.
- Ephemeral storage: granular session logs auto-delete (7–30 days); long retention only by opt-in.
- “Minimal data” privacy mode (no personalization, weekly auto-deletes).
- Settings UI showing what’s collected, why, retention, and a one-click delete.

Development

- Client-side encryption of stored data; keys derived per device (Web Crypto API).
- Use chrome.storage.local with extension isolation; disable Chrome sync for sensitive data.
- Enforce auto-expiration rules and integrity checks (SHA-256/HMAC); clear and warn on tampering.
- Sanitize/summarize data before any API call (no PII, send only category+platform+elapsed).
- Signed updates to prevent supply-chain injections.

Deployment

- Plain-language privacy policy, annual third-party audits, breach detection/response plan.
- Granular consent during onboarding (checkboxes for goals, tracking, AI).
- “Panic button” to delete all data immediately.
- Quarterly transparency reports.

Risk 3 Fairness & Bias (Productivity Culture / Ableism)

Design

- Reframe value as “intentionality,” not productivity. Provide diverse session intents (work, connect, mental-health break, browse).
- “Protected sessions” and “intervention pause” to respect user autonomy.
- Multiple tone options (supportive, gentle, neutral, direct); cumulative tracking opt-in only.
- Culturally diverse templates and defaults.

Development

- Context-aware settings per platform (user labels platform uses).
- Adaptive learning that honors overrides ask users before changing defaults; allow excluding outliers.
- Flexible escalation profiles (gentle → assertive) selectable by user.

Deployment

- Co-design and compensate for disability/neurodiversity communities in beta.
- Explicit limitations docs: not a clinical tool; may not suit everyone.
- “This doesn’t work for me” feedback button; regular review and quarterly debiasing updates.
- Ethical A/B testing with opt-out and safeguards for vulnerable groups.

Risk 4 Security (Data Poisoning / Adversarial Attacks)

Design

- Require 10+ sessions for baseline learning; show users the data used to generate defaults and allow exclusions.
- Limit collaborative filtering to verified cohorts and apply differential privacy.

Development

- Anomaly detection for training data and provenance logging (which sessions trained what).
- Prompt-injection defenses: strict input sanitization, template enforcement, keyword blocklists.
- Adversarial output checks (semantic similarity, pattern matching); strict API response schema + confidence thresholds.
- Rate limiting and verified opt-in before contributing aggregated data.

Deployment

- Real-time security dashboard, incident response playbook, quarterly tabletop exercises.
- Bug bounty and red-team adversarial testing before public launch.
- Publish post-mortems and advisories when incidents occur.

Risk 5 Reliability (Hallucinations & Cultural/Language Mistakes)

Design

- Constrain prompts: explicit “do not invent” instructions and strict length/structure requirements.
- Maintain a large human-written fallback library; default new users to pre-written messages for week 1.
- Make AI use opt-in with clear trade-offs explained.

Development

- Multi-layer output validation: length, keyword/time/person detection, sentiment thresholds, embedding similarity to the user’s goal, and a confidence cutoff. Fallback if checks fail.

- Language detection: do not send non-English input to the API without consent/translation; prioritize pre-written localized messages.
- Daily AI message caps to reduce exposure and cost.

Deployment

- Onboarding warnings about AI limits; “Report this message” with 24-hour triage.
- Human QA review for early deployment (sampled message audits), conservative staged rollout, and rapid rollback kill-switch for spikes in problematic outputs.
- Staged model upgrades only after safety benchmarking and native-speaker validation for any new language support.
- User education materials on hallucinations and how to respond.

Part 3: Ethical Reflection

Conducting this risk analysis fundamentally changed how I view ScrollSense. I originally saw it as a simple productivity tool and focused mainly on technical tasks, smooth blur effects, UI polish, and good prompts. But applying the sociotechnical systems paper, the malicious AI paper, and the NIST risk management framework made me realize ScrollSense is not just a technical system; it’s a sociotechnical one shaped by cultural assumptions and power dynamics. (2) What I once treated as an objective goal of “healthy” or “productive” scrolling is actually deeply contextual. The risk analysis made me see how the tool might unintentionally reinforce Western productivity norms, ableist assumptions, or neurotypical patterns, and how my design choices could harm users who don’t fit into those categories. The Framing Trap and Formalism Trap highlighted how easily I could turn complex human behavior into oversimplified metrics that don’t reflect lived reality.(1)

I also realized that ScrollSense’s architecture/personalization, behavioral tracking, and AI-generated nudges have dual-use risk built into it. The same features meant to support users could, if compromised, be repurposed for manipulation or targeted harm, as illustrated in the malicious AI paper. The NIST framework helped me understand that trustworthiness, privacy, and safety are not add-ons but core requirements. (2) This exercise showed me that thinking about risks early is essential, because once harmful assumptions are built into an AI system and deployed, they scale quickly and are extremely difficult to undo. Mistakes in judgment or design can create real psychological, cultural, or security harms especially for marginalized users.

The biggest ethical challenge for ScrollSense is balancing supportive intervention with respect for autonomy. There’s a thin line between helping users honor their intentions and creating a surveillance-like system that judges their behavior. This reflection pushed me toward designing with more humility offering flexibility, transparency, and user control, and acknowledging that ScrollSense will not be appropriate for everyone. Overall, this exercise taught me that responsible AI development is less about technical capability and more about understanding context, listening to affected communities, and being willing to slow down or reconsider design choices to prevent harm.

References:

1. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *ACM Digital Library Home*, 59–68. https://doi.org/10.1145/3287560.3287598
2. IBM Technology. (2025, August 9). *Mastering AI Risk: NIST's risk Management Framework explained* [Video]. YouTube. https://www.youtube.com/watch?v=0oeD2Wf25wY