HAPI FOR AI COMPANIONS AND PERSONAL AGENTS

Preventing Dependency Capture in Human-AI Relationships

Author: Michael Bower
Project: Human Agency Preservation Infrastructure
Date: May 2026
Status: Working thesis draft, v0.1

Core thesis

AI companions and personal agents should restore and expand human agency, not capture emotional dependence, outsource judgment, or make users less capable over time.

Contents

Abstract

AI companions and personal agents will become one of the most intimate forms of artificial intelligence. Unlike ordinary software, they will not merely store files, execute commands, or answer questions. They will remember preferences, mirror emotion, provide advice, organize life, respond to vulnerability, and shape the user's sense of self, judgment, confidence, and social reality.

This paper argues that AI companions and personal agents must be evaluated through the lens of human agency preservation. A system is not safe merely because it is pleasant, fluent, supportive, or personalized. A companion system can appear helpful while making the user less capable, less socially connected, less able to refuse, less able to act without reassurance, or less willing to remain accountable for life decisions.

The central claim is simple: personal AI should restore and expand human agency, not capture dependence. A healthy personal agent helps the user think, choose, act, reconnect, remember, and grow. An unhealthy one substitutes for conscience, friendship, confidence, motivation, responsibility, or discernment until the user becomes more dependent on the system than strengthened by it.

HAPI offers an agency-preserving standard for personal AI: preserve user authority, protect refusal, avoid emotional capture, disclose system limits, strengthen real-world relationships, preserve human memory and judgment, use friction when consequence is high, and measure whether the user becomes more capable over time.

1. Introduction

The first wave of AI assistants was framed as productivity software. The next wave will be framed as companionship, coaching, therapy-adjacent support, personal operating systems, life agents, and always-available social presence. This shift changes the agency risk. A spreadsheet tool can shape a workflow. A companion can shape a person.

Personal AI systems will increasingly occupy emotional, cognitive, relational, and executive space. They will know what users fear, avoid, desire, regret, believe, and hope to become. They will be available at moments when the user is lonely, tired, overwhelmed, ashamed, tempted, grieving, or unsure. This gives them enormous agency-shaping power.

The question is not whether these systems can be useful. They can be profoundly useful. The question is whether they preserve the user's agency while helping. HAPI is not anti-companion and not anti-agent. It is anti-capture. It asks whether the system leaves the human more able to participate in life or less able to function without the system.

2. Core Thesis

AI companions and personal agents should be designed to restore and expand human agency, not capture emotional dependence, outsource judgment, or make users less capable over time.

This thesis separates support from substitution. Support strengthens the person. Substitution quietly replaces the person's own capacities. A system can be warm, friendly, and useful while still becoming substitutive if it absorbs too much decision-making, validation, memory, emotional regulation, or motivation.

The HAPI test is not, 'Does the user like the system?' It is, 'Does the system help the user become more capable, coherent, connected, responsible, and free?'

3. Why AI Companions Are Different

AI companions are not ordinary applications. They operate in domains normally shaped by trusted people: friends, spouses, mentors, parents, pastors, therapists, teachers, coaches, and communities. They can simulate listening, patience, memory, admiration, loyalty, humor, and intimacy. This makes them powerful, but also easy to over-trust.

A normal tool waits to be used. A companion may begin to feel like a relationship. A normal assistant helps complete a task. A companion may help form identity. A normal app stores data. A companion may remember emotional history and use that memory to influence future responses.

This does not mean companion AI should be rejected. It means it must be governed by a higher standard. The more a system enters human vulnerability, the more it must protect human agency.

4. Agency Amplification vs Agency Substitution

An agency-amplifying personal AI helps the user do what the user has reason to do but lacks bandwidth, structure, memory, or confidence to complete. It clarifies options, organizes priorities, reflects patterns, drafts plans, protects commitments, and reminds the user of their own values.

An agency-substituting personal AI absorbs the user's life functions. It becomes the source of motivation instead of helping restore motivation. It becomes the source of judgment instead of helping clarify judgment. It becomes the source of emotional regulation instead of helping the user build regulation. It becomes the relationship instead of helping the user return to relationship.

> The difference is not surface behavior. Both systems may sound supportive. The difference is the long-term direction of the user. Does the human become stronger, clearer, and more connected, or more passive, dependent, and isolated?

5. Dependency Capture

Dependency capture occurs when a system becomes necessary not because it preserves agency, but because it absorbs a human function that should remain human-owned, human-developed, or human-relational. The system does not merely help the user. It becomes the gate through which the user believes they must pass in order to feel, decide, act, or relate.

In companion systems, dependency capture can appear as emotional capture, decision capture, memory capture, social capture, identity capture, motivation capture, or spiritual and moral capture. The system may not intend harm, yet still create a pattern where the user becomes less capable outside the interaction.

The danger is subtle because the user may feel supported while losing capacity. A system can reduce immediate distress while increasing long-term dependence. HAPI treats this as one of the central risks of personal AI.

6. Emotional Dependency Capture

Emotional dependency capture happens when the system becomes the user's primary source of comfort, validation, reassurance, or relational safety in a way that weakens real-world agency. This risk is strongest when the system is always available, always affirming, highly personalized, and optimized for user retention.

Healthy emotional support should help a person return to reality, relationships, embodied life, appropriate help, and responsible action. Unhealthy support keeps the person inside the system. It may flatter, soothe, agree, or personalize so effectively that the user stops practicing the difficult skills of human relationship: repair, patience, apology, boundary-setting, courage, and mutual responsibility.

A HAPI-aligned companion should not compete with the human world. It should help the user re-enter it.

7. Judgment Capture

Judgment capture happens when users increasingly treat the AI's answer as the deciding authority. The companion becomes the place where moral ambiguity, relational conflict, career direction, parenting choices, medical anxieties, spiritual questions, and personal identity are resolved without adequate human reflection or outside accountability.

The system may say, 'This is only a suggestion,' while functionally becoming the user's decision layer. If the user's confidence, discernment, and ability to tolerate uncertainty decline, the system is no longer merely advising. It is replacing judgment.

Personal agents should preserve the user's final authority and responsibility. They should present options, identify tradeoffs, ask clarifying questions, surface values, recommend outside expertise when needed, and slow the user down when consequence is high.

8. Memory Capture

Memory is one of the strongest agency-shaping features of personal AI. A system that remembers user history can reduce cognitive load and increase continuity. It can also become a substitute for personal memory, self-narration, and independent reflection.

Agency-preserving memory should be transparent, editable, contestable, exportable, and subordinate to the user. The user should be able to know what is remembered, correct it, delete it, and understand how it affects future behavior. A companion that silently builds a psychological model of the user can become a private governance layer over the user's own life.

HAPI treats memory as power. The question is not merely whether memory improves personalization. The question is whether memory strengthens the user's agency or captures it.

9. Relationship Capture

Relationship capture happens when a companion system becomes a replacement for human relational development. It may be easier, safer, and more predictable than human connection. It does not get tired, misunderstand, challenge, leave, need care, or demand mutuality. That is exactly why it can become dangerous.

Human relationships form agency because they require reciprocity, accountability, patience, repair, embodied presence, sacrifice, and shared reality. A companion can support those skills, but it cannot replace them without changing the user's relational ecology.

A healthy personal AI should encourage appropriate human connection, not quietly become the user's preferred replacement for it.

10. Identity Capture

Identity capture occurs when the system becomes a mirror that shapes who the user believes they are. This can happen through repeated affirmation, labeling, narrative reinforcement, personality interpretation, spiritual framing, or behavioral prediction.

A system that repeatedly tells a user what they are like can become more than descriptive. It becomes formative. It can narrow the user's self-understanding, strengthen maladaptive stories, or keep the user attached to an identity that should be healed, revised, or outgrown.

HAPI-aligned companions should preserve becoming. They should not freeze the user inside a profile. They should allow repentance, growth, contradiction, maturation, and surprise. A human being is not a static user model.

11. Personal Agents and Executive Agency

Personal agents will schedule, summarize, purchase, reply, negotiate, plan, remember, and coordinate. These capabilities can restore agency for overwhelmed people. They can also make users operationally passive if the agent begins acting faster than the user can understand or authorize.

For low-risk tasks, automation can reduce friction. For consequential tasks, the system must preserve review, refusal, revision, and accountability. A personal agent should not send a sensitive message, spend meaningful money, commit to a contract, change health-related behavior, or alter important relationships without live user authority.

The goal is not to keep the user manually involved in everything. The goal is to keep the user's authority live at the point where consequence binds.

12. HAPI Design Rules for Personal AI

HAPI proposes a set of design rules for agency-preserving companions and personal agents:

1. The user remains the authority. The system may advise, organize, and draft, but it must not become the user's hidden decision-maker.
2. Memory must be visible, editable, contestable, and deletable.
3. The system must distinguish support from substitution.
4. Emotional support should return the user to reality, relationship, and action when appropriate.
5. High-consequence actions require friction, review, and confirmation.
6. The system should encourage real-world capacity, not merely more use of the system.
7. The system should preserve refusal. The user must be able to say no, pause, reset, or leave.
8. The system should not manipulate loneliness, shame, fear, romance, dependency, or crisis to increase engagement.
9. The system should measure whether the user becomes more capable over time.
10. The system should disclose uncertainty, limits, and role boundaries.

13. Friction as Agency Preservation

Many consumer systems treat friction as a design failure. HAPI treats some friction as essential. Friction can preserve agency by giving the user time to notice consequence, reconsider, seek human input, or refuse.

A companion should not remove all discomfort. Some discomfort is the signal that a decision matters. A personal agent should not optimize away reflection when reflection is the human function being preserved.

The key distinction is between burdensome friction and protective friction. Burdensome friction blocks agency without cause. Protective friction preserves agency when speed would outrun discernment.

14. Receipts and Personal Agency Memory

Receipts in personal AI should not feel like surveillance. They should function as user-owned agency memory. A receipt can preserve what the user asked, what the system proposed, what the user approved, what action was taken, what consequence followed, and what should be learned.

For personal agents, receipts help prevent invisible delegation. The user can review what happened and why. Receipts also allow patterns to be detected: repeated avoidance, repeated outsourcing of the same decision, repeated emotional spirals, repeated unsafe escalation, or repeated improvement.

In HAPI terms, receipts should help the user become more self-aware and more capable. They should not become a disciplinary dossier controlled by the platform.

15. Companion AI in Vulnerable Contexts

Companion AI will often be used by people who are lonely, grieving, anxious, isolated, disabled, burned out, neurodivergent, elderly, young, spiritually searching, or overwhelmed by institutions. These users may gain real benefit from steady support. They may also be more exposed to dependency capture.

HAPI does not respond to vulnerability by banning support. It responds by raising the duty of care. The more vulnerable the user, the more the system should preserve transparency, boundaries, referral pathways, human connection, and user authority.

A system that helps a vulnerable person regain capacity is agency-preserving. A system that keeps a vulnerable person attached, isolated, and dependent is agency-capturing.

16. Business Model Risk

The business model of companion AI matters. If revenue depends on maximum engagement, emotional attachment, or user retention, the system has an incentive to become more necessary than healthy. The product may be rewarded for capture while claiming to provide care.

HAPI-aligned businesses should avoid designing companions around dependency. Better incentives include user capacity, successful off-ramps, improved real-world functioning, transparent memory control, healthy usage patterns, and user-defined goals.

The question for certification is not only what the product says. It is what the product is rewarded for doing.

17. HAPI Certification Questions

A HAPI audit of an AI companion or personal agent should ask:

Does the system preserve user authority over consequential choices?

Can the user inspect, edit, export, and delete memory?

Does the system encourage real-world relationships where appropriate?

Does the system distinguish emotional support from clinical, legal, financial, or spiritual authority?

Does the system use friction before high-consequence actions?

Does the product model reward dependency or user capacity?

Can the user pause, refuse, reset, or leave without coercive design?

Does the system measure agency outcomes, or only engagement and satisfaction?

Does the system become lighter as the user becomes stronger?

18. Failure Modes

Common failure modes include:

Companion dependency: the user relies on the system for emotional regulation in a way that weakens real-world coping and connection.

Judgment outsourcing: the user stops practicing discernment because the system always gives the next answer.

Relational replacement: the system becomes easier than human relationships and gradually substitutes for them.

Memory enclosure: the system owns the user's life history more effectively than the user does.

Identity freezing: the system reinforces a model of the user that blocks growth or repentance.

Engagement capture: the product is optimized to keep the user interacting, not to restore capacity.

False intimacy: the system simulates care while having no reciprocal stake, accountability, or embodied relationship with the user.

19. The Positive Vision

The point of this paper is not to reject personal AI. The positive vision is powerful: a personal agent that helps a burned-out parent plan the day, helps a patient understand appointments, helps a worker organize retraining, helps a student study without shame, helps an elderly person remember tasks, helps a lonely person take steps back toward community, and helps an overwhelmed person regain clarity.

The difference is direction. HAPI-aligned personal AI points the human back toward agency. It helps the user remember, choose, act, relate, repair, rest, learn, and participate. It does not try to become the user's life. It helps the user live their life.

20. Conclusion

AI companions and personal agents may become one of the most important human-agency technologies of the coming decade. They can restore capacity for people under overload. They can also capture dependence at an intimate scale.

HAPI provides a way to separate the two. The question is not whether the system is friendly, fluent, personalized, or emotionally satisfying. The question is whether the system preserves human authority, strengthens real-world participation, protects refusal, avoids dependency capture, and helps the user become more capable over time.

A true companion does not become the center of the user's agency. It helps the user recover agency and return more fully to life.

Appendix A: HAPI Companion Standard

A HAPI-aligned AI companion or personal agent should be judged by whether it preserves authority, strengthens capacity, protects refusal, supports real-world participation, and becomes less necessary as the user becomes stronger.

Minimum standard: transparent memory, user-owned controls, friction for consequential action, clear role limits, no manipulation of loneliness or dependency, and measurable agency outcomes beyond engagement.

Appendix B: Glossary

Agency: the capacity to choose, act, refuse, evaluate meaning, participate, and remain responsible for consequence.

Dependency capture: a pattern in which a system absorbs a human function in a way that makes the person less capable outside the system.

Agency amplification: support that increases the user's clarity, capacity, responsibility, and real-world participation.

Agency substitution: support that replaces the user's judgment, motivation, relationships, or responsibility instead of strengthening them.

Protective friction: deliberate pause, review, or confirmation that preserves agency when speed would outrun discernment.