Reflective Agency: Ethical and Empirical Framework for AI-Mediated Self-Reflection Systems

Minsol Kim[1], Wendy Wang[2], Jennifer Long[2], Rosalind Picard[1], Nathan Barczi[1], Pattie Maes[1]

[1]MIT Media Lab, Cambridge, MA, USA
[2]Wellesley College, Cambridge, MA, USA
minsol@mit.edu

Abstract

As artificial intelligence (AI) increasingly mediates self-reflective practices—from therapeutic conversational agents to personal informatics systems—it becomes crucial to examine how these technologies shape, rather than merely support, our self-understanding. This paper investigates the influence of AI-mediated journaling systems on *Reflective Agency*—the capacity to interpret and make meaning of one's experiences autonomously. Drawing on phenomenology and Aristotelian virtue ethics, we propose the **Reflective Agency Framework (RAF)**: a normative model with five design principles for such AI tools—*Internal Origination, Calibrated Responsiveness, Reflective Ambiguity, Transparency of Mediation, and Self-Continuity and Ethical Flourishing*.

We demonstrate RAF's relevance through a two-part empirical case study: first, a systematic feature-tension analysis of six widely used AI journaling applications; second, an exploration of user perceptions and responses to these tensions and principles. Mapping these insights back to RAF, we identify persistent conflicts between automation and autonomy, and offer design considerations that preserve the interpretive space essential for self-development. Our findings reveal that over-automation in AI-mediated reflection can erode agency. We urge designers and researchers to adopt and extend our framework to ensure future technologies genuinely support autonomy and meaningful self-discovery in digital well-being. Ultimately, we envision AI not as a guide, but as a quiet companion—helping individuals stay connected to their evolving sense of self and remain the primary agents in their self-discovery, thereby preserving what we define as *Reflective Agency*.

Introduction

As artificial intelligence (AI) becomes increasingly embedded in our daily lives, it is beginning to reach some of our most private and introspective spaces—ranging from therapeutic dialogue and personal journaling to emotional introspection (Jung et al. 2024; Kim et al. 2024a; Kocielnik et al. 2018; Nepal, Liu, and Lee 2024). But what happens when we start outsourcing our *inner voice* to these intermediaries? Can we remain as the primary agents of our self-discovery, or are we gradually relinquishing that role to machines?

A growing ecosystem of journaling applications and personal informatics (Li, Dey, and Forlizzi 2010) interfaces target mental health. These platforms now routinely gain insights about users from behavioral sensing—tracking activity and sleep patterns via wearable and mobile sensors (Whitmore et al. 2024; Nepal et al. 2024) and sentiment analysis of free-text entries (Kim et al. 2024a; Shickel et al. 2020). While such quantification can reveal latent patterns in mood and behavior, it also risks reifying complex affective experiences into reductive, system-readable formats (Cooney et al. 2018; Lupton 2016). In addition, prematurely generated summaries may truncate the narrative ambiguity requisite for rich self-interpretation (Di Lodovico, Houben, and Colombo 2025a; Reed et al. 2024). Moreover, overly assertive system interventions can undermine user intent by imposing authoritative perspectives. Such interventions may lead to nudge fatigue or erode trust in digital tools by dominating the interaction (Hussain 2025), and risk constraining users' interpretive space (Gaver, Beaver, and Benford 2003; Baumer 2015).

Nonetheless, there is growing evidence that thoughtfully designed AI tools can scaffold and deepen reflective practice rather than supplant it. For example, mixed-initiative personal informatics dashboards demonstrate how adaptive support helps users notice patterns without prescribing interpretations, fostering sustained engagement and deeper insight (Kocielnik et al. 2018; Song et al. 2025). Emotionally intelligent Conversational Agents (CAs) have demonstrated improvements in user engagement and well-being, particularly for individuals who struggle to journal consistently (Ghandeharioun et al. 2019; Zhou et al. 2020). Recent works have also tried to provided socially-situated XAI (Ehsan et al. 2021) Recent *Frictional AI* work also shows how intentional design can foster more reflection or engagement by introducing productive friction, rather than prioritizing frictionless automation (Natali 2024; Natali, Frischmann, and Cabitza 2024).

Yet despite the growing prevalence of AI-driven self-discovery tools, their impact on user agency during reflective activities remains largely unexplored. These tools reveal persistent tensions between automation and autonomy, frictionless usability and constructive ambiguity, that shape how reflection is mediated. In response, we propose the **Reflective Agency Framework (RAF)**: a set of five design

principles—*Internal Origination, Calibrated Responsiveness, Reflective Ambiguity, Transparency of Mediation,* and *Self-Continuity and Ethical Flourishing*. Grounded in phenomenology and Aristotelian virtue ethics, RAF attempts to address this important design challenge: how to design AI systems that enhance, rather than replace, the deeply personal work of self-reflection?

We make three contributions: (1) introduce RAF, a five-principle, philosophically grounded model towards designing AI systems that preserve reflective agency; (2) demonstrate RAF's practical relevance through case study of six widely used journaling applications and surface key design tensions; and (3) extract insights from user survey to understand user preference and perceptions around these design tensions relative to RAF's principles. Although the empirical focus of our current paper is on AI-mediated reflection systems, RAF's principles can help navigate foundational tensions that extend beyond reflection support, such as conversational AI, digital health and well-being applications, and personal informatics platforms wherever AI is present. Ultimately, we aim to provide actionable, ethically informed design considerations for the next generation of human-centered self-reflection technologies.

Philosophical Foundations: Mediated Selfhood and Ethical Becoming

Designing AI systems for self-reflection is not merely a technical challenge—it is a normative task that demands close attention to the kind of self such systems help cultivate, and how agency, growth, and meaning are co-constructed through technology. To build tools that preserve reflective agency, we draw on two complementary philosophical traditions: *phenomenology*, which interrogates the mediating structures of experience, and *Aristotelian virtue ethics*, which situates reflection within the arc of moral development. Together, these frameworks offer an understanding of the self not as a static entity to be optimized, but as a dynamic, interpretively open process of becoming.

Phenomenology: No Self Without World, No World Without Mediation

Phenomenology begins by rejecting the Cartesian separation of subject and object. Experience is not a passive reception of an objective world by a detached mind; rather, it is always situated, embodied, and mediated. As Ihde (1990) writes, "*There is no 'world-in-itself,' only a world that is experienced by someone, through some form of mediation*". This insight extends to selfhood: we do not uncover the self as an interior object, but encounter it through our relations with tools and environments.

Heidegger's concept of "*being-in-the-world*" (*Dasein*) emphasizes that identity emerges not in isolation, but through meaningful engagement with the world (Heidegger 1962). In his critique of modern technology, Heidegger warns that technological systems often *"enframe reality"*—transforming beings into *standing-reserve* (*Bestand*), resources to be ordered, quantified, and optimized (Heidegger 1977).

This warning is especially relevant in the context of AI-powered reflection tools. When such systems reduce inner life to sentiment scores, productivity metrics, or rigid emotional categories, they may risk turning our lived experience into data to be managed. We begin to relate to ourselves not as complex, evolving beings, but as systems to be monitored and improved—our thoughts and feelings treated as information rather than meaning.

Yet Heidegger also identifies within this danger a *saving power*: the potential for a different kind of revealing that resists reduction, control, and instrumental thinking. Heidegger (1977) writes, "*Where danger is, grows the saving power also*". This saving power refers to the possibility of *poetic revealing*—a mode of disclosure that does not operate through measurement, categorization, or optimization, but through openness, receptivity, and *letting-be*. In contrast to *enframing* (*Gestell*), which challenges the world to appear as *standing-reserve* (*Bestand*)—resources to be ordered and exploited—poetic revealing allows beings to come forth on their own terms, in their depth and singularity. Heidegger draws this notion from *poiēsis*, the ancient Greek term for bringing-forth, which underlies both artistic creation and natural growth. In this mode, truth is not imposed or extracted, but emerges through a process of unfolding and interpretation.

From Ihde's typology of relations to Heidegger's warning about enframing, we learn that AI tools never solely support reflection—they are participants in how reflection unfolds. Ihde extends this by describing four relational modes of human–technology interaction(Ihde 1990):

- **Embodiment relations**: the tool becomes part of the user's experience. *When I look through the glasses, I do not see them, but see through them at the world. The glasses become part of my embodied experience.*
- **Hermeneutic relations**: the tool mediates interpretation. *I do not see the thermometer itself as much as I read it to understand the temperature.*
- **Alterity relations**: the tool appears as a quasi-other. *Some technologies invite interaction as if a quasi-other… the robot seems to respond, the AI talks back.*
- **Background relations**: the tool influences experience ambiently. *Technologies like central heating or appliance noise influence us, without drawing our focus.*

AI-mediated self-reflection systems often operate across all these roles: as *embodied tools*, they structure reflective practice; as hermeneutic devices, they analyze and visualize thought; as *alterities*, they simulate companionship; as *background* agents, they subtly influence habits. Each mode shapes the reflective self differently, offering both opportunity and risk.

Central to Ihde's theory is the notion of *multistability*: technologies afford multiple stable uses, but they steer which meanings users adopt. In light of this, ethical reflection tools should aim to preserve this interpretive openness rather than impose fixed meanings, allowing for a plurality of responsible engagements.

Fostering this openness goes beyond algorithmic explainability. It involves what Ihde calls *technological disclosure*:

enabling users to sense how technology shapes thought and experience. AI reflection tools should be legible but not overbearing, supportive yet not substitutive.

Aristotelian Ethics: Prohairesis to Eudaimonia

While phenomenology reveals how reflection is always mediated through context and tools, Aristotelian ethics clarifies why reflective practice is ethically significant. In the *Nicomachean Ethics*, Aristotle links moral agency to *prohairesis* (deliberate, value-guided choice), *hexis* (stable character disposition), and *eudaimonia* (flourishing through cultivation of virtues over time) as the ultimate goal of human life (Aristotle 1908).

A key Aristotelian insight is that ethical action must originate from within. For Aristotle, moral action is voluntary when the agent is the source of action and aware of the circumstances: "*the moving principle is in the agent himself, he being aware of the particular circumstances*" (NE III.1, 1111a22).

This provides implications for AI tools: if systems preemptively interpret or reframe experience without the user's initiation or awareness, they may erode reflective agency. Even in complex or emotionally strained contexts, users must remain the source of interpretive authority.

Yet Aristotle also acknowledges the complexity of agency under constraint. Aristotle's notion of *mixed actions*—for example, a sailor who, during a storm, throws cargo overboard to save the ship—illustrates that actions taken under pressure can still be voluntary if they result from conscious choice guided by higher aims (NE III.1, 1110a11–15). This suggests that systems may justifiably assist users in emotionally difficult moments, as long as users preserve their reflective judgment. He distinguishes acting *in ignorance* from acting *because of ignorance*, the latter being truly involuntary (NE III.1, 1110b18–23). This highlights the risk of AI feedback that lacks context, potentially distorting reflection or undermining understanding.

For Aristotle, virtue is not innate but cultivated through habituation and sustained over time. *Ethical flourishing (eudaimonia)* is not a momentary state but a lifelong practice of integrating experiences, exercising judgment, and forming character. Reflection, in this sense, is not a task to be completed but an ongoing narrative process. AI systems that aim to support ethical growth should therefore attend to temporal coherence—helping users track evolving values, surface long-term patterns, and reinforce moral development. As Aristotle writes, "*Happiness... is something final and self-sufficient, and is the end of action*" (NE I.7, 1097b20–22), achieved through rational, value-aligned activity.

Together, phenomenology and virtue ethics offer a robust foundation for rethinking AI-mediated reflection—not as optimization, but as interpretive and ethical practice.

The Reflective Agency Framework

We translate these philosophical foundations into a practical framework for AI design: the Reflective Agency Framework (RAF). Comprising five interrelated principles, RAF provides actionable heuristics that help maintain the interpretive space essential for meaningful self-reflection. Developed through a synthesis of philosophical grounding and empirical findings, these principles aim to preserve the interpretive space essential for meaningful self-reflection. Each principle addresses a distinct yet overlapping dimension of reflective experience—from the initiation of thought to sustaining a coherent personal narrative over time.

Principle 1: Internal Origination

The user must remain the initiating source of reflection. Insights and meaning must arise from within, not be imposed externally.

Philosophical Grounding Aristotle's *prohairesis*—the capacity for deliberate, value-guided choice—and Heidegger's notion of authenticity both underscore that reflection must originate within the agent. When AI systems generate unsolicited insights or reframe user experience unprompted, they risk displacing this agency. Instead, systems should position users as the source of interpretive initiation.

State of the Art Empirical studies corroborate this concern: Angenius and Ghajargar (2023) found that users often perceived unsolicited AI prompts as intrusive during journaling, which they described as a sacred space. Likewise, Kim et al. (2024b) report that retrospective suggestions preserved users' sense of authorship more effectively than real-time interventions. While current evidence emphasizes the importance of user-initiated engagement, there is still a lack of trustworthy mechanisms for dynamically restoring interpretive control after interacting with system-led prompts and studies onlongitudinal effects on self-concept and skill cultivation.

Principle 2: Calibrated Responsiveness

AI systems should dynamically adapt their level of support based on the user's emotional and cognitive state, providing guidance when needed while stepping back when autonomy is preferred.

Philosophical Grounding Calibrated Responsiveness draws on Aristotle's notion of *phronesis*, or practical wisdom—the ability to discern how to act appropriately in complex, context-sensitive situations (Aristotle 1908). Ihde's analysis of *embodied* and *background* relations similarly points to the importance of fluid transitions in how technologies present themselves, sometimes receding into the background, other times coming to the fore (Ihde 1990). For reflective systems, this means that *timing* matters. If AI intervenes too readily, it may feel interruptive; if it remains silent, opportunities for support may be lost. Finding the appropriate moment requires adaptiveness to context and emotional and cognitive readiness.

State of the Art Recent systems have begun to adopt an intermittent model of engagement. Angenius and Ghajargar (2023) has presented systems stepping in only when users appear to be disengaged. The MindScape system (Nepal, Liu, and Lee 2024) varied the prompts in the journaling interface based on behavioral sensing on sleep and activity.

Code	Principle	Philosophical Foundation	Design Imperative
IO	Internal Origination	Aristotle (Voluntary Action, *Prohairesis*); Heidegger (Authenticity)	AI systems should allow the user to remain the initiating source of reflection. Insights and meaning must arise from within, rather than being imposed externally.
CR	Calibrated Responsiveness	Aristotle (*Phronesis*); Ihde (Embodied and Background Relations)	AI systems should dynamically adapt their level of support based on the user's emotional and cognitive state, providing guidance when needed while stepping back when autonomy is preferred.
RA	Reflective Ambiguity	Heidegger (Poetic Revealing); Ihde (Multistability, Hermeneutic Mediation)	AI systems should preserve the richness of user experience by embracing ambiguity and supporting multiple interpretations for interpretive expansion, rather than reductive conclusions.
TM	Transparency of Mediation	Ihde (Technological Disclosure); Heidegger (Enframing Awareness)	AI systems should make their interpretation processes transparent, allowing users to understand how outputs are generated and retain reflective authority.
SE	Self-Continuity and Ethical Flourishing	Aristotle (*Eudaimonia*, *Hexis*); Heidegger (*Dasein*'s Temporality)	AI systems should support sustained personal growth and coherent self-narratives by aligning with users' core values, while fostering critical self-reflection and cultivating self-understanding that is both ethically responsible and existentially meaningful.

Table 1: Reflective Agency Principles: Mapping Philosophical Foundations to Design Imperatives

Despite these responsive systems, many implementations still rely on surface-level heuristics without clear principles on calibration. Another key challenge lies in responsiveness: systems that adapt to users in real-time often depend on extensive personal data. For instance, employing a *digital twin* and modeling behavior can raise serious concerns around consent, surveillance, and autonomy (Vainionpää et al. 2023; Dewitte 2024). Moreover, overly responsive AI may reinforce user biases, amplify distorted thinking, or validate conspiratorial beliefs under the guise of empathy (Danry et al. 2025; Williamson and Prybutok 2024). Effective calibration requires more than technical precision—it should involve adaptive proposals users can accept, defer, or reject, so that AI support remains contextually appropriate while under user control.

Principle 3: Reflective Ambiguity

AI systems should preserve the richness of user experience by embracing ambiguity and supporting multiple interpretations for interpretive expansion, rather than reductive conclusions.

Philosophical Grounding When systems prematurely define or reductively conclude the meaning of experience they foreclose the uncertainty through which insight often emerges. Heidegger's notion of poetic revealing resists this reduction by emphasizing openness, emergence, and letting-be. Likewise, Ihde's concepts of hermeneutic mediation and multistability show that technologies are not neutral conduits but active participants in meaning-making. To foster authentic reflection, systems must sustain interpretive openness—inviting multiple meanings rather than delivering singular conclusions. Ambiguity here is not a confusion to be avoided, but a condition for depth, resonance, and growth.

State of the Art As an empirical support, Di Lodovico, Wolf, and Bardzell (2023) showed how ambiguous visualizations can encourage users to draw personal inferences. Recently, Di Lodovico, Houben, and Colombo (2025b) also revealed that incorporating ambiguity in design for self-tracking tools can allow users to interpret data in personally meaningful ways. Despite these benefits, ambiguity is often avoided in favor of clarity and efficiency in most widely-used applications. Few systems offer clear guidance on when or how to sustain ambiguity, and its effects on vulnerable users lack research.

Principle 4: Transparency of Mediation

AI systems should make their interpretive processes transparent, allowing users to understand how the outputs are generated and retain reflective authority.

Philosophical Grounding Transparency of Mediation builds on Ihde's argument that users should be able to perceive the structure of mediation itself—that is, to see how technology transforms experience (Ihde 1990). Heidegger's concern with *enframing* cautions that hidden processes may frame users' worlds without their knowledge (Heidegger 1977). As discussed in the previous section, invisible algorithmic logics can unintentionally distort user experience. If users do not understand how their data has been interpreted, they will not be able to meaningfully respond. This risks not just misunderstanding but a loss of user agency.

State of the Art Glinka and Müller-Birn (2023) show that when systems reveal their underlying assumptions, users engage more critically and thoughtfully. Similarly, Springer and Whittaker (2020); Muralidhar (2024) find that progressive disclosure—where system logic is gradually exposed—allows users to retain control without being overwhelmed. Yet current studies often focus on narrow domains or assume uniform user needs, limiting generalizability. There is little consensus on what makes explanations clear, helpful, or meaningful across diverse users.

Principle 5: Self-Continuity and Ethical Flourishing

AI systems should support sustained personal growth and coherent self-narratives by aligning with users' core values, while fostering critical self-reflection and cultivating self-understanding that is both ethically responsible and existentially meaningful.

Philosophical Grounding Selfhood unfolds through time. Aristotle's *hexis* links ethical character to repeated action and habit, while *eudaimonia* situates flourishing as a lifelong project (Aristotle 1908). Heidegger's account of *Dasein* emphasizes narrative temporality—our identities are shaped by how we integrate past, present, and future (Heidegger 1962). Reflection is central to this process: it enables individuals to examine their values, make deliberate choices, and refine their sense of Aristotle's *good*. AI systems that treat reflection purely as therapeutic relief or behavioral optimization risk flattening this richer horizon. Instead, they should preserve the interpretive space to grapple with value questions and sustain moral growth alongside personal meaning-making.

State of the Art Recent research has shown that AI systems can support temporally grounded and value-oriented reflection. In a 4-week in-home study, Maharjan et al. (2022) found that users often personalized and emotionally engaged with a speech-enabled agent for daily well-being reporting. MindScape (Nepal, Liu, and Lee 2024) generates prompts based on users' behavioral histories and Zulfikar et al. (2025) demonstrated that AI-generated suggestions based on memories improved mental health outcomes. Work on goal alignment (Mechergui and Sreedharan 2024) and survey-based value alignment for LLMs (Shen et al. 2023) underscores the importance of designing systems that behave consistently with user values. However, most current tools focus on short-term outcomes such as productivity, momentary memory support, or distress reduction, neglecting the cumulative arc of self-narrative and moral development. Designing for long-term ethical continuity requires mechanisms to maintain interpretive threads across interactions, help revisit earlier reflections in light of evolving commitments, and surface actions consistent with user values.

Principles in Action: A Case Study on AI Journaling Systems

The preceding sections introduced the five principles of the Reflective Agency Framework (RAF), tracing their philosophical groundings and supporting them with empirical evidence. In this section, we operationalize these principles by posing a concrete empirical question: *How do current AI journaling tools' design choices align with each ethical dimension—and what risks or opportunities do they pose?*

To explore this question, we adopted a two-step approach. First, we conducted a case study on the six most distinctive platforms—*Day One, Mindsera, Replika, Rosebud, Wysa, and Youper*—identifying common design tensions and investigating how these may support or compromise the aims of the Reflective Agency Framework. These platforms were selected based on public availability and sustained user adoption, and we narrowed the set to six applications that offered the greatest diversity in explicitly integrated AI features, representing different levels of AI-mediated self-reflection tools in the market. This enabled us to compare design tensions and patterns across the current landscape of AI-powered reflection support.

To map the design tensions and feature sets present in these tools, we reviewed official documentation, interfaces, and marketing materials, and actively engaged with each app, cataloging available features, user flows, and distinctive prompts. We then inductively coded features according to their relevance to reflective agency, with each tool mapped along key feature-tensions (user-driven vs system-driven) using *Day One* as a baseline for maximal user freedom (see Table 2). Building on this foundation, we aligned the mapped features and tensions with RAF.

Secondly, we conducted an exploratory user preference survey (full questions can be found in the Appendix.) with 20 respondents who has used at least one of the six AI-mediated journaling tools, balanced in age (24–54) and gender (11 women, 9 men; no respondents identified as non-binary), recruited via *CloudResearch Connect* (Hartman et al. 2023). The survey is intended mainly to provide initial insights to contextualize design tensions identified, rather than to generalize what users value in their journaling practices and how they interpret AI-mediated interventions for future large-scale empirical studies. To support a broader range of perspectives, we emphasized qualitative insights by including open-ended prompts, such as: *"In your opinion, what makes journaling meaningful to you? How could AI and technology help or hurt that meaning?"* Together, these insights help contextualize future directions for ethically grounded design.

By comparing the reflective affordances and user tensions in these widely adopted tools, our study is, to our knowledge, the first to empirically map this domain and reveal critical gaps in understanding how AI supports or impedes self-discovery. While individual platforms such as *Wysa, Replika,* and *MindScape* have been studied in isolation (Jung et al. 2024; Pentina, Hancock, and Xie 2023; Nepal et al. 2024), no peer-reviewed comparative analysis exists for the most widely used AI-powered journaling tools. Prior cross-application analyses, such as Haque et al. (2022), aggregate user feedback from app store reviews and social media to identify broad satisfaction and dissatisfaction themes but do not systematically map feature-level differences or examine how specific features influence reflective agency and self-discovery. Our case study addresses this gap by identifying design tensions and providing initial empirical support for understanding user perceptions. We sought perspectives that may challenge or conflict with our framework, using negative cases to test the RAF's robustness and inform future refinement of the framework.

Considerations in Internal Origination (IO) Survey responses emphasized the importance of maintaining control over both meaning-making and the reflection initiation. One participant wrote *"Journaling loses its value for me if someone or something tells me what I'm supposed to feel."* When

Tension Code	Reflective Autonomy (User-Led)	Systemic Guidance (System-Led)
IO-1: Initiation Support	Initiates reflection by user	Initiates reflection by AI
IO-2: Interpretation Support	User determines meaning	AI interprets and provides suggested meaning
CR-1: Emotional Sensitivity	Maintains emotional neutrality	Adapts to user's affective experience
CR-2: Cognitive Load Sensitivity	Makes no adjustment to mental state	Adapts to user's cognitive demands
RA-1: Interpretive Openness	Enables open-ended meaning-making	Provides interpretation or reductive summaries
TM-1: Reasoning Visibility	Provides explainable, transparent reasoning	Obscures reasoning (black-box AI)
TM-2: AI Disclosure	Explicitly acknowledges AI's role	Uses anthropomorphic design, potentially disguising AI as human
SE-1: Reflective Challenge	Encourages critical thinking	Offers validation and affirmation
SE-2: Narrative Continuity	Connects past to present experiences	Focuses on present-focused events
SE-3: Reflective Practice	Focuses on sustained reflection	Focuses on forming habitual logging

Table 2: Design Tensions from the Reflective Agency Framework (RAF): User-led vs System-led Ends

asked whether respondents preferred to initiate reflection, 16 of 20 participants selected "Always by me" or "Mostly by me," and four chose "Equal balance," suggesting a preference for self-directed reflection over AI involvement.

In response to *"How much should AI shape or guide your understanding of your own experiences?,"* nine participants indicated they would consider AI perspectives alongside their own, and another nine said they would often reflect on AI responses. In related qualitative responses, participants showed contrasting views. One participant stated, *"I would rely heavily on AI to help me interpret and make sense of my experiences,"* while another noted, *"I prefer to interpret everything on my own."* This range of responses reveals a key design tension around AI's role in both the initiation of reflection and the shaping of meaning, surfacing the need to find an optimal level of support for varying users.

Design Tension IO-1 : Initiation Support *How much should AI help initiate the reflection?*
Day One offers themed templates to guide entries with minimal AI mediation, where users primarily initiate their reflection. This freedom supports agency but can make it harder for some users to start reflecting. In contrast, *Mindsera* provides stage-based prompts and automated summaries. This can encourage structured introspection but risks shifting journaling into an externally framed activity, and may feel that they are provided with less freedom for already habituated users. Result shows that support should be provided based on users' current level of experience with journaling.

Design Tension IO-2 : Interpretation Support *How much should AI guide interpretations and meaning-making?* Apps like *Youper* and *Mindsera* take distinct approaches to interpretive control. *Youper* offers reflections only when prompted, emphasizing user affirmation and gentle suggestions. This preserves user authorship but may lead to a more passive and less engaging AI experience. *Mindsera*, by contrast, takes a proactive role, automatically generating summaries of mental states and mood visuals (e.g., "25% stress and 75% frustration"). While intended to support self-discovery, this approach risks reducing reflection to a quantized self. *Replika* offers another model of providing emotionally adaptive, context-aware conversational interactions that encourage users to reflect.

Our study responses reinforced this tension around origination: 13 out of 20 participants strongly or somewhat disagreed with the statement, *"If AI rephrased my journal entry, I would feel that the entry reflects my own thoughts"*, indicating a potential disconnect between AI input and users' sense of authenticity. However, participants also emphasized the importance of control over how content is recorded and shaped. 11 respondents chose, *"I want the final say on what gets recorded, but I'm open to AI suggestions,"* while five rejected any AI input and four preferred a more balanced collaboration. These responses suggest that users are more open to AI involvement if they can retain final authority over their reflections.

Considerations in Calibrated Responsiveness (CR) Open-ended responses revealed a consistent tension between the value of AI support and the need for emotional and cognitive alignment. Many participants described journaling as a tool for processing grief, stress, or insight—contexts where AI could help if attuned to the moment. One shared, *"Journaling is meaningful to me in helping to manage my grief. . . having AI point out repeated issues might help gain insight."* Others noted how AI might support during stress or overwhelm: *"It makes it easy to denoise the daily stressors,"* and *"AI prompts could give me the direction I need to elaborate on my feelings."* Yet even among those receptive to AI, respondents emphasized: *"It should act as a companion, not a substitute for your own inner voice."*

Level of AI Guidance	Count
High – I would rely heavily on AI to help me interpret and make sense of my experiences	1
Moderate – I would often reflect on AI's interpretations to better understand myself	8
Some – I'm willing to consider AI perspectives alongside my own	10
Minimal – I'm open to light suggestions, but mostly interpret myself	0
None – I prefer to interpret everything on my own	1

Table 3: Preferred level of AI involvement in guiding personal understanding (N = 20)

Design Tension CR-1: Emotional Sensitivity *How much should AI adapt to the user's emotional and cognitive state?* CAs' ability to adjust responses in real-time offers reflective support, unlike static journaling interfaces.*Replika* exemplifies this by tailoring its responses based on user mood, personality traits, and conversation history, potentially enabling

more meaningful engagement during periods of emotional vulnerability or cognitive fatigue. However, such flexibility can also risk misinterpreting user needs or overstepping interpretive boundaries. *Wysa*, by contrast, employs a less flexible structured message exchange pattern to deliver consistent and evidence-based mental health support. While this approach may enforce trust, it may limit responsiveness to subtle emotional or contextual changes.

One interesting finding is that our survey revealed varying levels of AI guidance needed based on their emotional state (Figure 1): respondents preferred more structured support for emotions like sadness and anxiety, while they desired more mixed support when feeling joy and inspiration. Notably, anger prompted a strong leaning toward a need for "some guidance". These findings suggest users generally prefer emotion-responsive systems, adjusting the level of guidance based on how they're feeling—favoring support in distress and autonomy in lighter or creative moods.

Design Tension CR-2: Cognitive Load Sensitivity *Should AI offer support only when requested by the user, or should it actively direct the flow and content of reflection?* A core challenge in AI responsiveness is calibrating when and how to guide reflection—avoiding both premature intervention and excessive passivity. CAs are often favored because they resemble natural dialogue. However, in apps like *Wysa* and *Replika*, users may perceive the AI as holding equal or greater authorship, which can shift the balance of input and interpretive control away from the user.

Rosebud may help mediate this imbalance by presenting optional prompts while allowing the user to write freely, with AI suggestions offered as invitations for deeper reflection. Similarly, in *Mindsera*, adaptive assistance is provided at key points in the journaling flow, but only when activated with the user's permission. These systems illustrate different strategies for calibrating AI-initiated guidance to preserve user authorship. However, challenges remain in calibrating responsiveness to cognitive load, which we find yet to be addressed, unlike emotional adaptation in current systems.

Considerations in Reflective Ambiguity (RA) Across both survey and open-ended responses, participants expressed a strong preference for journaling tools that protect interpretive openness. While AI support was often welcomed in principle, users voiced concern that such support should not take the form of fixed summaries or diagnostic outputs. Instead, AI was seen as most valuable when it invited deeper thought without asserting conclusions. As one participant explained, *"AI should provide potential explanations when prompted and acknowledge tendencies, but not make concrete observations and determinations for me."* Another echoed this sentiment: *"AI prompts to guide my writing could give me the direction I need to elaborate on my feelings."* These perspectives reflect a broader resistance to system-led meaning-making and a desire for ambiguity to remain an active, generative part of reflection.

Design Tension RA-1: Interpretive Openness *Should AI support multiple ways of understanding experience, or should it help categorize and provide summarized insights?* Several journaling apps position themselves as interpretive aids by offering post-entry summaries, visual themes, or emotional analyses. In *Mindsera* and *Rosebud*, AI-generated overviews are presented immediately after a journal is written. These features aim to reinforce insight but may preempt the user's own reflection, especially when delivered before they've had a chance to revisit or reframe their thoughts. Similarly, *Youper*'s analytics dashboard provides predictive feedback on emotional trends, representing complex internal experiences as discrete metrics. *DayOne* may not reduce meaning but also doesn't add new insights into understanding experience. Most of these analytical features categorize and condense meaning without inviting further reflection, particularly among systems that rely on the CAs.

Survey responses suggest strong user sensitivity to this trade-off. When asked whether they value journaling tools that leave room for personal interpretation, eight participants said "Strongly agree", nine participants said "Agree", and three responses were "Neither Agree nor Disagree", with none disagreed. Participants rated AI that *expanded* reflections (e.g., prompts) more positively than AI that *condensed* them (e.g., summaries) (Tables 4–5). Expansion aligns with the Reflective Ambiguity principle by preserving interpretive openness. The more mixed ratings for condensation suggest that user preference around AI features for condensing reflection may conflict with this principle, or that its role requires further investigation.

AI Expanding Reflection	Count
Very Helpful	15
Moderately Helpful	5
Very Unhelpful, Slightly Unhelpful, Neither helpful	0

Table 4: Helpfulness ratings for AI help expanding entries (e.g., interpretations, prompts)

AI Condensing Reflection	Count
Very Helpful	10
Moderately Helpful	5
Neither helpful nor unhelpful	4
Very Unhelpful	1
Slightly Unhelpful	0

Table 5: Helpfulness ratings for AI help condensing entries (e.g., summaries, tags)

Considerations in Transparency of Mediation (TM) We identified two levels of transparency: acknowledgment that the AI is not human during conversations or how provided suggestions and insights are generated. Survey responses revealed concerns about how AI communicates its presence and reasoning in reflective systems. When asked how AI should respond if it misinterpreted a journal entry, users preferred to correct the AI or have it apologize over asking a follow-up question to clarify an automatic update.

However, in terms of frequency of reminders, we had mixed results (Table 8). These mixed responses suggest that

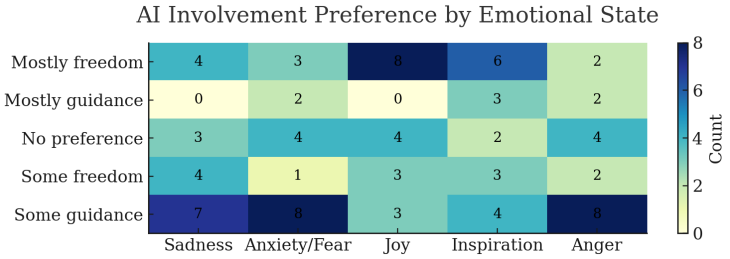

Figure 1: AI Involvement Preference by Emotional State. The heatmap shows the distribution of journaling guidance preferences across different emotional contexts. (N=20)

Preferred AI Response	**Count**
Allow me to easily correct the AI	11
Apologize and let me decide whether to update it	6
Never label or interpret my entries	2
It doesn't matter to me	1
Ask follow-up questions to clarify	0
Automatically update based on tone or keywords	0
Ignore it and continue as normal	0

Table 6: How participants prefer AI to handle misinterpreting a journal entry (N = 20)

Trust Factor	**Count**
Privacy (e.g., strict privacy policy)	16
Transparency (e.g., clear references or citations for AI-generated insights)	13
Open-source or verifiable technology	9
Expert validation (e.g., reviewed by mental health professionals or relevant experts)	8
Reputation (e.g., strong reviews or credible endorsements)	6
Explainability (e.g., reasoning of how AI generated insights)	6

Table 7: Factors that increase user trust in AI journaling tools (N = 20, multiple selections allowed)

Reminder Preference	**Count**
I prefer frequent reminders that I am interacting with AI	6
I prefer occasional reminders that I am interacting with AI	8
I don't mind whether I'm reminded or not	2
I prefer not to be reminded that I'm interacting with AI	3
I strongly prefer not to be reminded	1

Table 8: Participant preferences for being reminded they are interacting with AI (N = 20)

while many users value AI transparency, others find it disruptive. As one participant put it, AI can feel *"overbearing and misguiding,"* especially when it interrupts reflective flow or appears uninvited.

Design Tension TM-1: AI Reasoning Visibility *Can users see and understand how AI suggestions are generated, or is the logic hidden?*
Among the platforms, *Mindsera* stands out for offering an onboarding experience that introduces its AI features step by step. Before presenting the first round of AI-generated feedback, users see a screen explaining the *"Interactive Journaling"* interface. The *"Go Deeper"* tool is similarly contextualized, and the app notes that users can customize the experience. However, they do not offer the steps the AI takes to arrive at its interpretations. There is no separate option for users to see how the data is used, what inference process the AI applies, or why particular summaries or suggestions are surfaced. *Rosebud* follows a similar pattern. The app opens with a basic introduction to its journaling interface, but the origins of its AI feedback remain a black box. Once journaling begins, suggestions appear seamlessly, without explanation of how they were derived or whether they are based on prior entries, linguistic sentiment, or preset heuristics. For users, this opacity can undermine trust, especially when feedback contradicts their self-perception (Table 7).

Other platforms like *Replika*, *Wysa*, and *Youper* provided less context. Users are placed directly into interactions with CAs, without any background on how the system interprets or generates responses. This relates to one response: *"AI could hurt the meaning by being too overbearing and misguiding my reflection into areas that are not as relevant to me."* Transparency is often undermined by the prioritization of engaging user experience in current applications.

Design Tension TM-2: AI Disclosure *Is the AI identified as non-human, or does it blur the line between human and machine?*
None of the apps explicitly acknowledged AI use during user interactions. On their websites, *Mindsera* labels features such as *"Artworks," "Minds,"* and *"Go Deeper"* as AI-powered, while *Replika* markets itself as an AI companion, leaving users to infer its non-human nature. Within the interfaces, the AI interacts without identifying itself, often mimicking a human coach or partner. Other apps offer fewer cues: *Wysa* states on its website that AI handles most interactions but reveals this in-app only when asked; *Rosebud* briefly mentions "AI analysis"; and *Youper* makes no AI reference, restricting input to multiple-choice responses and further obscuring agency. This may risk users mistaking its suggestions for human reasoning or overestimating its authority. Participants shared this concern; one noted, *"excessive guidance reduces authenticity."*

Together, these findings suggest that transparency is not just a matter of labeling—it is a matter of *interactional ethics*. Although participants identified privacy and transparency as top trust factors (Table 7), and many preferred frequent or occasional reminders that they were interacting

with AI (Table 8), current applications fall short in addressing these expectations.

Considerations in Self-continuity and Ethical Flourishing (SE) Participant responses revealed a meaningful tension between using journaling for immediate emotional regulation and for long-term development. Participants expressed a desire for journaling tools that help them manage emotions, reduce stress, strengthen resilience, or practice self-compassion, while fewer prioritized developing character traits such as confidence and self-worth (Table 9).

Respondents described journaling as most impactful when it surfaced patterns and prompted deeper insight over time. One noted, *"that aha moment when I see a pattern of thought or behavior that is not serving me"*, while several said AI could assist by *"pointing out repeated issues"* or *"clarifying patterns"*. Others emphasized temporal self-review as most useful: *"being able to remember what I was thinking in certain periods helps me form goals"*. Conversely, participants also valued in-the-moment support, such as *"helping me further examine the emotions and thoughts I am having"*.

Lastly, one participant shared, *"Journaling can help create a sense of self-worth, purpose, and self-love. It can also show us how we treat people and how we can make interactions with them better."* This illustrates the potential for AI tools to cultivate values beyond the momentary support and toward compassion for others.

Development Goals for Journaling	Count
Managing emotions and reducing stress	11
Strengthening resilience and adaptability in challenging situations	11
Practicing self-compassion during difficult times	10
Becoming more present and mindful in daily life	10
Clarifying my sense of purpose and motivation	8
Cultivating gratitude and a more positive mindset	7
Reflecting on my values and making thoughtful decisions	6
Building confidence and a stronger sense of self-worth	6
None / Other	1

Table 9: Participant goals for character development through journaling (N = 20, multiple selections allowed)

Design Tension SE-1: Challenge vs. Affirmation *Should AI push users to think critically or prioritize affirmation?*
In current systems, the capacity to challenge users varies. Applications like *Wysa* and *Youper* often have fixed supportive scripts or CBT-style exercises. For example, when a user expresses ongoing stress, *Wysa* by default redirects them to a gratitude list—an approach that, while well-intentioned, may overlook opportunities to explore underlying causes or internal conflicts.

Survey insights show that most participants were open to discussing life challenges with AI, but preferred practical, reflective prompts over purely comforting responses. Commonly selected features included objective advice, insightful questions, and adaptive responses, while fewer participants chose comforting strategies like soothing words (Table 10). These preferences also align with the principle of *Calibrated Agency*—emphasizing the importance of tailoring support to a user's cognitive and emotional state.

Open-ended responses echoed this theme. One noted, *"AI could guide me toward thinking deeper about things,"* while another emphasized that journaling helps to *"uncover patterns, values, or truths that deepen your self-awareness or shift how you understand your emotions, actions, or goals."* Together, these responses suggest that reflective challenge beyond therapeutic purposes can foster growth and deeper reflection in ways that user affirmation alone cannot.

AI Assistance Preference	Count
Objective advice and strategies	11
Insightful questions for deeper reflection	10
A combination depending on my needs at the moment	10
Encouragement and validation	8
Help rewriting my story	7
Soothing, empathetic words	5
I would not discuss challenges in my life with AI	0
Other	0

Table 10: Participant preferences for AI assistance when discussing life challenges (N = 20, multiple selections allowed)

Design Tension SE-2: Narrative Continuity vs. Momentary Insight *Should systems help connect present with past reflections or focus on the momentary insights?*
Participants frequently highlighted the importance of seeing patterns emerge over time. Yet their ability to act on these insights depended on whether systems can make prior content accessible and meaningful. One participant said journaling helps *uncover patterns that deepen self-awareness,"* and another stated that *AI could help by surfacing things I never thought about,"* highlighting a desire for systems that connect insights over time.

Mindsera offers the most robust support for longitudinal reflection. Features such as *"Recurring Topics," "Emotional States,"* and *"Personality"* use AI to analyze historical entries and visualize recurring themes. *Rosebud* generates prompts on past content. These approaches help link present thoughts to prior patterns, potentially fostering sustained self-insights. In contrast, *Replika* retains past interactions but seldom references them. *Day One* and *Youper* provide minimal longitudinal scaffolding beyond basic calendar or mood-board features. *Wysa* keeps each session isolated.

This gap between user-preference and feature availability is notable. When asked about how often they revisit past journal entries, most participants indicated they do occasionally at best. Eight had looked back within the past month, four within the past week or day, and three rarely or never. This suggests a potential desire for longitudinal reflection.

Design Tension SE-3: Reflective Practice vs. Habitual Logging *How should systems support sustained self-reflection for long-term growth, or do they merely reinforce repetitive use?*
Supporting reflective continuity requires more than streak counts or push notifications—it demands keeping users connected to their evolving self-narrative. To understand which features can facilitate habit formation around self-reflection,

participants were asked to select all the features they perceive as helpful to develop a consistent journaling habit. Top features (Table 11) included reminders and dedicated journaling times, but feedback preferences (Table 12) varied widely, from real-time input to monthly summaries. This diversity highlights the need for adaptiveness rather than enforcing uniform engagement. As one participant put it, AI should *"track my thought processes"* and *"further examine the emotions and thoughts I'm having"*—linking present insights to past patterns without gamifying the process.

Habit-Building Feature	**Count**
Reminder & notification	13
Setting a dedicated journaling time	12
Encouraging feedback once journaling is completed	9
Daily streak tracking or rewards	8
Connecting to journaling community or sharing with friend	4
Features will not help / not needed	1
Other (please specify)	0

Table 11: Preferred features for building a consistent journaling habit (N = 20, multiple selections allowed)

Feedback Timing Preference	**Count**
Feedback while writing journal	7
Feedback instantly after I complete a journal entry	5
Daily insight/summary (once a day)	9
Weekly insight/summary (once a week)	9
Monthly insight/summary	6
Only when I ask for it	4
I don't want feedback from the AI	**2**

Table 12: Preference for when to receive AI feedback about journaling (N = 20, multiple selections allowed)

Among the six platforms, *Day One* supported in-app notifications and a calendar interface to encourage daily streaks. *Wysa* and *Replika* rely on daily reminders to re-initiate chat-based interaction. While helpful for habit formation, these features can veer into productivity framing, which may undermine reflective intention over time.

These findings suggest that existing features prioritize repetitive interaction over the user's reflective growth. Continuity in journaling is not only about repetition; it should help construct a coherent, values-oriented self-narrative over time. To support *Continuity and Ethical Flourishing*, AI-mediated tools should be able to remember and relate to the evolving arc of the person—supporting reflective challenge without overriding agency, and sustaining engagement without reducing reflection to a habit loop.

Limitations and Future Work

This paper introduces the Reflective Agency Framework (RAF), a philosophically grounded set of principles for AI-mediated systems aimed at preserving user agency. To explore its practical implications, we conducted a case study of six AI journaling tools and an exploratory survey. Although these offer early insights, the work is limited by a small and diverse sample with varied prior experience. Moreover, while user preferences help surface needs, they may not always align with long-term outcomes. Our aim is not to make universal claims, but to expose tensions, provoke design considerations, and highlight areas that warrant further study.

In addition, our findings reinforce that reflective needs and attitudes diverge by context, stage of change, and personal goals, underscoring the importance of viewing RAF as a living framework that adapts across sub-populations rather than a fixed set of prescriptions (Delgado et al. 2023; Ahrweiler et al. 2025). Future research should also examine how RAF relates to existing guidelines that emphasize frictionless experience, as it intentionally introduces "productive friction" where necessary to safeguard reflective autonomy (Sengers et al. 2005; Natali 2024).

Conclusion

As AI systems increasingly participate in our inner lives, the question is no longer whether they support introspection, but how they shape the evolving self. This paper introduced RAF, comprising five principles grounded in phenomenology and virtue ethics, to ensure that humans remain the primary agents of their own self-discovery.

Through a comparative analysis of journaling tools and an exploratory user study, we identified recurring design tensions that show how features intended to help users can also constrain reflective autonomy. As reflective technologies evolve, the challenge is not to automate self-understanding but to preserve the conditions that make it meaningful. Although our case study focused on journaling, the design tensions and empirical insights extend to a broader spectrum of systems, including intelligent coaches and digital companions (Song et al. 2025; Zhou et al. 2020), as well as learning assistants and self-tracking tools (Lupton 2016). Ultimately, technologies that augment self-reflection should advance Heidegger's notion of *Dasein* as self-becoming and self-understanding, while also supporting Aristotle's *hexis* and *eudaimonia* — human flourishing.

Ethical Statement

We emphasize that the RAF is not intended to promote AI journaling as superior than the proven value of traditional reflection methods. Pen-and-paper journaling has been shown to foster identity development (Lin et al. 2025) and deep self-reflection, while certain digital formats—such as positive-affect journaling—are more effective for reducing distress and enhancing well-being (Smyth et al. 2018). Instead, our framework is designed for the growing range of technologies where AI is already present and aims to ensure that agency, interpretive openness, and long-term flourishing remain protected even as automation expands (Sengers et al. 2005; Gaver, Beaver, and Benford 2003). In addition, we recognize that there is no one-size-fits-all solution, so following all aspects of our RAF framework may not be ideal for some cases; rather, each with a distinct focus and functionality may be better for a sub-population for different targeted purposes. We call for action to consider how their applications and design choices may impact user agency.

References

Ahrweiler, P.; Capellas, B. L.; Wurster, D.; Späth, E.; and Siqueiros García, J. M. 2025. Inclusive Technology Codesign for Participatory AI. In Ahrweiler, P., ed., *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society. Springer.

Angenius, J.; and Ghajargar, M. 2023. Designing with Respect for Reflection: Users' Attitudes Towards AI in Journaling. In *Proceedings of the Designing Interactive Systems Conference (DIS)*.

Aristotle. 1908. *The Nicomachean Ethics*. Oxford: Clarendon Press.

Baumer, E. P. 2015. Reflective informatics: Conceptual dimensions for designing technologies of reflection. *CHI*, 585–594.

Cooney, M.; Pashami, S.; Sant'Anna, A.; Fan, Y.; and Nowaczyk, S. 2018. Pitfalls of Affective Computing: How can the automatic visual communication of emotions lead to harm, and what can be done to mitigate such risks. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, 1563–1566. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

Danry, V.; Pataranutaporn, P.; Groh, M.; and Epstein, Z. 2025. Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 2023 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. New York, NY, USA: Association for Computing Machinery.

Dewitte, P. 2024. Better Alone Than in Bad Company: Addressing the Risks of Companion Chatbots Through Data Protection by Design. *Computer Law & Security Review*, 54(106019).

Di Lodovico, C.; Houben, S.; and Colombo, S. 2025a. How to Design with Ambiguity: Insights from Self-tracking Wearables. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Di Lodovico, C.; Houben, S.; and Colombo, S. 2025b. How to Design with Ambiguity: Insights from Self-tracking Wearables. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Di Lodovico, L.; Wolf, T.; and Bardzell, J. 2023. Ambiguity as Design Resource in Self-Tracking Visualizations. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.

Gaver, W. W.; Beaver, J.; and Benford, S. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, 233–240. New York, NY, USA: Association for Computing Machinery. ISBN 1581136307.

Ghandeharioun, A.; McDuff, D.; Czerwinski, M.; and Rowan, K. 2019. EMMA: An Emotion-Aware Wellbeing Chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–7.

Glinka, N.; and Müller-Birn, C. 2023. Transparent Ambiguity: Epistemic Humility in Reflective AI Tools. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Haque, A. U.; Winata, G. I.; Xu, P.; and Fung, P. 2022. A Cross-application Analysis of User Reviews for Mental Health Conversational Bots and Journaling Tools. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1235–1246. Seattle, WA, USA: Association for Computational Linguistics.

Hartman, R.; Moss, A. J.; Jaffe, S. N.; Rosenzweig, C.; Litman, L.; and Robinson, J. 2023. Introducing Connect by CloudResearch: Advancing online participant recruitment in the digital age. *PsyArXiv*.

Heidegger, M. 1962. *Being and Time*. New York: Harper and Row.

Heidegger, M. 1977. The Question Concerning Technology. In Lovitt, W., ed., *The Question Concerning Technology and Other Essays*, 3–35. New York: Harper and Row.

Hussain, S. R. 2025. The Personalization Paradox: When AI Stops Helping and Starts Overwhelming. Available at SSRN: https://ssrn.com/abstract=5223317 or http://dx.doi.org/10.2139/ssrn.5223317.

Ihde, D. 1990. *Technology and the Lifeworld: From Garden to Earth*. Bloomington: Indiana University Press.

Jung, G.; Choi, S.; Kang, Y.; and Kim, J. 2024. MyListener: An AI-Mediated Journaling Mobile Application for Alleviating Depression and Loneliness Using Contextual Data. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '24, 137–141. Association for Computing Machinery.

Kim, T.; Bae, S.; Kim, H. A.; Lee, S.-W.; Hong, H.; Yang, C.; and Kim, Y.-H. 2024a. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery.

Kim, T.; Shin, D.; Kim, Y.-H.; and Hong, H. 2024b. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in*

Computing Systems, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Kocielnik, R.; Xiao, L.; Avrahami, D.; and Hsieh, G. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2).

Li, I.; Dey, A. K.; and Forlizzi, J. 2010. A Stage-Based Model of Personal Informatics Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 557–566. ACM.

Lin, K.; Kawai-Yue, J.; Sklar, A.; Hecht, L.; Sterman, S.; and Tseng, T. 2025. Crafting a Personal Journaling Practice: Negotiating Ecosystems of Materials, Personal Context, and Community in Analog Journaling. Creativity and Cognition 2025, arXiv:2504.19767.

Lupton, D. 2016. *The Quantified Self*. Cambridge, UK: Polity Press. Relevant discussion in Chapters 1 and 4.

Maharjan, R.; Doherty, K.; Rohani, D. A.; Bækgaard, P.; and Bardram, J. E. 2022. Experiences of a Speech-enabled Conversational Agent for the Self-report of Well-being among People Living with Affective Disorders: An In-the-Wild Study. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(2): 1–29.

Mechergui, M.; and Sreedharan, S. 2024. Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10650–10658. AAAI Press.

Muralidhar, D. 2024. The Effect of Progressive Disclosure in the Transparency of Explainable Artificial Intelligence Systems. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 382–383.

Natali, C. 2024. Frictional AI: Designing Desirable Inefficiencies in Decision Support Systems for Knowledge Work. In *Proceedings of the 22nd European Conference on Computer-Supported Cooperative Work (ECSCW 2024): Doctoral Colloquium Contributions*. European Society for Socially Embedded Technologies (EUSSET).

Natali, C.; Frischmann, B. M.; and Cabitza, F. 2024. Stimulating Cognitive Engagement in Hybrid Decision-Making: Friction, Reliance and Biases (Workshop Preface). In *HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence, Malmö, Sweden*, CEUR Workshop Proceedings, vol. 3825.

Nepal, S.; Liu, S.; and Lee, J. 2024. MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–13. ACM.

Nepal, S.; Pillai, A.; Campbell, W.; Massachi, T.; Choi, E. S.; Xu, X.; Kuc, J.; Huckins, J. F.; Holden, J.; Depp, C.; Jacobson, N.; Czerwinski, M. P.; Granholm, E.; and Campbell, A. 2024. Contextual AI Journaling: Integrating LLM and Time Series Behavioral Sensing Technology to Promote Self-Reflection and Well-being using the MindScape App. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. New York, NY, USA: Association for Computing Machinery.

Pentina, I.; Hancock, T.; and Xie, T. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Comput. Human Behav.*, 140: 107600.

Reed, C. N.; Benito, A. L.; Caspe, F.; and McPherson, A. P. 2024. Shifting Ambiguity, Collapsing Indeterminacy: Designing with Data as Baradian Apparatus. *ACM Trans. Comput.-Hum. Interact.*, 31(6).

Sengers, P.; Boehner, K.; David, S.; and Kaye, J. J. 2005. Reflective design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, CC '05, 49–58. New York, NY, USA: Association for Computing Machinery. ISBN 1595932038.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large Language Model Alignment: A Survey. *arXiv preprint arXiv:2309.15025*.

Shickel, B.; Siegel, S.; Heesacker, M.; Benton, S.; and Rashidi, P. 2020. Automatic Detection and Classification of Cognitive Distortions in Mental Health Text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 275–280.

Smyth, J. M.; Johnson, J. A.; Auer, B. J.; Lehman, E.; Talamo, G.; and Sciamanna, C. N. 2018. Online Positive Affect Journaling in the Improvement of Mental Distress and Well-Being in General Medical Patients With Elevated Anxiety Symptoms: A Preliminary Randomized Controlled Trial. *JMIR Mental Health*, 5(4): e11290.

Song, I.; Park, S.; Pendse, S. R.; Schleider, J. L.; De Choudhury, M.; and Kim, Y.-H. 2025. ExploreSelf: Fostering User-driven Exploration and Reflection on Personal Challenges with Adaptive Guidance by Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery.

Springer, A.; and Whittaker, S. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Trans. Interact. Intell. Syst.*, 10(4).

Vainionpää, F.; Kinnula, M.; Kinnula, A.; Kuutti, K.; and Hosio, S. 2023. HCI and Digital Twins – A Critical Look: A Literature Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

Whitmore, N. W.; Chan, S.; Zhang, J.; Chwalek, P.; Chin, S.; and Maes, P. 2024. Improving Attention Using Wearables via Haptic and Multimodal Rhythmic Stimuli. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Williamson, S. M.; and Prybutok, V. 2024. The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. *Information*, 15(6): 299.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Comput. Linguist. Assoc. Comput. Linguist*, 46(1): 53–93.

Zulfikar, W.; Chiaravalloti, T.; Shen, J.; Picard, R.; and Maes, P. 2025. Resonance: Drawing from Memories to

Imagine Positive Futures through AI-Augmented Journaling. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Woodstock, NY: ACM.