Beyond Detection: A Critical Perspective on User Agency in Digital Mental Health

Shi Chen
schen74@uoguelph.ca
University of Guelph
Guelph, Ontario, Canada

Zhao Zhao
University of Guelph
Guelph, Canada
zzhao24@uoguelph.ca

Abstract

Digital mental health research has been dominated by a data-centric paradigm that treats fragmented sensor signals as objective ground truth, prioritizing predictive accuracy over meaningful user support. This paper challenges that paradigm through a critical perspective grounded in a purposive synthesis of 24 recent studies at the intersection of interactive systems and digital health. We argue that computational models systematically overlook the subjective, context-dependent nature of psychological experience. This misalignment is not a technical shortcoming but a structural consequence of how the problem has been framed. To address this, we introduce two original constructs. First, the *complementary evidence chain* reconceptualizes mental health assessment as the integration of subjective, behavioral, and physiological data, in which no single modality serves as ground truth and cross-modal divergence is treated as a meaningful signal rather than noise. Second, a *user-agency-centered architecture* operationalizes this reconceptualization through two mechanisms: *algorithmic veto power*, which positions users as the governing authority over their own data inferences, and *graceful degradation of inference*, which preserves system support when users restrict data access. Together, these constructs articulate a system-level transition from passive emotion recognition toward interpretable, participatory, and privacy-preserving mental health support. We conclude with a set of concrete design implications and open research directions for the interactive systems community.

CCS Concepts

• **Human-centered computing** → **User studies**; *Ubiquitous and mobile computing design and evaluation methods*; • **Applied computing** → Health informatics.

Keywords

digital mental health; user agency; complementary evidence chain; algorithmic veto power; affective computing; self-management

ACM Reference Format:
Shi Chen and Zhao Zhao. 2018. Beyond Detection: A Critical Perspective on User Agency in Digital Mental Health. In *Proceedings of Graphics Interface (GI '26)*. ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

GI '26, Kitchener-Waterloo, ON, Canada

ACM ISBN 978-1-4503-XXXX-X/2026/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Mental health challenges have become a pervasive global concern, affecting individuals' cognitive functioning, social participation, and overall quality of life [79]. According to the World Health Organization (WHO), nearly one in eight people worldwide live with a mental disorder, and prevalence has increased during and after the COVID-19 pandemic [55]. Despite this scale, a large proportion of those experiencing psychological distress do not receive professional care. Structural barriers limit access to services, and high costs and a shortage of trained professionals hinder timely intervention. Personal factors compound the problem further: stigma, privacy concerns, and fear of social consequences discourage help-seeking [58].

Recent advances in digital technology have created new opportunities to address this gap. The widespread adoption of smartphones and wearable devices has enabled continuous, unobtrusive collection of behavioral and physiological data in naturalistic settings [18, 63]. As generative AI (GenAI) evolves, large language models (LLMs) can interact with users in natural language and extract emotional cues from daily conversations [83]. These technologies capture diverse signals: activity patterns, sleep behaviors, heart rate variability (HRV), and linguistic expressions. Such signals support machine learning models for the early detection and monitoring of mental health conditions [57]. By contrast with traditional assessments, which are episodic and context-bound [50], these approaches promise longitudinal, ecologically valid insight into psychological states.

Yet technical capability has not translated into meaningful mental health support. Prior work has consistently shown that AI-driven mental health research remains focused on predictive performance, paying insufficient attention to user experience, interpretability, and real-world applicability [80]. Ethical concerns further complicate this landscape. Data agency, informed consent, and the over-medicalization of everyday emotional experience raise critical questions about how such systems should be designed and deployed [13]. The central question is no longer whether mental health states can be inferred from data, but how such inferences can be integrated into daily life in ways that genuinely serve the people they are meant to support.

1.1 Problem Statement

We argue that the dominant *data-centric* design paradigm in digital mental health is structurally misaligned with the nature of psychological experience. We define this paradigm by three interlocking commitments: (1) *inferential primacy*—the assumption that mental health states are objectively detectable through sufficiently rich sensor data; (2) *user passivity*—users are treated as data sources rather than interpreters of their own experience; and (3) *accuracy-centric evaluation*—system success is measured by classification performance on predefined labels rather than by outcomes that matter to the person. By *meaningful outcomes* we refer to outcomes that matter to everyday self-management. These include interpretive outcomes (self-understanding, trust calibration), behavioral outcomes (revised coping strategies), data-related outcomes (contextual labeling, selective sharing), lifestyle outcomes (changes in sleep or routine), and care-related outcomes (appropriate escalation, communication with clinicians). This paradigm is not merely a design preference. It is encoded into how datasets are constructed, how models are trained, and how systems are evaluated [13, 79]. It stands in contrast to an *interpretation-centered* alternative, which treats mental health understanding as an ongoing, context-sensitive process in which users actively participate. This paper develops the conceptual and design vocabulary for that alternative.

The consequences of this paradigm are visible at every level of system design. When system outputs diverge from users' subjective experiences, as is inevitable given the ambiguity of physiological and behavioral signals, they can produce confusion, mistrust, or disengagement [33, 62]. Emotion recognition models compound this by relying on predefined labels drawn from discrete or dimensional frameworks. These labels encode specific theoretical and cultural assumptions [8], and risk oversimplifying diverse affective experiences across populations [19]. Individual variability in baselines, coping strategies, and cultural norms is treated as statistical noise rather than as the meaningful signal it is [64].

1.2 Contribution and Scope

This is not a system paper and not a systematic review. It is a critical perspective paper that synthesizes contradictions in digital mental health research and translates them into a theoretical framework for user-governed interpretation of uncertain, multimodal mental health data.

This paper makes three contributions. **First**, we articulate a diagnostic critique of the data-centric paradigm. A structured synthesis of 24 recent Human-Computer Interaction (HCI) and digital mental health studies shows that its failure to support meaningful outcomes is structural rather than accidental. The failure stems from a tendency to treat mental health as an objective state detectable through increasingly sophisticated models. **Second**, we introduce the *complementary evidence chain* as a conceptual model for treating subjective, behavioral, and physiological evidence as partial and relational rather than hierarchical. The principle of *divergence-as-signal* follows directly from this model: cross-modal friction is preserved as a design resource rather than smoothed away. **Third**, we propose a *user-agency-centered architecture* that operationalizes the evidence chain through two mechanisms: *algorithmic veto power* and *graceful degradation of inference*. Together, these mechanisms position users as the governing authority over their own data and inferences, shifting digital mental health technologies from prediction-oriented toward interpretation-centered systems.

We address the following research question:

> **RQ** How can interactive system design support users in interpreting, contextualizing, and acting upon multimodal data about their psychological and behavioral states, in ways that enhance clearly specified self-management outcomes?

2 Background

2.1 Foundations of Emotion and Mental Health Measurement

Understanding how digital systems can support mental health requires a clear account of how emotions and psychological states are conceptualized and measured. Within psychology, two paradigms have shaped the study of emotion: discrete and dimensional models. Discrete theories, most prominently associated with Ekman [22], propose a set of biologically innate and universally recognizable basic emotions: happiness, sadness, anger, fear, disgust, and surprise. These categories have been widely adopted in early affective computing because they map readily onto classification tasks. Dimensional models represent emotions as continuous states along core axes; Russell's circumplex model plots affect along valence and arousal [59], capturing gradations and mixed emotional experiences that resist discrete categorization.

Despite their differences, both paradigms share a limitation: they treat emotions as entities that can be mapped onto observable signals. Contemporary psychological theory challenges this assumption. Barrett [8] argues that emotions are not directly measurable objects but constructed experiences arising from the interaction of physiological processes, contextual factors, and subjective interpretation. Any attempt to measure mental health must therefore rely on indirect indicators. Each such indicator provides only a partial and situated perspective — a fact that carries significant implications for computational system design.

Mental health assessment has traditionally relied on self-report methods: clinical interviews, standardized questionnaires, and ecological momentary assessments. Such approaches are valued for their interpretability and clinical grounding, but are limited by recall bias, social desirability, and their episodic nature [37, 76]. In response, researchers have increasingly turned to physiological and behavioral data as complementary sources. Continuous measurement of HRV, electrodermal activity (EDA), and sleep patterns offers physiological indicators of stress and arousal [31, 56]. Behavioral data from smartphones and wearables (e.g., mobility patterns, communication behaviors, and digital interaction traces) provide longitudinal insight into daily routines [60, 82].

The introduction of these data sources does not resolve the fundamental challenges of measurement. Even in combination, physiological, behavioral, and self-report signals offer only partial, situated perspectives on mental health. From an interaction design perspective, the critical problem is the persistent reductionism that treats a single modality as a definitive proxy for ground truth. This paper challenges that assumption by proposing a multi-modal framework that retains modality partiality as a design resource.

2.2 From Standardized Models to Personalized, User-Centered Design

The WHO defines self-care as the capacity of individuals, families, and communities to promote and maintain health with or without the support of health workers [54]. Within HCI, this framing has catalyzed interest in mental health technologies that move beyond passive monitoring. Systems are increasingly designed to foster interpretation and reflection, prioritizing personal agency and self-efficacy [72].

Several reviews have surveyed this shift. Pinge et al. [57] identified a recurring limitation in wearable-based stress detection: users are rarely involved in validating or interpreting predictions derived from their own physiological data. Thieme et al. [79] argued that predictive accuracy alone is insufficient without attention to interpretability, user experience, and clinical applicability. Sanches et al. [62] called for a move beyond detection-oriented systems toward designs that support user sensemaking and self-regulation. A parallel line of work has emphasized participatory and context-sensitive approaches. Stefanidi et al. [75] found that individuals and families are largely absent from the design of care technologies, and Majid et al. [47] argued for sustained user involvement across the full design and research cycle.

Inclusive Design [16] provides an additional foundation for this shift from standardized to personalized systems. Rather than treating user variability as an edge case to be accommodated after deployment, inclusive design positions variability in ability, context, values, and access needs as a primary design condition. This perspective is especially important in digital mental health, because psychological experience is shaped by culture, clinical status, age, stigma, socioeconomic position, and comfort with data sharing. Inclusive Design therefore complements participatory design. Participatory methods involve users in shaping the system, while inclusive design broadens the design target from an assumed "average user" to a heterogeneous population whose needs may conflict or change over time [17]. For the present framework, this means that personalization should not only tune model parameters, but also allow users to shape what data are collected, how inferences are weighted, what explanations are meaningful, and when support should involve other people.

Despite this growing consensus, two limitations persist. First, while prior work has identified the need for personalization and user involvement, few studies have examined how these principles are operationalized at the system level. It remains unclear what specific architectural and interaction-level commitments move user agency from aspiration to implementation. Second, the relationship between data modalities remains undertheorized. Systems that combine physiological, behavioral, and self-report data typically aim to maximize predictive agreement across signals. The possibility that *divergence* between modalities is itself a meaningful signal has received limited attention. Such discrepancies may warrant user interpretation rather than algorithmic smoothing.

Positioning relative to prior work. The framework we develop builds on, but is distinct from, several established HCI traditions. Sensemaking research [43, 84] and reflective design [67] establish that ambiguity can support interpretation [27]. They do not, however, specify how multimodal mental health evidence should be handled when subjective, behavioral, and physiological signals conflict. Participatory design and Inclusive Design establish the importance of involving diverse users, but do not by themselves define where user authority should enter an AI inference pipeline. Human-in-the-loop (HITL) and participatory AI approaches [4, 21, 65, 71] often focus on improving models through user feedback. By contrast, our framework focuses on governing the *interpretation and actuation* of mental health inferences. Explainable AI exposes model reasoning [3, 40, 44], but explanation alone does not give users authority to confirm, modify, reject, or restrict the consequences of an inference. The framework therefore contributes a domain-specific account of user agency in digital mental health: cross-modal divergence is treated as a design resource, and user authority is operationalized through veto checkpoints, situated evidence weighting, and graceful degradation under changing data access. These constructs are shaped by the distinctive properties of mental health data, such as its subjectivity, its context-dependence, and the ethical risks that misclassification poses to vulnerable populations [32]. Misclassification in this domain is not only an accuracy problem; it is a potential harm, and the framework addresses it structurally rather than through interface affordances alone.

3 Approach: Purposive Synthesis

This paper adopts a design-oriented conceptual synthesis, drawing on representative work across interactive systems, affective computing, and digital health to identify recurring assumptions and structural design tensions. The approach is suited to papers that develop and evaluate conceptual contributions such as frameworks, constructs, and design principles. It does not aim to produce statistical generalizations about the literature. It follows established precedents for critical perspective and vision papers in HCI [85].

3.1 Selection Strategy

Our search combined three conceptual dimensions: (1) HCI and interactive systems ("human-computer interaction," "interactive system"); (2) mental health and emotional well-being ("stress," "anxiety," "well-being"); and (3) self-management technologies ("wearable," "affective computing," "multimodal," "machine learning," "large language model"). We searched ACM Digital Library and IEEE Xplore as primary sources, complemented by JMIR and PubMed for coverage of digital mental health interventions and clinically grounded constructs. Searches were limited to English-language publications from 2015–2025 and were completed in March 2026.

3.2 Inclusion Criteria

Publications were selected if they: (1) focused on individual differences or personalized mental health support; (2) explicitly targeted mental health or emotional well-being; (3) addressed self-management, including self-tracking, reflection, or regulation; and (4) provided technical or interaction-level contributions enabling analysis of sensing, modeling, or feedback design. We excluded purely clinical or social studies without HCI relevance, papers unrelated to mental health self-care, and purely technical machine learning or signal-processing papers without interaction-level contributions.

3.3 Corpus and Analytical Lens

After iterative title/abstract screening and full-text review of 108 candidates, 23 papers met our inclusion criteria. One additional paper was identified through backward citation search, yielding a final corpus of 24 studies. The corpus is purposive rather than exhaustive: its value lies in theoretical coverage of the design space, not in statistical representativeness. We assess the corpus by theoretical coverage rather than statistical exhaustiveness. The four clusters in Table 2 identify recurring patterns across domains, target populations, methods, and data sources. The inclusion of qualitative studies, participatory design work, intervention studies, reviews, and adaptive modeling papers allows the framework to integrate interactional, social, and computational perspectives.

We analyzed the corpus through two analytical lenses. The first, *divergence-as-signal*, examines how systems handle inconsistencies across data modalities and whether such divergence is surfaced or used for reflection and adaptation. The second, *user agency*, examines the extent to which systems support user interpretation, control, and sensemaking rather than fully automated inference. Both lenses were developed iteratively in dialogue with the emerging literature and are formalized in the framework sections that follow.

We acknowledge that a purposive synthesis oriented around pre-specified analytical constructs carries an inherent risk of confirmation bias. Studies that operationalize divergence-as-signal and user agency well are more likely to be selected and foregrounded than those that do not. Rather than making statistical claims, we focus on identifying constructs that are already present but unarticulated in recent research. We demonstrate that these ideas are latent in the field, and provide a principled vocabulary that makes them explicit. Readers should interpret the synthesis as a theoretical argument illustrated by representative examples, not as an audit of the field. The framework should likewise be treated as a design vocabulary for future empirical work, not as evidence that these mechanisms have already been validated.

4 Synthesis: Four Design Tensions in Current Practice

Our analysis of the 24 reviewed studies revealed four recurring design tensions (See Table 2). Seven studies foregrounded self-tracking and interpretive agency. Five addressed temporal dynamics and changing user states. Seven situated mental health within broader socio-technical care contexts. Five examined adaptive modeling approaches that accommodate individual variability. These clusters are not mutually exclusive, but they provide an organizing structure for understanding how recent work moves away from passive detection toward user-centered interpretation.

4.1 Data as Subjective Artifact: Self-Tracking and Interpretive Agency

Seven of the reviewed studies primarily addressed the tension between treating mental health data as passively extractable objective signals and treating them as subjective artifacts that require user interpretation. These works reconceptualize self-tracked data as subjective artifacts that acquire meaning through user interpretation. Individual variability is positioned not as noise but as a resource for understanding lived experience.

Grimme et al. [28], studying women's health data practices, argue that difference should be understood as a resource rather than a failure, and advocate for designs that give users control over how their data are collected, interpreted, and shared. Severes et al. [68] reach a similar conclusion in the context of well-being self-tracking, framing wearable data as a tool for co-constructing personal well-being narratives rather than a medium for static monitoring. Majid et al. [46] examine self-tracking for bipolar disorder through Patient and Public Involvement methods and find that the complexity of lived experience demands deeply personalized systems with fine-grained control over data representation. Kelley et al. [38] demonstrate that generic tracking systems fail to meet the needs of university students and that designs must be tailored to specific life stages and developmental contexts.

Three additional studies extend this theme into intervention design. Cho et al. [14] identify highly individualized patterns, such as positive emotion-triggered behaviors, that cannot be captured by universal baselines, underscoring the need for systems that accommodate idiosyncratic user profiles. Kerber et al. [39] compare a self-guided intervention with passive tracking and report stronger outcomes on measures such as patient empowerment and mental health, suggesting that effectiveness in this literature is often operationalized through autonomy, engagement, and self-management capacity rather than sensing accuracy alone. Terp et al. [78], working in schizophrenia care, show that participatory design can support perceived empowerment, everyday manageability, and collaborative decision-making. In this context, "effective" should not be understood only as clinical symptom reduction or prediction accuracy, but as the extent to which a system supports users' capacity to interpret, negotiate, and act within everyday care practices.

4.2 Temporal Dynamics: Shifting User States and Contextual Adaptation

Five studies emphasized the instability of user states, needs, and coping strategies. These works demonstrate that static user profiles are insufficient for mental health self-management, because psychological states shift across time and context in ways that demand adaptive system responses.

Kornfield et al. [42] find that content generated by users during stable periods is more effective as an intervention resource than generic system prompts. By drawing on users' own past reflections, designs preserve personal authority and continuity. Burgess et al. [10] propose modular tool assemblages that users can dynamically reconfigure in response to evolving contexts. Hay et al. [30] frame wearable technologies not as fixed monitors but as resources for navigating changing health conditions. Majid et al. [47] and Wagener et al. [81] converge on the same finding: mental health self-management systems require continuous, context-sensitive interaction rather than fixed user representations.

4.3 Socio-Technical Context: Mental Health Beyond the Isolated User

Seven studies situated mental health self-management within broader socio-technical contexts, such as family, workplace, clinical, or cultural ecosystems, challenging approaches that treat the user as an isolated unit of analysis.

Stefanidi et al. [75] argue for co-design across the entire care ecosystem, including children, families, educators, and clinicians, rather than focusing on the individual in isolation. Slovak et al. [73] use technology probes to study emotional regulation within family contexts. Loke et al. [45] demonstrate that addressing emotional behavior in children requires therapist-guided, context-rich tools where meaning is constructed collaboratively. Burgess et al. [11] show that sociality is a fundamental component of depression self-management, with users actively curating their social interactions in response to emotional states. Jonathan et al. [36] and Choukou et al. [15] find that standalone digital interventions are insufficient without integration into clinical workflows. Haque et al. [29] demonstrate that culturally grounded, value-sensitive design is essential in low- and middle-income country contexts.

4.4 Adaptive Modeling: Accommodating Individual Variability in Computational Systems

Five studies addressed computational adaptation to individual variability. Saud et al. [65] demonstrate that HITL approaches enable users to iteratively label and reinterpret their own sensor data, allowing models to adapt to personal baselines rather than enforce population-level norms. Jain et al. [35] show that multimodal systems can capture context-dependent signals. For example, in sarcasm, the divergence across modalities carries significant semantic weight. These nuances would be lost under traditional agreement-maximizing fusion strategies. At the application level, Mendes et al. [48] and Gautam and Sreekumari [26] design conversational agents that personalize interactions by responding to user-specific behavioral and emotional patterns rather than generic templates. De Souza et al. [20] argue that mental health chatbots must account for individualized manifestations of illness, as standardized response strategies risk misalignment with users' lived experiences.

Across these studies, a pattern emerges of systems moving toward what we characterize as a *coherence-aware paradigm*: one that prioritizes interpreting cross-signal divergence and user-specific variation as informative, rather than enforcing uniformity through statistical averaging. The following section formalizes this observation into a design framework.

5 Toward a User-Agency-Centered Framework

Rather than treating individual variability as noise, our synthesis suggests that it is a critical resource for system design. This requires moving beyond agreement-maximizing models to embrace the situated, temporal, and social contexts that define the user experience. What is currently missing is a unifying framework that integrates these insights across system modalities and interaction layers. This section provides that framework, organized around two original constructs and summarized in Figure 3. Together, these components reposition digital mental health systems from pipelines that resolve uncertainty toward systems that surface, preserve, and support the interpretation of uncertainty.

The framework follows from the four tensions identified in the synthesis. If mental health data are interpretive artifacts rather than objective readouts, then systems must preserve users' ability to contextualize evidence. If user states change over time, then systems must support revisable interpretations rather than fixed profiles. If mental health is embedded in socio-technical contexts, then systems must clarify when data or interpretations are shared and who is expected to act on them. If computational systems must accommodate individual variability, then divergence across signals should be treated as potentially informative rather than automatically resolved. The complementary evidence chain and the user-agency-centered architecture translate these four observations into design constructs.

5.1 Construct 1: The Complementary Evidence Chain

The limitations outlined above point to a need for a new conceptual framing. We propose the *complementary evidence chain*:

> **Definition.** Mental health status is not a single variable but a latent construct composed of multiple heterogeneous signals, requiring inference through complementary evidence chains in which no single modality constitutes ground truth.

A complementary evidence chain conceptualizes mental health as a latent state that cannot be directly observed but must be inferred through the alignment of multiple partial signals (See Table 1). These signals fall into three broad categories [25]. *Subjective evidence* includes self-reports, interviews, and conversational data. It offers semantic richness and interpretability but is subject to bias and discontinuity [7]. *Behavioral evidence*, derived from digital traces such as activity patterns, communication behavior, and language use, provides continuous and ecologically valid observations. It often lacks clear semantic grounding [63]. *Physiological evidence*, including HRV, EDA, and other biosignals, offers bodily-state measurements but is typically context-free and difficult to interpret in isolation [66].

No single modality serves as a reliable proxy for mental health. Meaningful understanding emerges not from privileging one modality over others, but from examining how these sources relate to and constrain one another (See Figure 1). The complementary evidence chain does not imply that all evidence types should be treated equally in every context. The absence of a single ground truth does not mean that all signals have identical reliability, relevance, or ethical weight. A self-report may be more meaningful when the design goal is reflection, while physiological or behavioral signals may be more useful for detecting routine disruption or acute change. Evidence weighting should therefore be treated as a situated design problem, shaped by signal quality, temporal relevance, user trust, domain risk, and the user's own values. In some cases, the appropriate response is not to compute a fixed weighted average, but to surface the weighting dilemma itself: for example, "*Your sleep and activity patterns suggest strain, but your self-report does not. Which source feels more representative today?*" This preserves the

Table 1: Evidence Types in the Complementary Evidence Chain

Evidence Type	Data Source	Characteristics
Subjective	Questionnaires, self-reports, dialogues	Interpretable but subject to bias
Behavioral	Mobile phone usage, interaction logs, voice/text	Continuous but semantically weak
Physiological	Heart rate, skin conductance, eye movement	Measurement-precise but semantically ambiguous

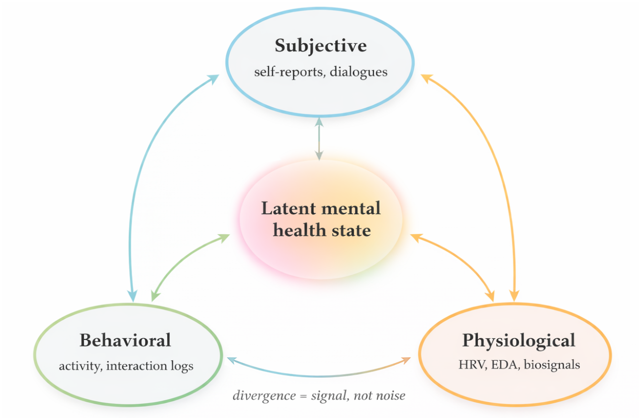

Figure 1: The complementary evidence chain. Mental health is inferred from relationships among subjective, behavioral, and physiological evidence. These evidence types are complementary but not automatically equal: their relevance and weight depend on signal quality, context, user values, and trust. Divergence between modalities is treated as a meaningful signal for reflection rather than noise to be smoothed algorithmically.

core argument of divergence-as-signal while avoiding the misleading assumption that all modalities are interchangeable or equally authoritative.

This framing has a specific computational implication. *Coherence*, within this framework, is not statistical agreement among heterogeneous signals. It is the algorithmic mapping of their relational divergence [51]. When a system detects friction, for instance, elevated physiological arousal contrasting with a benign self-report, that divergence is not noise to be smoothed over. A coherence-aware framework treats this friction as a primary signal, mapping the tension between an individual's lived experience and their physiological metrics. Rather than forcing a definitive state classification, the system uses this tension to prompt user reflection [23]. A worked instantiation of this logic appears in §5.3, Scenario 1, where divergence between physiological arousal, low mobility, and a calm self-report is surfaced to the user rather than collapsed into a single stress score.

This reframes the evolution of computational methods in digital mental health. Early single-modality classifiers [41, 74] gave way to multimodal fusion techniques [6], which were in turn extended by temporal modeling that treats intra-individual deviation as more informative than population-level averages [1]. The emergence of large language models introduces capacity to integrate low-level signals with high-level semantic interpretation [86]. Yet current systems remain constrained by data-centric assumptions. Models are evaluated on predictive accuracy rather than interpretive support [61], integration is fragmented, and users are rarely involved in validating inferences [47].

From an interaction design perspective, the complementary evidence chain is not only a methodological framework but a design-oriented concept. It foregrounds the need for systems that make relationships among forms of evidence visible and interpretable to users. Such visibility enables active participation in constructing one's own mental health understanding. Rather than positioning users as passive recipients of model outputs, such systems should support reflection, sensemaking, and negotiation among multiple information sources [30].

This framing carries three implications for system design. First, uncertainty should be explicitly represented rather than hidden, enabling users to engage critically with system outputs. Second, data must be interpreted in relation to users' daily life contexts. Third, mechanisms should be included that empower users to provide situated interpretations of their data. This approach grants users epistemic agency, integrating their subjective lived experience as a core component of the evidence chain. If the complementary evidence chain reframes mental health understanding as user-governed interpretation of partial signals, then the system architecture must operationalize that governance. Section 5.2 specifies the architectural commitments required.

5.2 Construct 2: User-Agency-Centered System Architecture

Building on this outcome-oriented framing, a user-agency-centered architecture should make explicit which type of outcome it is designed to support and how that outcome will be assessed. Recent HCI work has reframed digital mental health systems as interactive, ongoing services rather than static diagnostic tools [62, 75]. This framing reflects a broader recognition that mental health is a dynamic process involving self-awareness, interpretation, decision-making, and behavior change [47]. The role of computational systems is shifting from classifier to supportive partner embedded in everyday life.

This shift is tied to a redefinition of system-level goals. Rather than optimizing solely for predictive performance, contemporary systems are increasingly evaluated on their capacity to enhance *accessibility* for underserved populations [34], to support *user agency* over data [9], to preserve *privacy* [83], to enable *stepped care* [69], and to provide scalable *non-clinical support* [72]. Among these dimensions, *user agency* has emerged as a central organizing principle. It is visible in three design transformations: systems that provide interpretable feedback rather than opaque outputs [40]; user control over data collection, sharing, and usage, often described as *data sovereignty* [9]; and self-guided interventions in which users engage with therapeutic content at their own discretion [24]. Burgess et al. [11] provide supporting evidence by showing that users actively select communication channels based on emotional state, reinforcing the case for designs that support autonomy in when, how, and with whom users engage. These transformations align with emerging HCI design practices: participatory design [52], conversational interfaces powered by LLMs [49], context-aware embedded interventions, and longitudinal personalization techniques [77].

These developments signal an architectural shift: from a linear pipeline of data → model → prediction toward closed-loop, interactive processes (See Figure 2). This architecture operationalizes user agency through two mechanisms.

Algorithmic veto power. In Loop 1, users can contextualize, label, or reject physiological signals before they are transferred into the AI inference space. Similarly, AI-generated health plans should not automatically trigger interventions; they require user confirmation before becoming actionable. In this architecture, users are not endpoints of inference but governing participants in the loop. Coherence in the complementary evidence chain is therefore not a static calculation, but an ongoing negotiation between AI inference and user interpretation.

Three boundaries define algorithmic veto power. First, veto power applies to interpretation and actuation, not to data collection itself: users may reject an inference while the underlying sensor data remain available for longitudinal pattern recognition. Second, veto mechanisms must be tiered and lightweight. Low-stakes feedback may require no confirmation, significant recommendations may warrant brief confirmation, and irreversible actions should require explicit consent. Third, veto power is limited in safety-critical scenarios. However, escalation should not assume continuous clinical monitoring. Systems should distinguish user-facing crisis resources, pre-agreed emergency or care-team notifications, and clinician review only when such monitoring has been explicitly arranged. During onboarding and crisis interactions, systems should state who can see shared data, whether review is real-time or asynchronous, and what response time users should expect.

A graduated fail-safe protocol can operate independently of ordinary veto. Ambiguous signals should first trigger an explicit user check-in. If the user does not respond or risk estimates increase, the system should offer crisis resources and clarify who, if anyone, can see the data. If risk crosses a safety-critical threshold, the system should follow the pre-agreed escalation pathway rather than rely on user confirmation. All transitions should be logged for later user review.

Graceful degradation of inference. Users must have the authority to dynamically modulate their data fidelity without causing the system to fail, for instance, revoking access to continuous physiological tracking during periods of high vulnerability [53, 70]. When a user exercises this data sovereignty, the system should adapt to the resulting gaps rather than collapse or resort to coercive tactics such as locking features until permissions are restored. By leaning on remaining links in the complementary evidence chain, such as subjective self-reports or low-friction conversational inputs, the system transparently downshifts its analytical scope. Privacy, in this model, is not penalized with a loss of service but integrated as a functional, dynamic parameter of ongoing mental health management.

To concretize this architecture, Hay et al. [30] present the SHAW application, which replaces passive employee health tracking with a conversational interface for holistic reflection. The system enforces algorithmic veto power by mandating user oversight of AI health indexes. This ensures that workplace adjustments are preceded by personal reflection, thereby institutionalizing structural accountability.

Because this paper does not present a deployed system, the framework should be read as an implementation scaffold rather than a software architecture (See Table 3). To make the constructs actionable for HCI and healthcare practitioners, Table 3 translates each construct into operational design questions, possible implementation strategies, user-facing controls, and expected system behavior. These details are not intended to prescribe one technical solution, but to clarify where design decisions must be made when moving from conceptual framework to system development.

5.3 Illustrative Use Scenarios

Scenario 1: Cross-modal divergence in student stress self-management. A university student wears a smartwatch and occasionally completes short self-reports. The system detects elevated HRV-related arousal and reduced mobility, but the student reports feeling calm. Rather than classifying the student as stressed, the system presents the discrepancy: "*Your body shows signs of activation, while your self-report suggests calm. Is there a context that explains this?*" The student identifies that they had caffeine before an exam but do not feel emotionally distressed. The system records this interpretation and avoids over-pathologizing similar patterns in the future. The prompt targets the "blind spots" of digital phenotyping. By inviting reflection on physical, environmental, and emotional context, it captures factors that individual sensors are structurally unable to detect. The divergence is not a measurement error; it is a window into the gap between physiological signal and subjective experience, and the system makes that gap visible rather than collapsing it.

Scenario 2: Algorithmic veto in workplace well-being. An employee-facing well-being system infers possible burnout risk from sleep disruption and work communication patterns. Before suggesting workplace adjustments or sharing any summary with a manager or clinician, the system asks the user to confirm, modify, or reject the interpretation. The user explains that the pattern reflects temporary caregiving responsibilities rather than work-related burnout. The system updates its interpretation and offers private coping resources without escalating the inference.

Scenario 3: Graceful degradation under privacy restriction. A user temporarily disables physiological sensing during a vulnerable period. Instead of locking features or reducing support quality without explanation, the system transparently shifts to lighter-weight self-report and conversational check-ins. It explains that recommendations are now based on fewer evidence sources, and invites the user to decide whether and when to restore sensing.

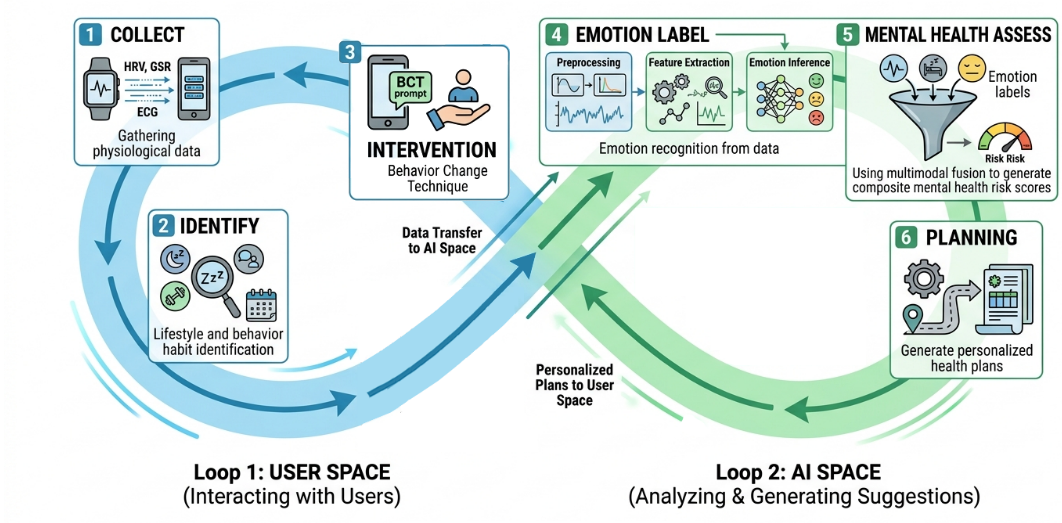

Figure 2: The closed-loop interactive process operationalizing user agency.

6 Discussion

6.1 Challenges

Despite rapid advances in sensing technologies, machine learning, and interactive systems, significant challenges remain in realizing effective and responsible digital mental health systems. These challenges are not merely technical. They concern how such systems are interpreted, trusted, and integrated into users' everyday lives.

Model complexity versus interpretability. Multimodal and large-scale models have improved predictive capabilities, but they often operate as opaque systems offering limited insight into how inferences are generated. In mental health contexts, where outputs influence users' self-perception and decision-making, this opacity can undermine trust and lead to harmful misinterpretation [40]. The challenge calls for approaches that make system reasoning accessible and meaningful to users, rather than approaches that optimize merely for accuracy.

Privacy and data agency. Mental health data, such as physiological signals, behavioral traces, and conversational content, are highly sensitive and deeply personal. The continuous and passive nature of data collection raises concerns about informed consent, data ownership, and secondary use [13]. Designing for privacy in this domain requires more than technical safeguards. It demands transparent mechanisms that allow users to understand, control, and negotiate how their data are collected and used. Graceful degradation of inference provides a structural answer to this tension, but its implementation in real-world systems remains an open design problem.

Individual and cultural variability. Mental health experiences are shaped by diverse personal, social, and cultural contexts, yet existing systems rely on generalized models that may not reflect this diversity [64]. Emotional constructs and their expression vary across populations, raising concerns about the universality of model assumptions and the risk of bias in both data and design [8]. Addressing this requires improved personalization techniques alongside participatory approaches that involve users in shaping how systems represent and respond to their experiences.

Long-term engagement and real-world effectiveness. Many digital mental health tools demonstrate promising results in controlled studies, but sustaining user engagement over time remains difficult. Improvements in predictive performance do not necessarily translate into meaningful behavioral or clinical outcomes [50]. This points to a need for evaluation frameworks that move beyond model-centric metrics toward measures of user experience, behavioral change, and well-being in naturalistic settings.

Balancing self-guided support with clinical integration. User-agency-centered systems emphasize autonomy and accessibility. They may not suffice, however, for individuals experiencing severe or acute conditions. Designing effective pathways that connect self-guided tools with professional care remains an open challenge. Systems must provide appropriate guidance without overstepping their scope — a boundary that is easy to state but difficult to operationalize [5].

Toward evaluation: specifying meaningful outcomes. A recurring limitation of conceptual frameworks in HCI is the absence of a clear path from principle to evidence. We propose three evaluation strategies appropriate to the constructs developed here. First, the complementary evidence chain can be evaluated through controlled comparison studies. Prototype systems that surface cross-modal divergence to users can be compared with agreement-maximizing baselines on measures of self-understanding, trust calibration, and reflection quality. Second, algorithmic veto power can be evaluated through diary or experience-sampling studies that track how often users exercise veto options, in what contexts, and with what downstream effects on engagement and perceived autonomy. Third, graceful degradation can be evaluated through data-scarcity simulation: progressively restricting available modalities and measuring whether the system's responses remain useful and whether users notice or object to the degradation. Together, these strategies would move the framework from conceptual proposal to empirically grounded design guidance. Figure 3 maps the framework's three layers (evidence inputs, inference, user agency); each evaluation strategy targets one layer.

Sensemaking, explainability, and the role of GenAI. A promising direction for implementing this framework lies at the intersection of sensemaking and GenAI. The primary objective of divergence-as-signal is to provoke sensemaking [84]. By presenting users with ambiguous data rather than a unified inference, the system creates the conditions for them to construct their own situated meaning. GenAI is well-suited to support this process because it can translate low-level sensor patterns into natural-language prompts that invite reflection rather than deliver verdicts [87]. The same fluency that makes LLMs effective communicators, however, also makes them susceptible to confident overstatement [83]. A system that frames an uncertain inference as a clear conclusion violates the epistemic commitments of the complementary evidence chain. The design challenge is not to make systems more conversational, but to ensure that their expressiveness is calibrated to the genuine uncertainty of the underlying evidence. To enforce this boundary, the LLM's reasoning process must be constrained through targeted prompting strategies and modular architecture. Rather than feeding the LLM raw data and asking it to interpret the user's state, the system must only supply the LLM with the *delta*, which is the mathematical or categorical gap between the two data streams. This constraint locates the LLM strictly within the inference layer of the framework (Figure 3); the agency layer remains the user's. By hardcoding the boundaries of automated processing to stop at the presentation of divergence, the system guarantees that the locus of interpretation remains with the user, turning algorithmic uncertainty into a catalyst for human insight.

6.2 Limitations

Several limitations bound the claims of this paper. As a conceptual synthesis, the framework has not been empirically validated. The constructs of the complementary evidence chain, divergence-as-signal, and algorithmic veto power are proposed as design principles grounded in existing evidence, not as empirically tested mechanisms. Their effects on user understanding, engagement, and clinical outcomes remain to be established. The paper also maintains a high-level focus on mental health self-management without distinguishing between general and clinical populations. Future research is needed to determine how these design implications shift when applied to specific care settings or high-severity contexts.

As a purposive synthesis, the corpus is also subject to selection bias. The two analytical lenses were developed prior to final corpus selection, increasing the likelihood that included studies support the argument rather than complicate it. Studies that defend data-centric inference, or that report negative findings from participatory design approaches, are likely underrepresented. The corpus draws predominantly from ACM and IEEE venues, skewing toward computer-science and HCI framings at the expense of clinical psychology, nursing informatics, and global health perspectives. Most reviewed work was conducted in high-income, English-speaking contexts with adult participants, constraining the generalizability of the proposed framework. These limitations do not invalidate the theoretical argument, but they establish clear boundaries on its current scope and point directly to the empirical validation work that must follow. We treat these limitations not as reasons to generalize cautiously but as design constraints. Because data-centric, clinical, and non-Western perspectives may be underrepresented, the framework emphasizes evidence weighting, monitoring expectations, and graceful degradation rather than assuming that user participation is always sufficient or universally desired.

6.3 Design Implications

The framework yields five implications for practitioners designing digital mental health systems (See Figure 3).

Surface divergence, do not suppress it. Systems should expose cross-modal inconsistencies as interpretive prompts rather than resolve them algorithmically. Designers should specify what counts as a meaningful conflict and determine whether it should trigger user reflection, an uncertainty disclosure, or care-related action.

Treat evidence weighting as user-shaped and context-dependent. Different data sources carry different weights depending on the goal. Whether the system relies on a user's input or a sensor's signal, the user should be able to see the reasoning and override it. These overrides should apply only to the current situation, not become a universal setting.

Implement veto checkpoints at inference boundaries. Veto mechanisms should be placed before consequential transitions: from sensing to labeling, from labeling to recommendation, from recommendation to notification, and from private interpretation to sharing or escalation. At each checkpoint, users should be able to confirm, modify, reject, defer, or annotate the inference.

Degrade gracefully under data scarcity. Designers should define the minimum viable evidence chain for each system function. When data access is restricted, the system should disclose which evidence is missing, lower the confidence or scope of recommendations, and continue offering lower-burden forms of support rather than fail silently.

Design for ecosystem, not just the individual. Systems should distinguish among private reflection, selective sharing, asynchronous review, and active monitoring. Interfaces should make clear who can see an interpretation, who is expected to act on it, and what response time users should expect.

7 Conclusion

This paper has argued that digital mental health systems should move from data-centric inference toward interaction-centered support. Through a purposive synthesis of 24 studies, we identified four recurring tensions: data as subjective artifact, temporal instability, socio-technical context, and adaptive modeling. We translated these tensions into a framework built on two constructs: the *complementary evidence chain*, from which the principle of *divergence-as-signal* follows, and a *user-agency-centered architecture* operationalized

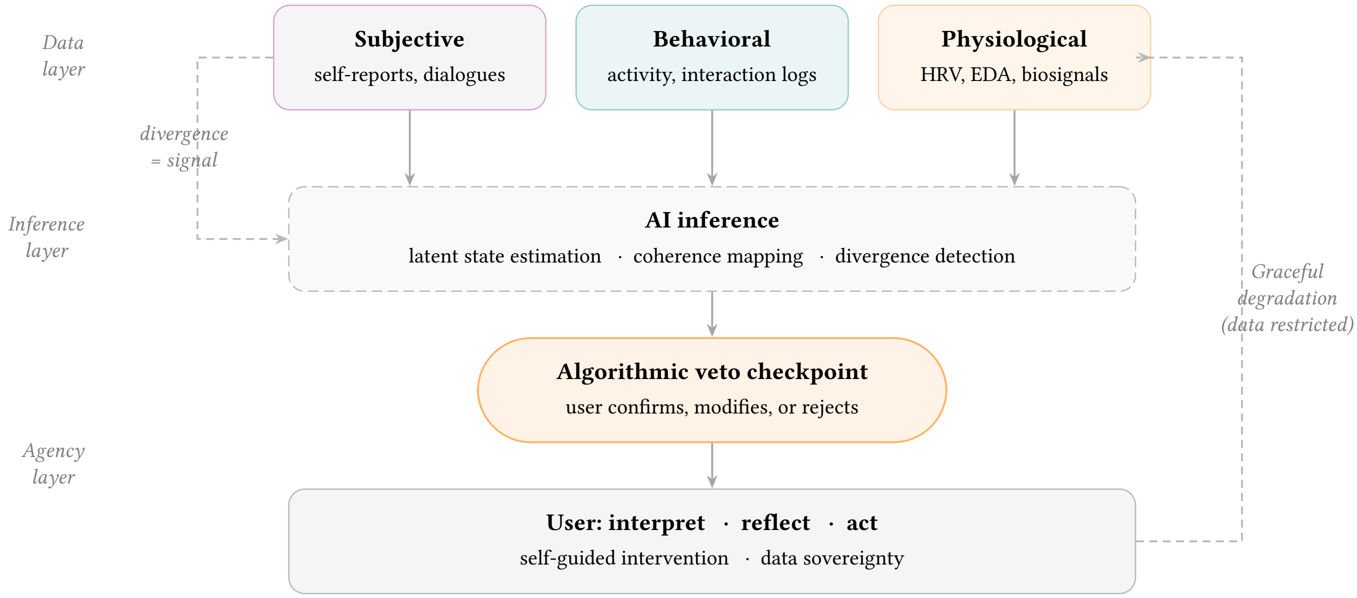

Figure 3: User-agency-centered framework. Evidence from three modalities feeds an AI inference layer (dashed border = computational and uncertain). The algorithmic veto checkpoint gives users authority to confirm, modify, or reject inferences before any system action is taken. The dashed return path illustrates graceful degradation: when data access is restricted, the system falls back to lighter-weight evidence rather than failing. Divergence across modalities (left annotation) is treated as a meaningful signal surfaced to the user, not smoothed away algorithmically.

through *algorithmic veto power* and *graceful degradation of inference*. The framework does not claim to solve mental health inference through better fusion alone. It reframes conflicting evidence as a design opportunity. Systems should help users interpret uncertain signals, contest automated inferences, and control how data are acted upon or shared. The value of digital mental health systems lies not only in their ability to detect patterns, but in their capacity to enable reflection, support decision-making, and empower individuals in everyday life [2, 4, 12, 88].

Future work should empirically test whether these mechanisms improve self-understanding, trust calibration, engagement, and appropriate care escalation in real-world deployments. Four directions are immediate. First, the complementary evidence chain requires prototype evaluations that explicitly surface cross-modal divergence and compare reflective outcomes against agreement-maximizing baselines. Second, algorithmic veto power needs operational specification through comparative analysis of veto mechanisms (tiered consent models, proactive confirmation dialogues), measuring effects on engagement and longitudinal data quality. Third, the reviewed studies disproportionately focus on adult populations in high-income, English-speaking contexts; extending this work to adolescents, older adults, and users in low- and middle-income settings is essential. Fourth, longitudinal in-the-wild deployments are needed to understand how user agency and system coherence evolve over weeks and months of real-world use.

References

[1] Z W Adams, E A Mcclure, K M Gray, C K Danielson, F A Treiber, and K J Ruggiero. 2017. Mobile devices for the remote acquisition of physiological and behavioral biomarkers in psychiatric clinical research. *Journal of psychiatric research* 85 (2017), 1–14. doi:10.1016/j.jpsychires.2016.10.019

[2] J.E. Allen, C.I. Guinn, and E. Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14, 5 (1999), 14–23. doi:10.1109/5254.796083

[3] Marco Almada. 2019. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on artificial intelligence and law*. 2–11. doi:10.1145/3322640.3326699

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[5] Luke Balcombe and Diego De Leo. 2022. Human-computer interaction in digital mental health. In *Informatics*, Vol. 9. MDPI, 14. doi:10.3390/informatics9010014

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443. doi:10.1109/tpami.2018.2798607

[7] Jakob E. Bardram, Mads Frost, Károly Szántó, and Gabriela Marcu. 2012. The MONARCA self-assessment system: a persuasive personal monitoring system for bipolar patients. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. Association for Computing Machinery, New York, NY, USA, 21–30. doi:10.1145/2110363.2110370

[8] L F Barrett. 2017. *How emotions are made: The secret life of the brain.* Pan Macmillan. doi:10.7202/1064926ar

[9] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D Mekler. 2023. How does HCI understand human agency and autonomy?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3544548.3580651

[10] Eleanor R Burgess, Sean A Munson, David C Mohr, and Madhu C Reddy. 2025. What's In Your Kit? Mental Health Technology Kits for Depression Self-Management. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19. doi:10.1145/3706598.3713585

[11] Eleanor R Burgess, Kathryn E Ringland, Jennifer Nicholas, Ashley A Knapp, Jordan Eschler, David C Mohr, and Madhu C Reddy. 2019. "I think people are powerful" The Sociality of Individuals Managing Depression. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–29. doi:10.1145/3359143

[12] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. doi:10.1145/3359206

[13] S Chancellor and M De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* 3, 1 (2020). doi:10.1038/s41746-020-0233-7

[14] Minseo Cho, Doeun Park, Myounglee Choo, Doug Hyun Han, and Jinwoo Kim. 2025. Adolescent self-reflection process through self-recording on multiple health metrics: qualitative study. *Journal of medical Internet research* 27 (2025), e62962. doi:10.2196/62962

[15] Mohamed-Amine Choukou, Funminiyi Olatoye, Reg Urbanowski, Maurizio Caon, and Caroline Monnin. 2023. Digital Health Technology to Support Health Care Professionals and Family Caregivers Caring for Patients With Cognitive Impairment: Scoping Review. *JMIR Mental Health* 10 (2023), e40330. doi:10.2196/40330

[16] P John Clarkson, Roger Coleman, Simeon Keates, and Cherie Lebbon. 2013. *Inclusive design: Design for the whole population.* Springer Science & Business Media.

[17] Roger Coleman, John Clarkson, and Julia Cassim. 2016. *Design for inclusivity: A practical guide to accessible, innovative and user-centred design.* crc Press.

[18] V P Cornet and R J Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics* 77 (2018), 120–132. doi:10.1016/j.jbi.2017.12.008

[19] Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* Association for Computing Machinery, New York, NY, USA, 353–369. doi:10.1145/2998181.2998220

[20] Paula Maia de Souza, Isabella da Costa Pires, Vivian Genaro Motti, Helena Medeiros Caseli, Jair Barbosa Neto, Larissa C Martini, and Vânia Paula de Almeida Neris. 2022. Design recommendations for chatbots to support people with depression. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems.* 1–11. doi:10.1145/3554364.3559119

[21] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17).* Association for Computing Machinery, New York, NY, USA, 278–288. doi:10.1145/3025453.3025739

[22] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200. doi:10.1080/02699939208411068

[23] Yara El-Tawil, Aneesha Sampath, and Emily Mower Provost. 2026. What You Feel is not What They See: on Predicting Self-Reported Emotion from Third-Party Observer Labels. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 16212–16216. doi:10.1109/ICASSP55912.2026.11463580

[24] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785. doi:10.2196/mental.7785

[25] Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J Oedegaard, and Jim Tørresen. 2018. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing* 51 (2018), 1–26. doi:10.1016/j.pmcj.2018.09.003

[26] Abikrit Gautam and Prasanthi Sreekumari. 2026. HopeBot: An AI-Powered Mental Health Chatbot Built to Support College Students. In *2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC).* 1–5. doi:10.1109/ICAIC67076.2026.11395838

[27] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03).* Association for Computing Machinery, New York, NY, USA, 233–240. doi:10.1145/642611.642653

[28] Sophie Grimme, Susanna Marie Spoerl, Susanne Boll, and Marion Koelle. 2024. My data, my choice, my insights: women's requirements when collecting, interpreting and sharing their personal health data. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* 1–18. doi:10.1145/3613904.3642851

[29] Munirul M Haque, Dipranjan Das Dipal, Masud Rabbani, Md Ishrak Islam Zarif, Anik Iqbal, Shaheen Akhter, Shahana Parveen, Mohammad Rasel, Basana Rani Muhuri, Tanjir Soron, and others. 2020. Towards developing a mobile-based care for children with autism spectrum disorder (mCARE) in low and middle-income countries (LMICs) like Bangladesh. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC).* IEEE, 746–753. doi:10.1109/compsac48688.2020.0-170

[30] Kiersten Hay, Larissa Pschetz, Billy Dixon, Belinda Steffan, Jakov Jandric, Thanasis Tsanas, Kate Sang, and Wendy Loretto. 2024. "Working it Out": Exploring How Digital Technologies Could Support Healthy Ageing at Work. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction.* 1–16. doi:10.1145/3679318.3685357

[31] Jennifer A Healey and Rosalind W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* 6, 2 (2005), 156–166. doi:10.1109/tits.2005.848368

[32] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 conference on designing interactive systems.* 95–99. doi:10.1145/3064663.3064703

[33] K Hoemann, F Xu, and L F Barrett. 2019. Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental psychology* 55, 9 (2019). doi:10.1037/dev0000686

[34] Angel Hsing-Chi Hwang, Dan Adler, Meir Friedenberg, and Qian Yang. 2024. Societal-scale human-AI interaction design? How hospitals and companies are integrating pervasive sensing into mental healthcare. In *Proceedings of the 2024 CHI conference on human factors in computing systems.* 1–16. doi:10.1145/3613904.3642793

[35] Jaishree Jain, Saroj Kushwaha, Updesh Kumar Jaiswal, Sunil Kumar, Eshank Jain, and Mani Dublish. 2025. Interactive Emotion Recognition System: AI-Driven Emotional Assistance. In *2025 5th International Conference on Advancement in Electronics & Communication Engineering (AECE).* IEEE, 561–564. doi:10.1109/aece67531.2025.11386645

[36] Geneva K. Jonathan, Cynthia A. Dopke, Tania Michaels, Andrew Bank, Clair R. Martin, Krina Adhikari, Rachel L. Krakauer, Chloe Ryan, Alyssa McBride, Pamela Babington, Ella Frauenhofer, Jamilah Silver, Courtney Capra, Melanie Simon, Mark Begale, David C. Mohr, and Evan H. Goulding. 2021. A Smartphone-Based Self-management Intervention for Bipolar Disorder (LiveWell): User-Centered Development Approach. *JMIR Mental Health* 8 (2021), e20424. doi:10.2196/20424

[37] Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. 2004. A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306, 5702 (2004), 1776–1780. doi:10.1126/science.1103572

[38] Christina Kelley, Bongshin Lee, and Lauren Wilcox. 2017. Self-tracking for Mental Wellness: Understanding Expert Perspectives and Student Experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).* Association for Computing Machinery, New York, NY, USA, 629–641. doi:10.1145/3025453.3025750

[39] André Kerber, Ina Beintner, Sebastian Burchert, and Christine Knaevelsrud. 2023. Effects of a Self-Guided Transdiagnostic Smartphone App on Patient Empowerment and Mental Health: Randomized Controlled Trial. *JMIR Mental Health* 10 (2023), e45068. doi:10.2196/45068

[40] P Klasnja, E B Hekler, S Shiffman, A Boruvka, D Almirall, A Tewari, and S A Murphy. 2015. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health psychology* 34, S (2015). doi:10.1037/hea0000305

[41] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31. doi:10.1109/t-affc.2011.15

[42] Rachel Kornfield, Renwen Zhang, Jennifer Nicholas, Stephen M Schueller, Scott A Cambo, David C Mohr, and Madhu Reddy. 2020. "Energy is a finite resource": designing technology to support individuals across fluctuating symptoms of depression. In *Proceedings of the 2020 CHI Conference on Human factors in Computing systems.* 1–17. doi:10.1145/3313831.3376309

[43] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-Centered Approaches to Fair and Responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20).* Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3334480.3375158

[44] Ian Li, Anind Dey, Jodi Forlizzi, Kristina Höök, and Yevgeniy Medynskiy. 2011. Personal informatics and HCI: design, theory, and social implications. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 2417–2420. doi:10.1145/1979742.1979573

[45] Lian Loke, Aaron Blishen, Carl Gray, and Naseem Ahmadpour. 2021. Safety, connection and reflection: Designing with therapists for children with serious emotional behaviour issues. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–17. doi:10.1145/3411764.3445178

[46] Shazmin Majid, Richard Morriss, Grazziela Figueredo, and Stuart Reeves. 2022. Exploring self-tracking practices for those with lived experience of bipolar disorder: Learning from combined principles of Patient and Public Involvement and HCI. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22).* Association for Computing Machinery, New York, NY, USA, 1907–1920. doi:10.1145/3532106.3533531

[47] Shazmin Majid, Stuart Reeves, Grazziela Figueredo, Susan Brown, Alexandra Lang, Matthew Moore, and Richard Morriss. 2021. The extent of user involvement in the design of self-tracking technology for bipolar disorder: literature review. *JMIR Mental Health* 8, 12 (2021), e27991. doi:10.2196/27991

[48] Carla Mendes, Rafael Pereira, Luis Frazao, José Cb Ribeiro, Nuno Rodrigues, Nuno Costa, Joao Barroso, and António MJ Pereira. 2024. Emotionally Intelligent

Customizable Conversational Agent for Elderly Care: Development and Impact of Chatto. In *Proceedings of the 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*. 208–214. doi:10.1145/3696593.3696619

[49] Adam S Miner, Arnold Milstein, and Jefferey T Hancock. 2017. Talking to machines about personal mental health problems. *Jama* 318, 13 (2017), 1217–1218. doi:10.1001/jama.2017.14151

[50] D C Mohr, M Zhang, and S M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 13 (2017), 23–47. doi:10.1146/annurev-clinpsy-032816-044949

[51] Ivan Moura, Ariel Teles, Davi Viana, Jean Marques, Luciano Coutinho, and Francisco Silva. 2023. Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: a systematic literature review. *Journal of biomedical informatics* 138 (2023), 104278. doi:10.1016/j.jbi.2022.104278

[52] Mariesa Nicholas, Penny Hagen, Kitty Rahilly, and Nathalie Swainston. 2012. Using participatory design methods to engage the uninterested. In *Proceedings of the 12th Participatory Design Conference: Exploratory Papers, Workshop Descriptions, Industry Cases-Volume 2*. 121–124. doi:10.1145/2348144.2348183

[53] Jakob Nielsen. 1994. *Usability engineering*. Morgan Kaufmann.

[54] World Health Organization. 2022. *WHO guideline on self-care interventions for health and well-being, 2022 revision*. World Health Organization.

[55] World Health Organization. 2022. *World mental health report: Transforming mental health for all*. World Health Organization.

[56] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[57] Anuja Pinge, Vinaya Gad, Dheryta Jaisighani, Surjya Ghosh, and Sougata Sen. 2024. Detection and monitoring of stress using wearables: a systematic review. *Frontiers in Computer Science* 6 (2024), 1478851. doi:10.3389/fcomp.2024.1478851

[58] M E Rodríguez-Rivas, A J Cangas, L A Cariola, J J Varela, and S Valdebenito. 2022. Innovative technology-based interventions to reduce stigma toward people with mental illness: systematic review and meta-analysis. *JMIR Serious Games* 10, 2 (2022). doi:10.2196/35099

[59] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161. doi:10.1037/h0077714

[60] S Saeb, M Zhang, C J Karr, S M Schueller, M E Corden, K P Kording, and D C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015). doi:10.2196/jmir.4273

[61] Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. 2024. Multimodal machine learning in mental health: a survey of data, algorithms, and challenges. *arXiv preprint arXiv:2407.16804* (2024).

[62] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, and others. 2019. HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–17. doi:10.1145/3290605.3300475

[63] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 671–676. doi:10.1109/acii.2013.117

[64] Z Sarsenbayeva, G Marini, N Van Berkel, C Luo, W Jiang, K Yang, and Goncalves. 2020. Does smartphone use drive our emotions or vice versa? A causal analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[65] Conrado Santos Saud, Vivian Motti, Luciano Neris, Tais Bleicher, and Vania Paula de Almeida Neris. 2024. Identification of possible depressive profile through mobile sensor data in university students. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*. 1–13. doi:10.1145/3702038.3702096

[66] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 400–408. doi:10.1145/3242969.3242985

[67] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph Kaye. 2005. Reflective Design. In *Proceedings of the 4th Decennial Conference on Critical Computing*. 49–58. doi:10.1145/1094562.1094569

[68] Beatriz Severes, Paulo Bala, Patricia Piedade, Eduardo Gomes, Augusto Esteves, and Mary Barreto. 2025. Making Sense of Our Data: Exploring Well-Being Self-Tracking Through Creative Collaboration. In *Proceedings of the 2025 Conference on Creativity and Cognition*. 182–197. doi:10.1145/3698061.3726950

[69] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–29. doi:10.1145/3613904.3642761

[70] Ben Shneiderman. 1987. Designing the user interface strategies for effective human-computer interaction. *SIGBIO Newsl.* 9, 1 (March 1987), 6. doi:10.1145/25065.950626

[71] Ben Shneiderman. 2022. Human-centered AI: ensuring human control while increasing automation. In *Proceedings of the 5th Workshop on Human Factors in Hypertext* (Barcelona, Spain) *(HUMAN '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 2 pages. doi:10.1145/3538882.3542790

[72] Petr Slovak and Sean A Munson. 2024. HCI contributions in mental health: A modular framework to guide psychosocial intervention design. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21. doi:10.1145/3613904.3642624

[73] Petr Slovák, Nikki Theofanopoulou, Alessia Cecchet, Peter Cottrell, Ferran Altarriba Bertran, Ella Dagan, Julian Childs, and Katherine Isbister. 2018. "I just let him cry... Designing Socio-Technical Interventions in Families to Prevent Mental Health Disorders. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–34. doi:10.1145/3274429

[74] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* 3, 1 (2011), 42–55. doi:10.1109/t-affc.2011.25

[75] Evropi Stefanidi, Johannes Schöning, Sebastian S Feger, Paul Marshall, Yvonne Rogers, and Jasmin Niess. 2022. Designing for care ecosystems: A literature review of technologies for children with ADHD. In *Proceedings of the 21st Annual ACM Interaction Design and Children Conference*. 13–25. doi:10.1145/3501712.3529746

[76] Arthur A Stone, Saul S Shiffman, and Marten W DeVries. 1999. Ecological momentary assessment. (1999).

[77] Norbury A. Perez-Rodriguez M. M. Olmos P. M. & Artés A. Sükei, E. 2021. Predicting emotional states using behavioral markers derived from passively sensed data: data-driven machine learning approach. *JMIR mHealth and uHealth* 9 (2021), e24465. doi:10.2196/24465

[78] Malene Terp, Rikke Jørgensen, Birgitte Schantz Laursen, Jan Mainz, and Charlotte D. Bjørnes. 2018. A Smartphone App to Foster Power in the Everyday Management of Living With Schizophrenia: Qualitative Analysis of Young Adults' Perspectives. *JMIR Mental Health* 5 (2018), e10157. doi:10.2196/10157

[79] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53. doi:10.1145/3398069

[80] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–50. doi:10.1145/3564752

[81] Nadine Wagener, Yvonne Rogers, and Jasmin Niess. 2025. The Self-Care Technology Process Model (SCTpm): A Framework for Designing VR-Based Self-Care Technology. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 2237–2254. doi:10.1145/3715336.3735743

[82] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 3–14. doi:10.1145/2632048.2632054

[83] Synthia Wang, Yuwei Cheng, Austin Song, Sarah Keedy, Marc Berman, and Nick Feamster. 2025. Can LLMs address mental health questions? A comparison with human therapists. *arXiv preprint arXiv:2509.12102* (2025).

[84] Karl E. Weick. 1995. *Sensemaking in Organizations*. Sage.

[85] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. 1–10. doi:10.1145/2601248.2601268

[86] Wenxuan Xu, Arvind Pillai, Subigya Nepal, Amanda C Collins, Daniel M Mackin, Michael V Heinz, Tess Z Griffin, Nicholas C Jacobson, and Andrew Campbell. 2025. LENS: LLM-Enabled Narrative Synthesis for Mental Health by Aligning Multimodal Sensing with Language Models. *arXiv preprint arXiv:2512.23025* (2025).

[87] Yuxin Xu, Mengqiu Cheng, and Anastasia Kuzminykh. 2024. What makes it mine? Exploring psychological ownership over human-AI co-creations. In *Proceedings of the 50th graphics Interface conference*. 1–8. doi:10.1145/3670947.3670974

[88] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376301

Table 2: Structured Overview of the 24 Reviewed Studies

Study cluster	Papers	Domain	Target users	Methods	Data sources	Analytical role
Self-tracking and interpretive agency	Grimme et al. [28]; Severes et al. [68]; Majid et al. [46]; Kelley et al. [38]; Cho et al. [14]; Kerber et al. [39]; Terp et al. [78]	self-tracking, well-being, bipolar disorder, schizophrenia, student mental health	women, students, people with lived experience, young adults	qualitative study, participatory design, RCT, interviews	self-report, personal health data, app use, wearable/self-tracking data	Establishes why mental health data should be treated as interpreted artifacts rather than objective signals
Temporal dynamics	Kornfield et al. [42]; Burgess et al. [10]; Hay et al. [30]; Majid et al. [28]; Wagener et al. [81]	depression, workplace health, VR self-care, bipolar disorder	people with depression, workers, self-care users	design study, qualitative inquiry, literature review, framework development	user reflections, technology kits, interaction logs, wearable/contextual data	Establishes why static user profiles and one-time labels are insufficient
Socio-technical context	Stefanidi et al. [75]; Slovak et al. [28]; Loke et al. [45]; Burgess et al. [11]; Jonathan et al. [36]; Choukou et al. [15]; Haque et al. [29]	ADHD, family mental health, child therapy, depression, bipolar disorder, cognitive impairment, autism care	children, families, therapists, caregivers, patients	co-design, probes, user-centered design, scoping review	interviews, care practices, family interaction, app data	Establishes why interpretation and action may involve caregivers, clinicians, workplaces, or families
Adaptive modeling	Saud et al. [65]; Jain et al. [35]; Mendes et al. [48]; Gautam & Sreekumari [26]; de Souza et al. [20]	sensor-based depression detection, emotion recognition, chatbots, elderly care, student mental health	students, elderly users, chatbot users	HITL modeling, system design, conversational agent design	mobile sensor data, multimodal emotion signals, dialogue data	Establishes why computational systems should adapt to personal baselines and divergent signals

Table 3: Operationalizing the framework

Framework construct	Operational design question	Possible implementation	User-facing control	System behavior
Complementary evidence chain	What evidence sources are available and how reliable are they in this context?	Maintain metadata for each modality: recency, quality, confidence, user trust, and contextual relevance	Let users inspect which signals contribute to an inference	Avoid presenting a single definitive mental-health label when evidence is weak or conflicting
Situated evidence weighting	Which evidence should matter more in this moment?	Use context-dependent weighting rather than fixed fusion; allow weights to shift by goal, risk, and user preference	Ask users which source feels more representative when signals diverge	Record user weighting preferences as contextual annotations, not universal rules
Divergence-as-signal	When do signals conflict meaningfully?	Detect discrepancies across subjective, behavioral, and physiological streams	Present divergence as a reflective prompt rather than a diagnosis	Trigger sensemaking prompts instead of smoothing discrepancies into one score
Algorithmic veto checkpoint	When should an inference require user confirmation?	Place checkpoints before notifications, recommendations, or sharing; define separate pre-agreed protocols for crisis escalation	Confirm, modify, reject, defer, or add context	Store the user response and adapt future inference thresholds or explanations
Graceful degradation	What happens when data are missing or restricted?	Define minimum viable evidence chains for each feature	Let users pause or restrict modalities without losing all support	Downshift to lower-confidence or lighter-weight support and disclose the limitation
Crisis escalation boundary	When can user veto be bypassed?	Use pre-agreed safety protocols and monitoring expectations	Explain during onboarding who monitors what and how quickly	Offer immediate resources, notify agreed contacts, or escalate only through predefined pathways