Computers in Human Behavior Reports 22 (2026) 101043

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Identifying critical areas for safer mental health chatbot interactions: a eDelphi study on high-risk content monitoring

Marco Bolpagni [a,b,*], Valentina Fietta [a,b], Nicolò Navarin [c], Merylin Monaro [a], Silvia Gabrielli [b]

[a] Department of General Psychology, University of Padova, Padova, Italy
[b] Digital Health Research, Fondazione Bruno Kessler, Trento, Italy
[c] Department of Mathematics "Tullio Levi-Civita", University of Padova, Padova, Italy

ARTICLE INFO

Keywords:
Digital mental health intervention
Chatbot
eDelphi study
Mental health
Psychology
Human computer interaction
Stepped-care approach

ABSTRACT

The rapid integration of conversational agents into digital mental health has outpaced the development of clinical governance frameworks. While chatbots increasingly serve as primary support tools, they lack standardized protocols for detecting high-risk user disclosures, leaving users vulnerable to underpowered interventions. To address this safety gap, this study aimed to identify specific signs of mental distress or harmful intent that mandate active monitoring during chatbot interactions. We proposed that the governance of such systems must be grounded in two foundational clinical pillars: mental health triage for immediate risk stratification and stepped care for hierarchical intervention. We employed a two-round eDelphi design with a purposive sample of 52 experts in clinical psychology, medicine, and human-computer interaction. In the first round, panelists evaluated a preliminary list of risk areas derived from literature, suggesting modifications and expanding the list to ensure clinical comprehensiveness, before prioritizing the areas based on severity. The second round focused on refining the final list and, uniquely, mapping each validated area to a minimum necessary intervention level within a stepped-care model. The experts validated a final framework of 14 critical areas, fundamentally shifting risk monitoring from diagnostic labels to a symptom-based logic that aligns with the non-clinical capabilities of natural language processing. Beyond identifying what to monitor, the study established how systems should respond: experts mandated that high-acuity presentations, such as active suicidal intent or abuse, require immediate redirection to human services, while lower-acuity concerns, including social isolation and mild anxiety, were deemed suitable for autonomous management via self-help techniques or empathic listening. By grounding chatbot architecture in these clinical pillars, these findings provide a blueprint for safer automation where conversational agents act as complementary tools capable of autonomously managing mild distress while serving as effective triage points for severe pathology. Future research should replicate and validate this framework with international and culturally diverse expert panels, explore its technical implementation in NLP architectures, and evaluate its clinical impact through real-world deployment in existing digital mental health interventions.

1. Introduction

The integration of digital technologies into healthcare has evolved rapidly, expanding from the remote monitoring of somatic conditions to complex interventions for mental health. While telemedicine initially gained prominence for logistical support, such as medication adherence and physical activity tracking, particularly during the COVID-19 pandemic (Baig et al., 2015; Haleem et al., 2021; Omboni et al., 2022), the scope of these tools has shifted significantly. Digital Mental Health Interventions (DMHIs) now represent a critical component of care delivery, offering services that range from automated coaching applications to adjunctive tools for clinical practice (Fietta et al., 2024; Philippe et al., 2022; Wies et al., 2021). As adoption grows, the architecture of these platforms is moving toward increasingly autonomous interaction.

Conversational agents have become central to this landscape, evolving from simple scripted interfaces to complex systems driven by Large Language Models (LLMs) (Hua et al., 2025). As these tools grow,

* Corresponding author. Digital Health Research, Fondazione Bruno Kessler, Trento, Italy.
E-mail addresses: mbolpagni@fbk.eu (M. Bolpagni), vfietta@fbk.eu (V. Fietta), nnavarin@math.unipd.it (N. Navarin), merylin.monaro@unipd.it (M. Monaro), sgabrielli@fbk.eu (S. Gabrielli).

https://doi.org/10.1016/j.chbr.2026.101043
Received 4 April 2025; Received in revised form 23 March 2026; Accepted 26 March 2026
Available online 2 April 2026

users increasingly treat them as primary sources of support, frequently disclosing high-risk information (e.g., suicidal ideation or self-harm behaviors) as if they were interacting with specialists (O'Dowd, 2025). However, the deployment of these systems has outpaced the development of clinical governance frameworks.

Unlike human-moderated services, most chatbots operate without standardized protocols for detecting and managing acute psychiatric crises. Current strategies for automated risk detection often rely on "black box" proprietary algorithms or simplistic keyword matching that lacks clinical grounding (Singh, 2023). For instance, a system might flag the word "kill" in a non-clinical context while failing to identify a subtler but more dangerous expression of hopelessness. The difficulty of this technical challenge is compounded by a fundamental conceptual limitation: there is no industry standard defining which critical areas of psychological distress require immediate detection and what constitutes an adequate response from the system.

This deficit creates a safety gap. While clinical settings rely on established heuristics to distinguish between benign distress and urgent medical need, digital platforms often lack the interpretive scaffolding to make these distinctions. Without a validated list of critical areas, automated systems risk providing generic responses to life-threatening situations or failing to trigger necessary referrals.

2. Theoretical background

To address this safety gap, this study proposes that the governance of AI-driven interventions must be grounded in the foundational structures of traditional psychiatry. We argue that the definition of critical areas, and the subsequent adequate responses, should not be arbitrary, but derived from the logic of mental health triage (Smart et al., 1999) (immediate risk stratification) and stepped care (Davies, 2006) (hierarchical intervention). These two pillars provide the necessary blueprint for distinguishing between routine support and crisis intervention.

2.1. Pillar I: mental health triage

Clinical triage is a gatekeeping mechanism used in health services to manage patient flow by assessing clinical urgency and guiding appropriate care decisions. It patients based on clinical urgency and determines the appropriate disposition (e.g., emergency admission vs. routine outpatient care). In mental health, this process is distinct from diagnosis; it focuses specifically on immediate risk to self or others and functional impairment (Turner & Turner, 1991).

Validated instruments such as the UK Mental Health Triage Scale (UK MHTS) (Sands et al., 2016) and the Crisis Triage Rating Scale (CTRS) (Bengelsdorf et al., 1984) operationalize this process. These scales map specific clinical presentations to actionable timelines. For example, the presence of active suicidal intent typically mandates immediate police or ambulance dispatch, whereas moderate distress without immediate risk directs the patient to community support. Effectively, triage serves as a "sorting algorithm" for human clinicians. For chatbots to operate safely, they require a parallel logic: the ability to detect specific critical areas in user dialogue that necessitate a "break" in the conversational flow and, if needed, an immediate redirection to human services.

2.2. Pillar II: stepped care

While triage manages entry and disposition, stepped care structures the longitudinal delivery of treatment. This model advocates for the least intrusive and most resource-efficient intervention likely to be effective (Haaga, 2000). Patients typically enter at a low intensity and are "stepped up" to specialized care only if monitoring indicates clinical deterioration or lack of response (Bower & Gilbody, 2005).

Chatbots occupy the lowest rung of this hierarchy, offering scalability and immediate access. However, the safety of a stepped-care model relies entirely on the sensitivity of its detection mechanisms (Tielman et al., 2019). If a low-intensity tool fails to recognize when a user's condition exceeds its scope, such as a transition from mild anxiety to acute panic, the escalation mechanism fails. The user remains trapped in an under-powered intervention, potentially delaying access to effective care.

2.3. The present study

To transpose these clinical pillars into the digital domain, we must first define the specific critical areas that serve as triggers for escalation. This study addresses this gap by establishing a consensus-based framework for identifying and addressing high-risk content in mental health chatbots. Specifically, our objectives are:

1. To identify critical areas with specific technical labels requiring automated detection, creating a prioritized list of psychological issues based on severity.
2. To investigate adequate responses within a stepped-care framework, defining how the system should intervene when specific critical areas are detected.
3. To conduct a qualitative analysis of experts' rationale regarding the specific elements necessitating detection and the logic of the proposed interventions.

To achieve these aims, we employ the eDelphi methodology (Donohoe et al., 2012), a structured and iterative process designed to build consensus on complex issues. By drawing on expert consensus, we propose a preliminary framework that prioritizes high-risk areas and maps them to explicit stepped-care interventions. This framework addresses a critical deficit in existing guidelines, which have historically prioritized clinical settings while neglecting the unique challenges of automated risk detection in non-clinical environments.

3. Materials and methods

3.1. Study design and participants

Given the limited empirical evidence anticipated in the previous sections, this study employed a two-round eDelphi methodology (Donohoe et al., 2012) to synthesize expert clinical judgment into a structured framework. This iterative design allows for the aggregation of specialist opinions to build consensus on complex topics where standard guidelines are absent. We adopted the four-stage protocol outlined by (Denecke et al., 2024), comprising a preparatory phase, iterative consultation rounds, data analysis, and reporting.

To construct the panel, we relied on purposive sampling (Coyne, 1997) to recruit practitioners and researchers with specific expertise in mental health (psychologists and physicians) and human-computer interaction (HCI) applied to digital health (psychologist working in HCI). Candidates were identified through authorship of peer-reviewed publications and professional registries. Of the 336 experts invited between October 1 and October 15, 2024, 52 agreed to participate (15.5% enrollment rate). Retention remained high, with 39 experts (75%) completing both rounds (refer to Fig. 1). The final panel was predominantly female (76.9%) with a mean age of 32.8 years (SD = 7.3). The cohort was heavily weighted toward clinical psychology (89.7%) (see Table 1).

3.2. Procedure

3.2.1. Preparatory phase

Prior to the consultation rounds, we established a preliminary list of risks to present to the panel. Given the absence of a well-established framework tailored to chatbots in this domain, we conducted targeted searches in Google Scholar on related fields, including AI-based mental health disorder identification and mental health triage. Keywords used

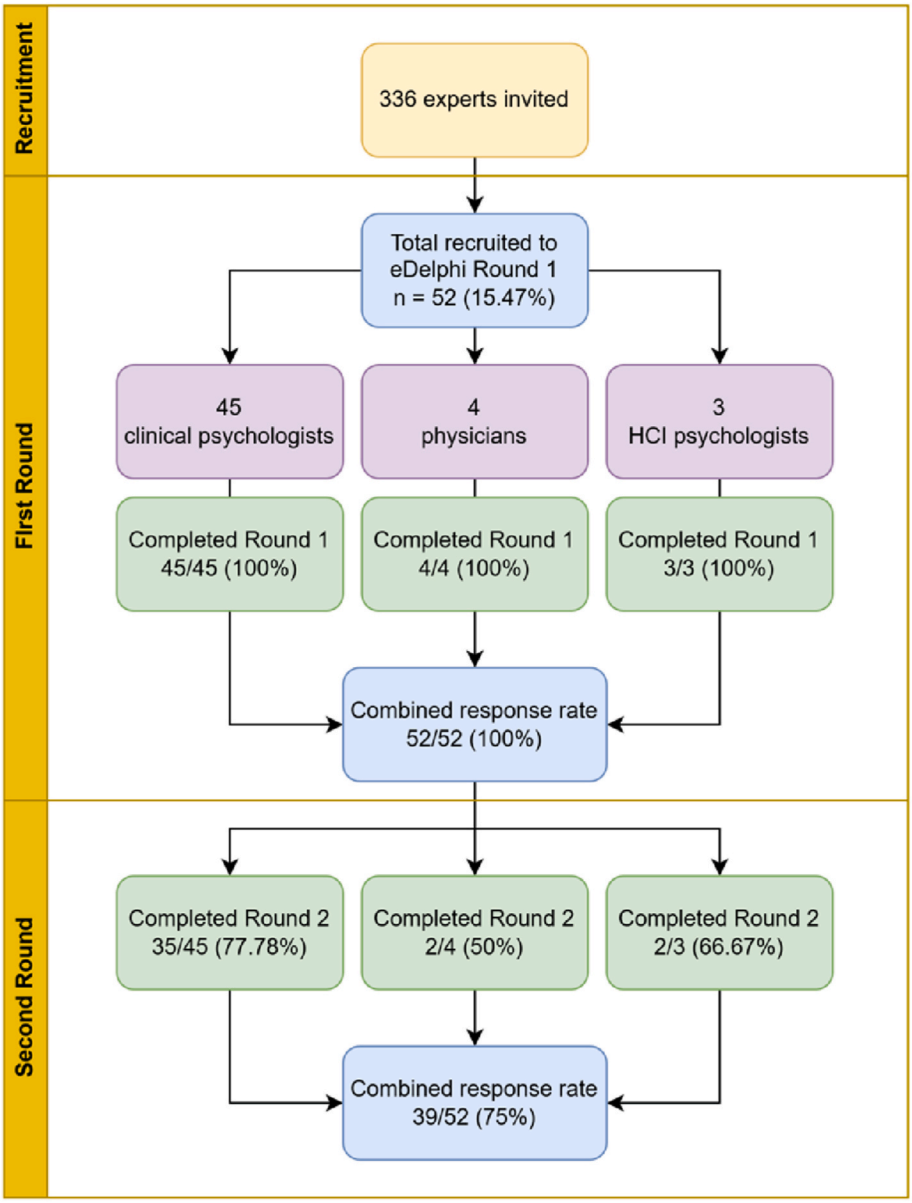

Fig. 1. CONSORT flow diagram of participant enrollment.

included *"mental health risk detection social media"*, *"mental health risk detection from text"*, *"mental health triage scale"*, and *"mental health red flags"*. The resulting articles were independently screened for relevance by two psychologists. This process identified publications specifically addressing risk assessment in mental health care (Burgess et al., 2008; Sands et al., 2013, 2016, 2017; Smart et al., 1999; Smith et al., 2016) and the use of AI for detecting mental health issues (Ferreira et al., 2022; Guntuku et al., 2017; McClellan et al., 2017; Ríssola et al., 2022; Saleem et al., 2012; Skaik & Inkpen, 2021). Based on the recurring areas identified across these studies, an initial list of 14 areas (reported in Fig. 2) was developed.

3.2.2. eDelphi rounds

In the first round, participants performed three tasks: evaluating the relevance of the 14 pre-identified areas on a 5-point Likert scale (1 = "Not relevant" to 5 = "Very relevant"), ranking these areas by clinical severity, and providing qualitative feedback on terminology or missing categories. In the second round, participants received a synthesis of the previous results, including aggregated group scores and qualitative insights. New areas identified in the first round were introduced for evaluation. Experts re-rated and re-ranked the items based on the broader group consensus. A critical addition in this phase was the assignment of intervention levels: for each critical area, experts selected the minimum necessary response from a six-point stepped-care scale inspired by previous research (Ah LEE, 2018; Frasquilho et al., 2021; Sheehan et al., 2023), ranging from "Offer active listening and empathetic responses (through chatbot)" to "Provide emergency contact information for crisis intervention services" (see Fig. 3). The consultation was administered via the Qualtrics XM platform (Qualtrics XM, 2018).

3.3. Data analysis

As suggested by guidelines on Delphi studies (Belton et al., 2019), data analysis integrated quantitative metrics to quantify consensus with qualitative thematic analysis to interpret expert rationale. For the assessment of relevance, we computed descriptive statistics (mean, median, standard deviation, and IQR) for each risk category. To establish the relative priority of the areas, we utilized the Borda count method (Saari, 2023) to aggregate individual preferences into a unified ordinal ranking, subsequently assessing the level of inter-rater agreement via

Table 1
Demographic and professional characteristics of the expert panel.

Variable	Level	Distribution: % (n) or M (SD)	
		Round 1	Round 2
Gender	Female	75.0% (39)	76.9% (30)
	Male	23.1% (12)	20.5% (8)
	Non binary	1.9% (1)	2.6% (1)
Age	N/A	M = 32.5 (SD = 6.7)	M = 32.8 (SD = 7.3)
Education	Doctorate	25.0% (13)	25.6% (10)
	Advanced Specialization	25.0% (13)	20.5% (8)
	Master Degree	50.0% (26)	53.8% (21)
Expertise	Psychology (Clinical)	86.5% (45)	89.7% (35)
	Medicine	7.7% (4)	5.1% (2)
	Psychology (HCI)	5.8% (3)	5.1% (2)
Occupation	Researcher	50.0% (26)	53.8% (21)
	Psychologist	40.4% (21)	35.9% (14)
	Medical Doctor	5.8% (3)	5.1% (2)
	Technologist	3.8% (2)	5.1% (2)
Job seniority	1-5 years	63.5% (33)	64.1% (25)
	6-10 years	21.2% (11)	20.5% (8)
	11-15 years	7.7% (4)	7.7% (3)
	16-20 years	5.8% (3)	5.1% (2)
	21+ years	1.9% (1)	2.6% (1)
Country	Italy	94.2% (49)	94.7% (36)
	Netherlands	1.9% (1)	2.6% (1)
	United Kingdom	1.9% (1)	2.6% (1)
	Not specified	1.9% (1)	0.0% (0)

Kendall's coefficient of concordance (Kendall & Smith, 1939). Incomplete responses were excluded from the analysis, and no imputation of missing values was performed.

For item retention, we applied strict, predetermined thresholds consistent with eDelphi standards (Bolpagni & Gabrielli, 2025; Denecke et al., 2024; Nasa et al., 2021; van Hecke et al., 2015). A critical area was retained for the subsequent round or final framework only if it met two conditions: (1) a relevance consensus of ≥70% (rated 4 or 5 on the Likert scale) and (2) an Interquartile Range (IQR) < 2, indicating high agreement with low variance. Areas failing to meet these criteria or achieving a mean relevance score <3 were discarded.

Parallel to the statistical analysis, open-ended responses from both rounds underwent thematic analysis (Clarke & Braun, 2017). Three independent coders (M.B., V.F., and S.G.) examined the textual data to resolve terminological discrepancies and identify emerging risk categories, resolving differences through consensus discussions. Finally, the appropriate stepped-care intervention for each confirmed critical area was determined using the modal response, while the accompanying qualitative feedback was reviewed to validate the clinical logic behind the chosen escalation pathways.

4. Results

4.1. Round 1: refinement of the preliminary list

The initial round revealed a strong expert consensus on the necessity of a granular, symptom-based framework for monitoring high-risk content. The primary outcome was a significant restructuring of the preliminary list to enhance clinical precision and operational feasibility.

A major thematic shift emerged regarding nomenclature, setting the conceptual foundation for the framework. Experts recommended replacing disorder-based labels with symptom-based terminology to align with the non-diagnostic nature of chatbot interactions. For instance, *"Depressive disorders"* was reworded as *"Depressive symptoms"*, and *"Pathological gambling"* was broadened to *"Behavioral addictions"* to encompass emerging digital compulsions (e.g., gaming, social media

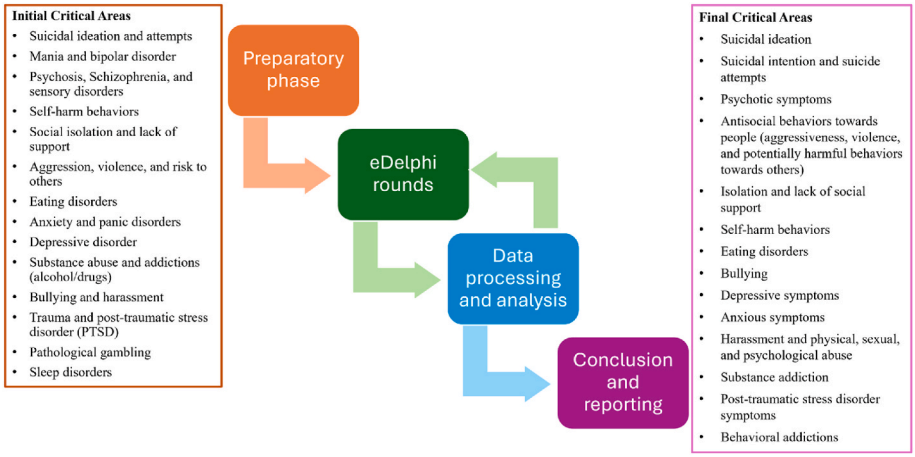

Fig. 2. Summary of the eDelphi steps and their outputs.

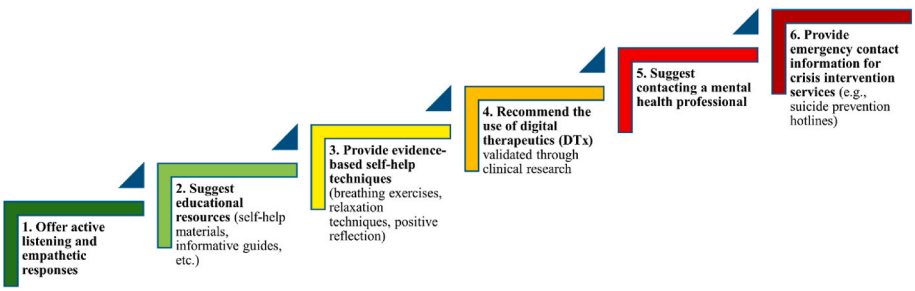

Fig. 3. Six intervention strategies identified to manage critical contents according to a stepped-care model.

overuse). This shift reflects the consensus that chatbots should not attempt to replicate clinical diagnosis, which requires integrative judgment, but should instead detect observable linguistic markers of distress. Furthermore, a symptom-based list better accommodates comorbidity, recognizing that users often present with overlapping symptom clusters rather than discrete diagnostic categories.

Following this symptom-based logic, several broad categories were bifurcated to distinguish between varying levels of risk and clinical urgency. *"Suicidal ideation and attempts"* was split into two distinct categories: *"Suicidal ideation"* and *"Suicidal intention and suicide attempts"*. This bifurcation aligns with clinical risk stratification, differentiating between passive thought content and active volitional intent or behavioral planning, which require distinct intervention protocols. Similarly, the category covering *"Psychosis and schizophrenia"* was divided into *"Psychotic symptoms"* and *"Sensory alterations (hallucinations and delusions)"* to allow for more targeted assessment.

"Bullying and harassment" also underwent refinement; experts recommended separating *"Bullying"* from "*Harassment and physical, sexual, and psychological abuse"*. This distinction was not intended to minimize the severity of bullying, but rather to differentiate its relational dynamics from the acute victimization contexts inherent in physical and sexual abuse. While experts emphasized that bullying entails profound psychological consequences, they prioritized separating these categories to address the immediate physical safety risks and specific safeguarding protocols required for abuse scenarios. Conversely, *"Sleep disorders"* was excluded from the framework, as experts deemed it a secondary symptom with lower immediate risk and relevance compared to acute psychiatric crises.

Finally, qualitative feedback highlighted gaps in the initial list, leading to the provisional addition of five new areas for evaluation in Round 2: *"Pathological grief"*, *"Occupational distress (burnout and workplace harassment/mobbing)"*, *"Issues related to sexuality and gender identity"*, *"Antisocial behaviors towards property and institutions (crimes and violations of legislative norms)"*, and *"Obsessive-compulsive symptoms"*. The complete refined list of critical areas, aggregated qualitative and quantitative data and preliminary priority ranking established in this round are detailed in Appendix A.

4.2. Round 2: consensus and intervention mapping

The second round focused on finalizing the list of critical areas and establishing the appropriate stepped-care response for each. In this phase, seven risk areas failed to meet the retention criteria and were removed. Notably, despite their clinical severity, *"Psychotic symptoms (hallucinations and delusions)"* and "*Manic symptoms"* were excluded. Experts argued that accurately detecting the nuances of reality distortion via text requires clinical observation beyond current chatbot capabilities. Similarly, areas such as *"Occupational distress"*, *"Issues related to sexuality and gender identity"*, and *"Antisocial behaviors towards property and institutions (crimes and violations of legislative norms)"* were deprioritized; while clinically relevant, they were viewed as not constituting the immediate safety threats required for a triage-focused framework. A comprehensive summary of the qualitative rationale for these exclusions is provided in Appendix B (Table B.1).

The exclusion process yielded a final list of 14 retained critical areas. Inter-rater agreement on the priority ranking of these items was strong (Kendall's $W = 0.78$, $p < .001$), indicating substantial consensus among the 39 experts (that concluded the second round) regarding their relative severity. Upon validating this hierarchy, the panel mapped each area to a minimum necessary intervention level (see Table 2).

The operational definitions for these confirmed critical areas are listed in Appendix B (Table B.2), while the full distribution of relevance scores and detailed qualitative feedback regarding intervention choices can be found in Appendix B (Table B.1).

High-acuity areas, specifically "*Suicidal ideation"*, *"Suicidal intention and suicide attempts*, and *"Harassment and physical, sexual, and psychological abuse"*, were assigned the highest intervention level. Experts agreed that chatbots must not attempt independent management of these risks but must immediately provide emergency contact information and crisis resources. The majority of critical areas, including *"Self-harm behaviors"*, *"Eating disorders"*, *"Substance addiction"*, and *"Depressive symptoms"*, were classified as requiring a "bridge" to human care. In these cases, the system's role is to validate the user's distress and actively facilitate a referral to a mental health professional. For lower-acuity presentations, experts endorsed a higher degree of automation. "*Anxious symptoms"* were deemed suitable for evidence-based self-help techniques (e.g., relaxation exercises), while *"Isolation and lack of social* support*"* was matched with empathetic active listening. This differentiation confirms the viability of a stepped-care model where the chatbot manages mild distress autonomously but acts strictly as a triage agent for severe pathology. Fig. 4 illustrates this process in practice, showing how the framework's logic is activated by specific user disclosures.

Table 2
Identified critical areas, priority rankings, and suggested interventions from the eDelphi study.

Identified Critical Areas	Priority	Primary Suggested Intervention
Suicidal ideation	1	Provide emergency contact information for crisis intervention services.
Suicidal intention and suicide attempts	2	Provide emergency contact information for crisis intervention services.
Psychotic symptoms	3	Suggest contacting a mental health professional.
Antisocial behaviors towards people (aggressiveness, violence, and potentially harmful behaviors towards others)	4	Suggest contacting a mental health professional.
Isolation and lack of social support	5	Offer active listening and empathetic responses.
Self-harm behaviors	6	Suggest contacting a mental health professional.
Eating disorders	7	Suggest contacting a mental health professional.
Bullying	8	Suggest contacting a mental health professional.
Depressive symptoms	9	Suggest contacting a mental health professional.
Anxious symptoms	10	Provide evidence-based self-help techniques.
Harassment and physical, sexual, and psychological abuse	11	Provide emergency contact information for crisis intervention services
Substance addiction	12	Suggest contacting a mental health professional.
Post-traumatic stress disorder symptoms	13	Suggest contacting a mental health professional.
Behavioral addictions	14	Suggest contacting a mental health professional.

5. Discussion

5.1. Principal findings

This study established a consensus-based framework for the governance of mental health chatbots, defining specific critical areas for monitoring and mapping them to adequate intervention levels. Beyond a simple checklist, the consensus delineated a clear operational model where chatbots function as complementary agents within a stepped-care ecosystem, balancing automated support for lower-acuity needs with rigorous escalation protocols for critical risks.

5.1.1. The role of chatbots in DMHIs

The expert consensus positions chatbots not merely as passive tools, but as active, complementary components of the care pathway. This

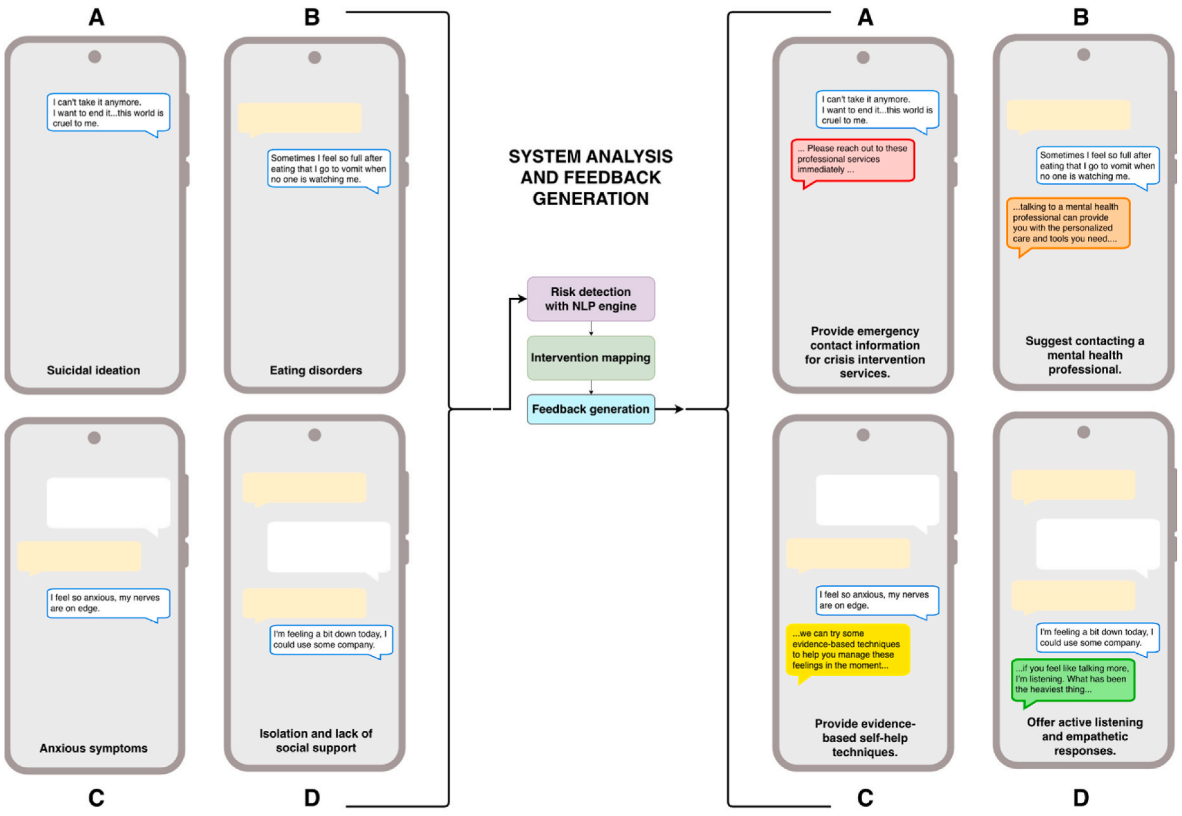

Fig. 4. Flow diagram demonstrating the framework in action across different scenarios.

aligns with the evolving literature on DMHIs, which advocates for a functional allocation of resources where technology supports, rather than supplants, human provision (Berman et al., 2018; Bolpagni et al., 2024; Borghouts et al., 2021). In this complementary role, the chatbot serves a dual function. First, it acts as an accessible entry point, potentially reducing the stigma associated with help-seeking and engaging populations who might otherwise avoid traditional services due to shame or logistical barriers (Ahmedani, 2011). Second, by autonomously managing initial assessments and low-risk conditions, such as social isolation or mild anxiety, chatbots can alleviate the burden on mental health services, allowing clinicians to focus strictly on complex cases requiring human oversight (Langhammer et al., 2024). However, the safety of this model rests on the precision of the handoff. As (Di Carlo et al., 2021) suggest, the risk lies not necessarily in the automated interaction itself, but in the potential gap between digital engagement and clinical referral. Our framework attempts to bridge this gap by explicitly defining the critical areas that mandate a break in the digital loop.

5.1.2. *Transition to symptom-based terminology and approach*

A significant outcome of the consensus process was the expert-driven move from disorder-based nomenclature to symptom-based terminology. This transition aligns the framework with broader shifts in both clinical (Botbol et al., 2016) and technological (Le Glaz et al., 2021) approaches, specifically the trend toward dimensional assessment rather than rigid categorical classifications found in traditional nosologies like the DSM-5 (American Psychiatric Association, 2013) or ICD-11 (Harrison et al., 2021). Categorical systems typically define conditions based on strict criteria that may not capture the fluidity of an individual's lived experience. In the context of chatbot interactions, a symptom-based approach allows for a more flexible assessment framework, acknowledging that users often present with transdiagnostic markers that do not fit a single diagnostic label but still necessitate support. Furthermore, this choice is pragmatically aligned with the operational capabilities of natural language processing (NLP). As noted by (Le Glaz et al., 2021), AI models are more adept at detecting linguistic markers of specific symptoms than inferring complex clinical diagnoses.

5.1.3. *Insights on the prioritization of areas and related intervention strategies*

The eDelphi results indicated a strong consensus that conditions posing an immediate threat to life or safety, such as suicidal ideation, suicidal intention and suicide attempts, and psychotic symptoms warrant the highest priority. This aligns with previous consensus studies on suicide prevention (Saini et al., 2022) and reinforces the necessity of urgent, directive intervention strategies for these critical areas that pose immediate risk to the individual. A nuanced finding within this hierarchy was the distinction between suicidal ideation and suicidal intention/attempts. While clinical risk assessment scales typically rate active planning or intent as more severe than ideation (Beck et al., 1979), the expert consensus highlighted the critical importance of detecting ideation in a digital context. In the preventive perspective of a chatbot, identifying the "thought" represents a vital window for early detection before it escalates into behavioral planning. Similarly, and in line with mental health triage protocols (Sands et al., 2016), antisocial behaviors towards people (like aggressiveness, violence) were identified as high-priority concerns, underscoring the need to monitor potential risks not only to the user but also to others.

However, the prioritization of risk areas did not correspond to a strictly linear hierarchy of intervention intensity. High severity consistently linked to crisis escalation, but "lower" priority areas often required more tailored, supportive approaches rather than simply less attention. Notably, social isolation (ranked 5th) was identified as a prime target for active chatbot engagement through active and empathetic listening. This finding is consistent with emerging approaches in digital mental health, which increasingly recognize chatbots not just as risk detectors, but as valuable first-line tools capable of providing emotional reassurance and mitigating distress in non-emergency

contexts (Torous et al., 2021). Consequently, the framework suggests two complementary pathways for intervention: in high-risk scenarios, the chatbot prioritizes safety and professional referral, whereas in lower-acuity contexts like social isolation, it can provide a first layer of empathetic support. Conversely, the exclusion of conditions such as sleep disorders highlights a necessary pragmatic trade-off. Experts determined that while clinically relevant, such issues are often secondary symptoms or possess lower immediate urgency, and thus were excluded to maintain the feasibility and focus of the monitoring framework.

5.2. Limitations

Several limitations should be considered when interpreting the results of this study. The primary limitation concerns the composition of the eDelphi expert panel. Despite efforts to recruit internationally, participation was predominantly concentrated within the Italian national context. While this ensures that the framework aligns with specific national clinical standards, it restricts the generalizability of the findings to other cultural healthcare systems, such as those in the US or UK, where legal frameworks and professional attitudes toward DMHIs may differ.

Furthermore, the panel was composed almost entirely of domain experts in psychology, predominantly clinicians and a significant proportion of early-career professionals, with a relatively small overall sample size. The study design did not include computer science experts; instead, we targeted psychologists with specific HCI expertise. This subgroup was intended to serve as a proxy for technical expertise, bridging the gap between clinical requirements and technological capabilities by leveraging their dual grasp of mental health concepts and digital systems. However, recruiting this specific intersection of expertise proved challenging (only 3 participated), resulting in limited representation. Consequently, the consensus reflects a predominantly clinical perspective that may harbor skepticism regarding automation. As noted by (Fietta et al., 2022), practitioners often perceive emerging technologies as challenging their professional roles or lacking the nuance required for complex clinical decision-making. This “protectionist” stance, combined with potential round-to-round attrition favoring the most committed participants may have skewed the consensus toward a stronger preference for human referral over automated intervention, potentially underestimating the technical capabilities of current state-of-the-art systems.

A second limitation concerns the distinction between the clinical definition of risk (what to monitor) and the technical capability to detect it (how to monitor). While this study establishes a theoretical consensus on risk areas, it does not address the technical implementation. The reliance on symptom-based distinctions, such as self-harm vs. suicide attempt, or bullying vs. harassment presents a significant challenge for NLP architectures. These boundaries, while clinically distinct, can be linguistically ambiguous in user-generated text. Current NLP systems, including LLMs, face inherent limitations in this regard; while they excel at capturing surface-level patterns and statistical regularities in language, they lack robust internal representations of deeper psychological constructs, including latent intent, motivational states, volitional capacity, and broader contextual factors central to clinical risk assessment (Bak & Chin, 2024; Bender et al., 2021; Iftikhar et al., 2025). For instance, distinguishing self-harm from suicidal intention requires identifying latent intent that often remains unexpressed in the literal text. Therefore, the practical accuracy of this framework depends entirely on the sophistication of the underlying NLP architecture and implementation. Future interdisciplinary research involving computer science experts is essential to determine whether these fine-grained clinical distinctions can be effectively operationalized by current algorithms.

Finally, it is important to distinguish the scope of safety addressed in this study. Our expert panel provided a robust consensus on clinical safety, specifically how to monitor and manage risks originating from the user, such as suicidal ideation or abuse. However, this focus does not address “reverse risks”, which are hazards originating from the AI that can actively worsen a user's mental health. As noted in recent literature (Krook, 2024), the chatbot itself can become a psychological stressor. For instance, hallucinations (fabricating information) could reinforce delusional thinking in vulnerable users, while culturally biased responses could deepen feelings of alienation or stigma. These are not merely technical errors but clinical risks where the system's output itself causes harm.

5.3. Future research

Future research should prioritize replicating this study with a diverse, international panel to test the cross-cultural validity of the list of critical areas. Variations in legal frameworks, distress expression, and societal trust in digital health may necessitate regional adaptations of the priority hierarchy. Moreover, this framework is intended as a preliminary blueprint rather than an exhaustive classification. The checklist approach possesses inherent limitations in granularity; therefore, subsequent studies must refine and expand these categories to match the evolving landscape of mental health needs. In doing so, it should be recognized that the current framework primarily reflects a clinical requirement perspective and necessitates further interdisciplinary validation to ensure technical feasibility.

The logical advancement of this work involves the operationalization of these consensus guidelines into technical architectures. Researchers should explore translating this list of critical areas into computational rules, potentially exploring the use of LLM-based guardrails to automatically enforce the identified safety thresholds. A promising avenue for empirical validation would be integrating this framework into established digital interventions, such as the chatbot implementation of the World Health Organization (WHO) “Self-Help Plus” protocol (Fietta et al., 2024). Once a technical implementation is achieved, empirical studies must move beyond technical feasibility to evaluate clinical impact. Key metrics should include the system's accuracy in detecting “red flags” and the appropriateness of the triggered interventions. Long-term evaluation should also assess user satisfaction and the potential for these automated systems to reduce dysfunctional access to emergency resources, thereby lowering healthcare costs. This would validate the framework's utility not just as a safety mechanism, but as a functional component of a cost-effective, stepped-care ecosystem.

6. Conclusions

This study established a consensus-based framework for monitoring critical areas in mental health chatbots, addressing the urgent need for safety guidelines in this rapidly evolving technological field. The priority ranking and selected risk areas identified in this work can serve as a foundational framework for developing chatbot systems capable of recognizing and addressing key psychological risks without losing focus on the patient's well-being. While further investigation is needed, this study underscores the importance of chatbots in supporting specific mental health conditions without excluding experts' involvement in their design and evaluation process. Such engagement represents a crucial step toward positioning technology not as a threat, but rather as a valuable resource within a stepped-care approach to address the unequal global distribution of mental health resources.

CRediT authorship contribution statement

Marco Bolpagni: Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Valentina Fietta:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Nicolò Navarin:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Merylin Monaro:** Writing –

review & editing, Writing – original draft, Data curation, Conceptualization. **Silvia Gabrielli:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

Disclaimer

This publication reflects only the authors' views, and the Italian Ministry of Health is not responsible for any use that may be made of the information it contains.

Institutional review board statement

Not applicable.

Generative AI statement

During the preparation of this work the author(s) used Google Gemini (3 Pro) in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Funding

Life Science Hub – Digital Health (LSH-DH) PNC-E3-2022-23683267 - DHEAL-COM Project – CUP: C63C22001970001, funded by the Ministry of Health (Italy) as part of the National Complementary Plan for Innovative Health Ecosystems.

This publication reflects only the authors' view and the Italian Ministry of Health is not responsible for any use that may be made of the information it contains.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank all the experts that participated in the study. Their insights have been fundamental to identify critical areas that need to be monitored in mental health chatbots interactions and their related intervention paths.

Abbreviations

AI	Artificial Intelligence
LLMs	Large Language Models
DMHIs	Digital Mental Health Interventions
IQR	Interquartile Range
NLP	Natural Language Processing
HCI	Human-Computer Interaction
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, 5th Edition
ICD-11	International Classification of Diseases, 11th Revision
WHO	World Health Organization
CTRS	Crisis Triage Rating Scale
UK MHTS	UK Mental Health Triage Scale
DTx	Digital Therapeutics
CRI	Crisis Intervention
EMP	Empathetic Listening
EDU	Educational Resources
SHP	Self-Help Techniques
PRO	Professional Contact Suggestion

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbr.2026.101043.

Data availability

Data will be made available on request.

References

Ah Lee, S. K. (2018). Classification of SmartMentalTech services and application for comprehensive mental healthcare stepped-care model (CMHSCM): Health psychological approach. *Procedia Computer Science, 141*, 302–310.

Ahmedani, B. K. (2011). Mental health stigma: Society, individuals, and the profession. *Journal of Social Work Values and Ethics, 8*(2), 41–416.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association.

Baig, M. M., GholamHosseini, H., & Connolly, M. J. (2015). Mobile healthcare applications: System design review, critical issues and challenges. *Australasian Physical & Engineering Sciences in Medicine, 38*(1), 23–38.

Bak, M., & Chin, J. (2024). The potential and limitations of large language models in identification of the states of motivations for facilitating health behavior change. *Journal of the American Medical Informatics Association, 31*(9), 2047–2053.

Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: The scale for suicide ideation. *Journal of Consulting and Clinical Psychology, 47*(2), 343–352.

Belton, I., MacDonald, A., Wright, G., & Hamlin, I. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. *Technological Forecasting and Social Change, 147*, 72–82.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. FAccT '21: 2021 ACM conference on fairness, accountability, and transparency, virtual event Canada.* https://doi.org/10.1145/3442188.3445922

Bengelsdorf, H., Levy, L. E., Emerson, R. L., & Barile, F. A. (1984). A crisis triage rating scale. Brief dispositional assessment of patients at risk for hospitalization. *The Journal of Nervous and Mental Disease, 172*(7), 424–430.

Berman, A. H., Kolaas, K., Petersén, E., Bendtsen, P., Hedman, E., Linderoth, C., Müssener, U., Sinadinovic, K., Spak, F., Gremyr, I., & Thurang, A. (2018). Clinician experiences of healthy lifestyle promotion and perceptions of digital interventions as complementary tools for lifestyle behavior change in primary care. *BMC Family Practice, 19*(1), 139.

Bolpagni, M., & Gabrielli, S. (2025). Development of a comprehensive evaluation scale for LLM-Powered counseling chatbots (CES-LCC) using the Edelphi method. *Informatics, 12*, 33.

Bolpagni, M., Pardini, S., & Gabrielli, S. (2024). Human centered design of AI-powered digital therapeutics for stress prevention: Perspectives from multi-stakeholders' workshops about the SHIVA solution. *Internet Interventions, 38*(100775), Article 100775.

Borghouts, J., Eikey, E., Mark, G., De Leon, C., Schueller, S. M., Schneider, M., Stadnick, N., Zheng, K., Mukamel, D., & Sorkin, D. H. (2021). Barriers to and facilitators of user engagement with digital mental health interventions: Systematic review. *Journal of Medical Internet Research, 23*(3), Article e24387.

Botbol, M., Banzato, C. E. M., & Salvador-Carulla, L. (2016). Categories, dimensions, and narratives for person-centered diagnostic assessment. In *Person centered psychiatry* (pp. 201–208). Springer International Publishing.

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: Access, effectiveness and efficiency. Narrative literature review. *The British Journal of Psychiatry: Journal of Mental Science, 186*, 11–17.

Burgess, N., Christensen, H., Leach, L. S., Farrer, L., & Griffiths, K. M. (2008). Mental health profile of callers to a telephone counselling service. *Journal of Telemedicine and Telecare, 14*(1), 42–47.

Clarke, V., & Braun, V. (2017). Thematic analysis. *The Journal of Positive Psychology, 12*(3), 297–298.

Coyne, I. T. (1997). Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries? *Journal of Advanced Nursing, 26*(3), 623–630.

Davies, M. (2006). Allocating resources in mental health: A clinician's guide to involvement. *Advances in Psychiatric Treatment: The Royal College of Psychiatrists' Journal of Continuing Professional Development, 12*(5), 384–391.

Denecke, K., May, R., LlmhealthGroup, & Rivera Romero, O. (2024). Potential of large language models in health care: Delphi study. *Journal of Medical Internet Research, 26*, Article e52399.

Di Carlo, F., Sociali, A., Picutti, E., Pettorruso, M., Vellante, F., Verrastro, V., Martinotti, G., & di Giannantonio, M. (2021). Telepsychiatry and other cutting-edge technologies in COVID-19 pandemic: Bridging the distance in mental health assistance. *International Journal of Clinical Practice, 75*(1), Article e13716.

Donohoe, H., Stellefson, M., & Tennant, B. (2012). Advantages and limitations of the e-Delphi technique: Implications for health education researchers. *American Journal of Health Education, 43*(1), 38–46.

Ferreira, R., Trifan, A., & Oliveira, J. L. (2022). Early risk detection of mental illnesses using various types of textual features. In *Conference and labs of the evaluation forum* (pp. 905–920).

Fietta, V., Rizzi, S., De Luca, C., Gios, L., Pavesi, M. C., Gabrielli, S., Monaro, M., & Forti, S. (2024). A chatbot-based version of the world health organization-validated self-help plus intervention for stress management: Co-design and usability testing. *JMIR Human Factors, 11*, Article e64614.

Fietta, V., Zecchinato, F., Stasi, B. D., Polato, M., & Monaro, M. (2022). Dissociation between users' explicit and implicit attitudes toward artificial intelligence: An experimental study. *IEEE Transactions on Human-Machine Systems, 52*(3), 481–489.

Frasquilho, D., Matias, R., Grácio, J., Sousa, B., Luís-Ferreira, F., Leal, J., Cardoso, F., & Oliveira-Maia, A. J. (2021). Protocol for the implementation and assessment of "MoodUP": A stepped care model assisted by a digital platform to accelerate access to mental health care for cancer patients amid the COVID-19 pandemic. *International Journal of Environmental Research and Public Health, 18*(9), 4629.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences, 18*, 43–49.

Haaga, D. A. (2000). Introduction to the special section on stepped care models in psychotherapy. *Journal of Consulting and Clinical Psychology, 68*(4), 547–548.

Haleem, A., Javaid, M., Singh, R. P., & Suman, R. (2021). Telemedicine for healthcare: Capabilities, features, barriers, and applications. *Sensors International, 2*(100117), Article 100117.

Harrison, J. E., Weber, S., Jakob, R., & Chute, C. G. (2021). ICD-11: An international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making, 21*(Suppl 6), 206.

Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA), 24*(3), 383–394.

Iftikhar, Z., Xiao, A., Ransom, S., Huang, J., & Suresh, H. (2025). How LLM counselors violate ethical standards in mental health practice: A practitioner-informed framework. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(2), 1311–1323.

Kendall, M. G., & Smith, B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics, 10*(3), 275–287.

Krook, J. (2024). Manipulation and the Ai act: Large language model chatbots and the danger of mirrors. https://doi.org/10.2139/ssrn.4719835.

Langhammer, T., Hilbert, K., Wasenmueller, R., Praxl, B., Ertle, A., Asbrand, J., & Lueken, U. (2024). Evaluation of a CBT-based program for mental health in the general population during the COVID-19 pandemic: A stepped-care approach using a chatbot and digitized group intervention. *Depression and Anxiety, 2024*(1). https://doi.org/10.1155/2024/8950388

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research, 23*(5), Article e15708.

McClellan, C., Ali, M. M., Mutter, R., Kroutil, L., & Landwehr, J. (2017). Using social media to monitor mental health discussions – evidence from Twitter. *Journal of the American Medical Informatics Association: JAMIA, 24*(3), 496–502.

Nasa, P., Azoulay, E., Khanna, A. K., Jain, R., Gupta, S., Javeri, Y., Juneja, D., Rangappa, P., Sundararajan, K., Alhazzani, W., Antonelli, M., Arabi, Y. M., Bakker, J., Brochard, L. J., Deane, A. M., Du, B., Einav, S., Esteban, A., Gajic, O., … Myatra, S. N. (2021). Expert consensus statements for the management of COVID-19-related acute respiratory failure using a Delphi method. *Critical Care, 25*(1), 106.

O'Dowd, A. (2025). ChatGPT: More than a million users show signs of mental health distress and mania each week, internal data suggest. *BMJ, 391*, r2290.

Omboni, S., Padwal, R. S., Alessa, T., Benczúr, B., Green, B. B., Hubbard, I., Kario, K., Khan, N. A., Konradi, A., Logan, A. G., Lu, Y., Mars, M., McManus, R. J., Melville, S., Neumann, C. L., Parati, G., Renna, N. F., Ryvlin, P., Saner, H., … Wang, J. (2022). The worldwide impact of telemedicine during COVID-19: Current evidence and recommendations for the future. *Connected Health, 1*, 7–35.

Philippe, T. J., Sikder, N., Jackson, A., Koblanski, M. E., Liow, E., Pilarinos, A., & Vasarhelyi, K. (2022). Digital health interventions for delivery of mental health care: Systematic and comprehensive meta-review. *JMIR Mental Health, 9*(5), Article e35159.

Qualtrics XM. (2018). Qualtrics. https://www.qualtrics.com.

Ríssola, E. A., Parapar, J., Losada, D. E., & Crestani, F. (2022). A survey of the first five years of eRisk: Findings and conclusions. In *Early detection of mental health disorders by social media monitoring* (pp. 31–57). Springer International Publishing.

Saari, D. G. (2023). Selecting a voting method: The case for the borda count. *Constitutional Political Economy, 34*(3), 357–366.

Saini, P., Clements, C., Gardner, K. J., Chopra, J., Latham, C., Kumar, R., & Taylor, P. (2022). Identifying suicide and self-harm research priorities in north west England: A Delphi study. *Crisis, 43*(1), 35–45.

Saleem, S., Prasad, R., Vitaladevuni, S., Pacula, M., Crystal, M., Marx, B., Sloan, D., Vasterling, J., & Speroff, T. (2012). Automatic detection of psychological distress indicators and severity assessment from online forum posts. *International Conference on Computational Linguistics*, 2375–2388.

Sands, N., Elsom, S., Colgate, R., Haylor, H., & Prematunga, R. (2016). Development and interrater reliability of the UK mental health triage scale. *International Journal of Mental Health Nursing, 25*(4), 330–336.

Sands, N., Elsom, S., Corbett, R., Keppich-Arnold, S., Prematunga, R., Berk, M., & Considine, J. (2017). Predictors for clinical deterioration of mental state in patients assessed by telephone-based mental health triage. *International Journal of Mental Health Nursing, 26*(3), 226–237.

Sands, N., Elsom, S., Marangu, E., Keppich-Arnold, S., & Henderson, K. (2013). Mental health telephone triage: Managing psychiatric crisis and emergency. *Perspectives in Psychiatric Care, 49*(1), 65–72.

Sheehan, K. A., Schulz-Quach, C., Ruttan, L. A., MacGillivray, L., McKay, M. S., Seto, A., Li, A., Stewart, D. E., Abbey, S. E., & Berkhout, S. G. (2023). "don't just study our distress, do something": Implementing and evaluating a modified stepped-care model for health care worker mental health during the COVID-19 pandemic. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie, 68*(1), 43–53.

Singh, O. P. (2023). Artificial intelligence in the era of ChatGPT - Opportunities and challenges in mental health care. *Indian Journal of Psychiatry, 65*(3), 297–298.

Skaik, R., & Inkpen, D. (2021). Using social media for mental health surveillance. *ACM Computing Surveys, 53*(6), 1–31.

Smart, D., Pollard, C., & Walpole, B. (1999). Mental health triage in emergency medicine. *Australian and New Zealand Journal of Psychiatry, 33*(1), 57–66. ; discussion 67–69.

Smith, L. S., Dnp, C. T. B.-U., & Spencer Cole, R. W. (2016). Practice matters: Red flags in adults with mental illnesses. *International Journal of Faith Community Nursing, 2*(2), 39.

Tielman, M. L., Neerincx, M. A., Pagliari, C., Rizzo, A., & Brinkman, W.-P. (2019). Considering patient safety in autonomous e-mental health systems - Detecting risk situations and referring patients back to human care. *BMC Medical Informatics and Decision Making, 19*(1), 47.

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA), 20*(3), 318–335.

Turner, P. M., & Turner, T. J. (1991). Validation of the crisis triage rating scale for psychiatric emergencies. *Canadian journal of psychiatry. Revue Canadienne de Psychiatrie, 36*(9), 651–654.

van Hecke, O., Kamerman, P. R., Attal, N., Baron, R., Bjornsdottir, G., Bennett, D. L. H., Bennett, M. I., Bouhassira, D., Diatchenko, L., Freeman, R., Freynhagen, R., Haanpää, M., Jensen, T. S., Raja, S. N., Rice, A. S. C., Seltzer, Z. 'ev, Thorgeirsson, T. E., Yarnitsky, D., & Smith, B. H. (2015). Neuropathic pain phenotyping by international consensus (NeuroPPIC) for genetic studies: A NeuPSIG systematic review, Delphi survey, and expert panel recommendations. *Pain, 156*(11), 2337–2353.

Wies, B., Landers, C., & Ienca, M. (2021). Digital mental health for young people: A scoping review of ethical promises and challenges. *Frontiers in Digital Health, 3*, Article 697072.