GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2026) 38(3) 113-122

Psicothema

https://www.psicothema.com/es • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Psicothema

Article

A Multilevel Framework for the Safe and Ethical Use of AI in Mental Health Practice

Ignacio Pedrosa

International University of La Rioja (UNIR) (Spain)

ARTICLE INFO

Received: 04/11/2025
Accepted: 05/03/2026

Keywords:
Digital mental health
Psychological ethics
Artificial intelligence
Psychological regulation
Digital wellbeing

ABSTRACT

Background: Incorporating Artificial Intelligence (AI) into mental health practice represents a profound transformation of psychological care, enabling novel forms of assessment, intervention, and support. However, this shift also introduces substantial ethical and psychological risks, including erosion of autonomy, algorithmic bias, threats to digital rights, and strain on clinical judgment and therapeutic relationships. This article proposes PRISM, a multilevel framework designed to guide the ethical integration of AI in mental health across individual, technological, and institutional dimensions. **Method:** An integrative conceptual review was conducted drawing on Self-Determination Theory, information ethics, data justice, and international guidelines from the APA, WHO, EU, and UNESCO. The framework introduces the concept of psychological regulation of AI, extending ethical oversight to the subjective and relational impacts of algorithmic mediation. **Results:** PRISM articulates four operational principles (digital autonomy, affective transparency, algorithmic justice, and relational care) along with practical tools, including psychological impact indicators and a Psychological Model Card, to support ethical auditing and clinical digital wellbeing. **Conclusions:** A psychology-based regulatory approach is essential to ensure dignity, autonomy, and ethical sustainability in digital mental health. PRISM provides an operational bridge between technological governance and psychological practice, supporting responsible, person-centred AI use in clinical contexts.

Marco Multinivel para el Uso Seguro y Ético de la Inteligencia Artificial en la Práctica de la Salud Mental

Palabras clave:
Salud mental digital
Ética psicológica
Inteligencia artificial
Regulación psicológica
Bienestar digital

RESUMEN

Antecedentes: La integración de la Inteligencia Artificial (IA) en la práctica de la salud mental representa una transformación de la atención psicológica, ofreciendo nuevas formas de evaluación, intervención y apoyo. Pero también introduce riesgos éticos y psicológicos sustanciales como la erosión de la autonomía, el sesgo algorítmico, amenazas a los derechos digitales y la tensión en el juicio clínico y las relaciones terapéuticas. Se propone un marco multinivel (PRISM) diseñado para guiar la integración ética de la IA en la salud mental mediante dimensiones individuales, tecnológicas e institucionales. **Método**: Se realizó una revisión conceptual integradora basada en la Teoría de la Autodeterminación, ética de la información, justicia de datos y directrices internacionales (APA, OMS, UE y UNESCO). **Resultado**: PRISM articula cuatro principios (autonomía digital, transparencia afectiva, justicia algorítmica y cuidado relacional) junto con herramientas prácticas, incluyendo indicadores de impacto psicológico y una ficha modelo, para apoyar la auditoría ética y el bienestar digital clínico. **Conclusiones**: Un enfoque regulador psicológico es fundamental para garantizar la dignidad, autonomía y sostenibilidad ética en la salud mental digital. PRISM proporciona un puente operativo entre la gobernanza tecnológica y la práctica psicológica, apoyando el uso responsable y centrado en la persona en contextos clínicos.

Cite as: Pedrosa, I. (2026). A multilevel framework for the safe and ethical use of AI in mental health practice. *Psicothema, 38*(3), 113-122. https://doi.org/10.70478/psicothema.2026.38.11
Corresponding author: Ignacio Pedrosa, ignacio.pedrosa@unir.net

Over the past two decades, mental health has experienced accelerated digitalization. The advent of Artificial Intelligence (AI) has not only represented a technological revolution but also an epistemological and ethical reconfiguration of applied psychology (Dehbozorgi et al., 2025). The increasing deployment of platforms, mobile applications, machine learning approaches, simulated patients, and therapeutic chatbots has transformed psychological literacy, assessment, prevention, and treatment (Sanz et al., 2025; Torous et al., 2025; Wang et al., 2025).

Recent studies have introduced the concept of digital compassion, defined as the capacity of AI systems to modulate empathetic responses and emotional support in mental health contexts without replacing human interaction (Abou Hashish, 2025; Wiljer et al., 2019). Unlike earlier digital tools, such as traditional smartphone apps, which functioned primarily as passive instruments, AI operates as a mediating entity capable of simulating cognitive and affective processes, thereby reshaping core psychological practices (Wiljer et al., 2019). This shift calls for a specifically psychological form of regulation, as existing biomedical or legal frameworks (such as the EU AI Act) prioritize systemic safety while neglecting the subjective impact of algorithmic mediation and its effects on the therapeutic relationship.

Current AI applications in mental health include test construction, assessment, therapeutic support, and clinical decision assistance, ranging from risk detection models to conversational agents and documentation systems (Cheng, 2024; Dehbozorgi et al., 2025; Suárez-Álvarez et al., 2026).

Although these tools can enhance accuracy and accessibility, their opacity may undermine patient autonomy and the therapeutic alliance when limitations are not adequately communicated (Theunissen & Browning, 2022).

At the same time, AI offers substantial benefits, including improved diagnostic accuracy, increased access for underserved populations, and reduced administrative burden, allowing greater clinical focus on patient needs (Milasan & Scott-Purdy, 2025; Yıldız, 2025a).

Nevertheless, these advances introduce unresolved risks, such as algorithmic opacity, automated bias, sensitive data governance challenges, and the potential dehumanization of therapeutic relationships (Adler et al., 2024; WHO, 2023).

Gap Between Technological Development and Psychological Reflection

The development of AI has significantly outpaced the capacity of existing ethical and regulatory frameworks to address its psychological implications. Current regulatory efforts, including the EU AI Act (Regulation EU 2024/1689) and WHO guidelines (2023, 2025), remain largely shaped by a technical–biomedical bias that relegates ethical-psychological concerns to a secondary role. The EU AI Act's risk-based approach prioritizes systemic safety but lacks specific criteria for assessing the mental impact of AI systems on user subjectivity or the integrity of the therapeutic relationship.

Similarly, although the APA (2025) and WHO (2025) stress the importance of human oversight, they do not resolve what has been described as the oversight paradox: the inherent contradiction of requiring meaningful human control over agentic AI systems explicitly designed for autonomous goal achievement (Kyrychenko et al., 2026).

This limitation is compounded by the fragmented interaction between the GDPR, the Digital Services Act, and the AI Act, which produces an accountability paradox in clinical contexts (Kyrychenko et al., 2026). A single biased algorithmic output may trigger multiple regulatory responses without providing a coherent mechanism for reporting or mitigating complex psychological harm. By privileging technical transparency over affective transparency current frameworks risk generating moral fatigue among practitioners and mistrust among patients (Theunissen & Browning, 2022).

This regulatory gap underscores the need for a specialized psychological approach capable of integrating disparate legal requirements into a person-centred clinical reality.

Emerging Psychological Effects

Recent studies indicate that sustained interaction with AI-based therapeutic agents produces mixed psychological outcomes (Torous et al., 2025; Wang et al., 2025). While these systems enhance accessibility, skill acquisition, and anonymity, they are also associated with emotional dependency, reduced perceived control, digital moral fatigue, and confusion between human and artificial relationships.

Within this context, algorithmic anxiety (i.e., the distrust and uncertainty surrounding automated decision-making) has emerged as a relevant construct in social and clinical psychology (Li & Liu, 2025). Such subjective effects, often invisible to conventional efficacy metrics, directly affect psychological wellbeing, self-efficacy, and therapeutic adherence, with meaningful consequences for mental health outcomes.

These emerging effects highlight the need for evaluative approaches that extend beyond performance and safety indicators. A psychological regulatory perspective is required to assess AI-mediated mental health interventions through the lenses of subjectivity, relational dynamics, and ethics of care. From this standpoint, four core questions become central: how patient autonomy is preserved in automated environments; what limits should govern algorithmic supervision of behaviour; how explainability and algorithmic justice can be ensured in clinical contexts; and which ethical–digital competencies are required of mental health professionals. Addressing these questions requires an integrative framework capable of translating ethical principles into operational criteria for clinical practice, professional training, and institutional governance in digital mental health.

This work pursues a dual objective. First, it develops an integrative theoretical synthesis of the main ethical-psychological dilemmas associated with the application of AI in mental health. Second, it proposes a model articulated around three interdependent axes (i.e. individual, technological, and institutional) to guide policy design, professional training, and ethical best practices in the digital era. Unlike existing bioethical, medical, or legal frameworks (Tang et al., 2025), no equivalent concept has yet been defined in recent systematic reviews of digital health from a psychological perspective.

The study adopts an integrative conceptual review methodology, designed to synthesize diverse theoretical perspectives and empirical findings into a unified regulatory framework. The search was conducted across three primary high-impact databases: APA PsycInfo, PubMed, and Scopus, supplemented by grey literature from international regulatory bodies (EU, WHO, APA, UNESCO). The search strategy utilized combinations of Boolean operators and keywords, including: ('*Artificial Intelligence*' OR '*Generative AI*') AND ('*Mental Health*' OR '*Psychology*') AND ('*Ethics*' OR

'*Regulation*' OR '*Digital Rights*') AND ('*Algorithmic Bias*' OR '*Autonomy*'). The scope was primarily limited to the period from 2020 to 2025 to ensure the inclusion of the most recent developments in large multi-modal models and international AI governance.

Inclusion criteria focused on a) peer-reviewed articles discussing the psychological impact of AI on the therapeutic relationship or user subjectivity, b) official ethical guidelines from psychological associations, and c) international legislative frameworks for AI health technologies. Exclusion criteria were applied to purely technical or biomedical reports that did not address the human-AI interaction or the ethical-psychological dimension (Hoose & Kralikova, 2024).

Theoretical Framework: Psychology Facing Artificial Intelligence

The development of AI in mental health represents a historical inflection point comparable to the emergence of psychoanalysis or cognitive-behavioural therapy in their respective eras. Unlike previous paradigmatic shifts, psychology now shares agency with non-human systems capable of simulating cognitive and affective functions (Sanz et al., 2025). This unprecedented condition compels a reconsideration of the discipline's epistemological foundations and its role as guarantor of psychic integrity within increasingly digitalized clinical contexts.

Several psychological paradigms provide conceptual tools for understanding human-AI interaction. Self-Determination Theory (Deci & Ryan, 2000) posits that wellbeing depends on the satisfaction of autonomy, competence, and relatedness that AI systems may either support or undermine depending on their design and implementation. This perspective underscores the need to distinguish psychological regulation from broader approaches such as applied ethics or digital bioethics, which primarily prioritize systemic safety and regulatory compliance (Dencik & Sanchez-Monedero, 2022; EU, 2024/1689; Torous et al., 2025). In contrast, psychological regulation evaluates AI-mediated care through its impact on subjectivity, relational dynamics, and basic psychological needs, reframing regulation from external rule compliance to the preservation of psychological integrity in digital clinical environments.

While traditional ethical frameworks focus on systemic risk mitigation, psychological regulation targets specific constructs such as algorithmic anxiety (Yang & Sundar, 2025), the distress arising from unpredictable automated processes, and digital moral fatigue among practitioners (Epstein et al., 2019). By prioritizing understandability as an outcome of transparency and interpretability, this approach seeks to prevent the depersonalization of care (Padalkar et al., 2024) and to safeguard the therapeutic bond in technologically mediated interactions.

From a humanistic perspective, authenticity and empathy remain central to the therapeutic process. The partial delegation of these dimensions to algorithms raises critical questions about relational authenticity, perceived agency, and the limits of technological mediation in clinical care.

At a structural level, Social Cognitive Theory (Bandura, 2001) introduces the notion of digital self-efficacy, understood as perceived control over automated processes. Similarly, Information Ethics (Floridi, 2019) emphasizes the protection of autonomy, dignity, and agency as intrinsic values of the informational ecosystem, while the Data Justice perspective (Dencik & Sanchez-Monedero, 2022) and international guidelines (WHO, 2023) extend these concerns to equity and democratic participation. Together, these frameworks provide an essential ethical–psychological foundation for embedding human values into technological design.

These perspectives converge on a central premise: AI systems are not neutral. Their architectures embody implicit values that shape identity, perceived control, emotional experience, and professional responsibility. Consequently, AI transforms not only diagnostic and therapeutic procedures but also the structure of the clinical relationship itself. Before proposing a regulatory framework, it is therefore essential to analyse the ethical and psychological dilemmas that emerge when psychological practice becomes technologically mediated.

Emerging Ethical and Psychological Dilemmas

Building on these theoretical foundations, the following sections identify the primary ethical and psychological challenges that justify a dedicated regulatory framework.

Algorithmic Bias and Psychological Justice

A central ethical concern is algorithmic bias, which may replicate historical discrimination embedded in clinical and population datasets. Adler et al. (2024) showed that AI models in psychiatric diagnostics often underrepresent women and racial minorities, reducing diagnostic accuracy and undermining patients' epistemic trust.

Psychological regulation should include bias audits addressing both statistical and experiential dimensions, posing questions such as: *How does the patient perceive the system's fairness?* and *How does this perception influence therapeutic adherence and the alliance?*

Autonomy and Digital Consent

Autonomy, a cornerstone of psychological ethics, is challenged by the technical complexity of AI. Many users consent to data processing without a full understanding of algorithmic inference mechanisms, highlighting the need for cognitively and emotionally intelligible consent protocols (APA, 2025).

Therapeutic Relationship and Dehumanization

Studies on therapeutic chatbots (e.g., Replika) indicate that while accessibility improves, users may develop emotional dependency and experience reduced relational authenticity (Laestadius et al., 2024). Misalignment between AI responses and the user's affective-cognitive context exacerbates this disconnect (Loechner et al., 2025).

This emotional disconnect affects not only patients but also mental health professionals, who must manage interactions mediated by systems incapable of genuine affective responsiveness. This not only affects patients but also contributes to digital moral fatigue in practitioners, who must mediate interactions with systems incapable of genuine empathy. Ethical frameworks should therefore integrate the intersubjective experience of both patients and clinicians, combining the protection of patient wellbeing with the ethical supervision of algorithmic systems.

Privacy, Surveillance, and Psychological Wellbeing

Privacy research demonstrates that perceived constant surveillance negatively impacts emotional wellbeing and cognitive performance (Solove, 2021). Mental health monitoring tools analyzing text or voice can provoke exposure anxiety, heightening self-consciousness and behavioural inhibition.

Jointly, these dilemmas reveal that regulatory challenges are not merely technical, they are profoundly psychological. They underscore the necessity of a framework that safeguards autonomy, equity, and wellbeing while accounting for the relational dynamics inherent in AI-mediated care.

The PRISM Framework: A Psychological Regulatory Model for AI in Mental Health

The regulation of AI in mental health cannot be restricted to legal or technical mechanisms alone. It requires the integration of psychological principles that guide the design, implementation, and evaluation of intelligent systems in ways that uphold human dignity and wellbeing.

The proposed framework, hereafter referred to as the PRISM Framework -Psychological Regulation and Integrity in Smart Mental Health-, is grounded in four integrative principles, derived from both classical psychological ethics and contemporary research on AI:

- **Digital autonomy:** the individual's capacity to understand and make informed decisions regarding technologies that influence their mind, emotions, or behaviour.
- **Affective transparency:** the requirement for AI systems to clearly communicate their limitations in emotional understanding, to prevent false attributions of empathy or intentionality.
- **Algorithmic justice:** the ethical obligation to prevent bias through systematic auditing and the participatory inclusion of vulnerable or underrepresented groups in the design and validation process.
- **Relational care:** the preservation of authentic human presence and the quality of the therapeutic relationship as central components of psychological intervention.

This framework is not merely a normative proposal, but a conceptually grounded model based on the intersection of SDT and aspects linked to transparency and interpretability for understandability (Deci & Ryan, 2000; Joyce et al., 2023). The theoretical rigor of PRISM resides in the premise that psychological well-being in digital environments depends on the satisfaction of innate needs for autonomy, competence, and relatedness (Deci & Ryan, 2000). This framework primarily addresses the professional stakeholder, focusing on the ethical and psychological challenges clinicians face when interacting with AI.

Together, these principles constitute the foundation of the PRISM Framework, which is structured around three interdependent axes (i.e. individual, technological, and institutional) forming an ethical-psychological architecture designed to protect digital rights and promote mental health wellbeing (Table 1). Empirically, the Algorithmic Justice dimension is validated by findings demonstrating that smartphone-sensed behaviours are unreliable predictors across diverse populations. For instance, phone usage patterns that signify depression in younger cohorts can represent healthy engagement in older adults (Adler et al., 2024). This reliability gap requires the Individual Axis of the framework, which integrates trait-based assessments to calibrate trust and mitigate digital moral fatigue (Yang & Sundar, 2025). Furthermore, the Technological Axis adopts the operational definition of understandability. It posits that for a system to be clinically valid, its computational structures must stand in close correspondence with human clinical heuristics, such as abductive and inductive inference (Joyce et al., 2023). By shifting from post-hoc explanations to intrinsic interpretability, PRISM ensures that the mental impact of algorithmic mediation remains traceable and ethically auditable. Finally, the Institutional Axis is supported by implementation science models that emphasize the role of digital navigators in maintaining the human connection, a factor shown to significantly enhance the efficacy of digital interventions compared to unguided tools (Torous et al., 2025).

Individual Axis: Wellbeing, Digital Autonomy, and Agency

This axis represents the core of the proposed framework, as mental health is conceptualized as a dynamic balance among self-determination, perceived competence, and human connection (Deci & Ryan, 2000). The integration of AI introduces novel psychological determinants of digital wellbeing, including the following constructs:

- **Perceived digital autonomy:** individual's sense of control over the intelligent systems with which they interact. A deficit in this domain is associated with algorithmic anxiety and feelings of helplessness (Yang & Lu, 2026).
- **Algorithmic anxiety:** emotional response to the opacity or unpredictability of algorithmic processes, frequently observed in users who delegate clinical decisions to automated systems (Yang & Sundar, 2025).
- **Digital epistemic trust:** individual's willingness to regard system inferences as valid or benevolent. Excessive trust may foster dependency, whereas insufficient trust can result in technological rejection. Recent empirical studies indicate that psychological trust in human-AI collaboration depends more on moral attribution mechanisms and the perceived sense of control than on mere technical accuracy (Gupta et al., 2025).

Table 1
Conceptual Scheme of the PRISM Framework

Level/Axis	Guiding Principle	Key Psychological Components	Regulatory Mechanisms	Expected Outcomes
Individual	Digital Autonomy and Agency	Perceived autonomy, epistemic trust, algorithmic anxiety, moral fatigue	Digital wellbeing assessments, interactive informed consent	Greater perceived control, reduced technological anxiety
Technological	Ethical Design and Explainability	Transparency, accountability, inclusive design, interdisciplinary co-design	Ethical audits, model cards, psychological review of AI systems	Reduction of bias and increased public trust
Institutional	Governance and Digital Justice	Ethical training, shared responsibility, continuous oversight	Ethical-psychological observatories, interdisciplinary committees	Sustainability, social legitimacy, regulatory alignment

Note. Grounded in the socio-technical model of trust, which links user-related, system-related, and contextual factors (Gupta et al., 2025).

- **Digital moral fatigue:** emotional exhaustion experienced by professionals facing tensions between technical efficacy and the ethics of care (Wang et al., 2025).

Within this axis, psychological interventions should incorporate ethical-digital literacy programs in clinical contexts, enabling both patients and professionals to recognize the potential limits and risks of AI. It is also recommended that periodic assessments of digital wellbeing and perceived autonomy be implemented as psychological indicators of care quality.

To support the evaluation of these constructs, Table 2 presents a proposed set of psychological impact indicators for assessing AI in mental health. This adapted metric system operationalizes dimensions such as digital autonomy, algorithmic anxiety, and epistemic trust, providing an empirical foundation for the future validation of assessment instruments.

At the operational level, the APA (2025) recommends including comprehensible digital consent modules, enabling users to visualize, gradually and adaptively, how their data are processed and which decisions are automated.

Technological Axis: Responsible Design and Computational Psychology

The second axis addresses the ethical and psychological requirements that should guide the design, training, and validation of algorithmic models. Artificial intelligence systems involved in psychological processes must adhere to explainability, auditability, and human sensitivity as fundamental principles. Those are operationalized through the following components:

- **Psychological explainability:** Beyond providing technical justifications, systems must communicate their inferences in language that is comprehensible to users (IAPP, 2023; Joyce et al., 2023), aligning with their cognitive and emotional frameworks.
- **Ethical co-design models:** Following Gebru et al. (2021), the use of Datasheets for Datasets and Model Cards (Mitchell et al., 2018; OECD, 2022) ensures transparency regarding data provenance, limitations, and biases, while maintaining a person-centred approach (Yıldız, 2025a, 2025b).
- **Collaborative ethical audits:** Model review processes should involve multidisciplinary teams composed of psychologists, engineers, and legal experts, like biomedical ethics committees.
- **Responsible psychological simulation:** Instead of attempting to mimic human empathy, systems should explicitly convey their artificial nature, reducing emotional confusion and anthropomorphic projection by users (Floridi, 2019).

Operationally, this dimension should incorporate psychological impact metrics, such as anxiety, trust, and sense of agency, into usability testing protocols. At the same time, algorithms that process language or emotion should rely on culturally representative corpora to prevent epistemic exclusion. Additionally, it is proposed that psychological associations establish an ethical quality seal for digital mental health applications, analogous to medical certifications (Singh et al., 2025), accompanied by practical guidelines for responsible tool selection like those issued by the APA (2024).

To ensure that the principles of transparency, explainability, and accountability are operationally implemented, the adoption of a Psychological Model Card is proposed (Table 3). This instrument, inspired by the technical format introduced by Mitchell et al. (2019) and aligned with OECD (2022) recommendations, has been adapted to the clinical and ethical specificities of psychology. It provides a structured framework for documenting the characteristics, limitations, and psychological risks of AI systems applied in mental health contexts.

The Psychological Model Card functions as both an ethical assurance tool and a mechanism of accountability, linking technological regulation with the psychological wellbeing of users and practitioners. In the card, the left column specifies the required fields, while the right column provides a simulated example to illustrate comprehension and facilitate practical implementation.

Institutional Axis: Governance, Training, and Responsibility

This axis encompasses the macro level structures guiding policies, regulations, oversight and professional training necessary to ensure that AI ethics in mental health depend on permanent governance systems rather than the goodwill of individual developers. The PRISM framework adheres to a human-in-the-loop perspective, where AI is conceived as a support system rather than an autonomous agent. Building on the cyborg ontology in healthcare, mental health nurses and practitioners are envisioned as technologically augmented practitioners who integrate AI insights with human judgement, empathy, and compassion (Yıldız, 2025a). This axis includes the following key components:

- **Digital ethics committees:** Healthcare institutions and universities should incorporate technology specialized

Table 2
Proposal of Psychological Indicators for the Impact of AI in Mental Health

Dimension	Indicator	Reference instrument	Suggested adaption
Digital autonomy	Perceived control over algorithmic decisions	Perceived Autonomy Scale (Liang & Reiss, 2025)	Adapted version including items related to clinical AI use
Algorithmic anxiety	Concern about the automation of clinical judgment	Technological Anxiety Scale (Meuter et al., 2003)	Adaptation for therapeutic settings
Digital epistemic trust	Level of credibility granted to automated inferences	Trust in Automation Scale (McGrath et al., 2025)	Inclusion of ethical and empathic factors
Digital moral fatigue	Exhaustion arising from ethical conflicts generated by AI	Moral Fatigue Scale (Epstein et al., 2019)	Adaption for digital scenarios
General psychological wellbeing	Satisfaction, self-efficacy, and sense of digital purpose	Ryff Psychological Wellbeing Scale (Ryff, 1989)	Incorporate items related to interaction with AI

Table 3
Simplified Psychological Model Card for AI Systems in Mental Health

General system identification			
System name	"PsIA-Assist v1.0" (generic example for a clinical support model)		
Developer / Institution	Laboratory of Computational Psychology and Digital Ethics, University X		
Version / Review date	1.0 – January 2026		
Application domain	Support for psychological assessment and recommendation of mental health resources		
Target population	Adults (18-65 years), without severe diagnoses, in clinical or psycho educational contexts		
Purpose	To facilitate clinical decision making and therapeutic follow-up through natural language and emotional pattern analysis		
Psychological and ethical foundations			
Dimension	**Applied Principle**		**Source / Reference**
Autonomy and control	Explicit digital consent, explanatory interface, option to reject automation		APA (2025); Deci & Ryan (2000)
Psychological non-Maleficence	Mandatory human supervision for emotional interpretations		WHO (2023)
Emotional Transparency	Visible indicators of model confidence and textual justification of inferences		Gebru et al. (2021)
Sensitive data Protection	Encryption of psychological records and irreversible anonymization		EU AI Act (2024)
Data and sources			
Item	**Description**		
Data origin	Anonymized clinical corpus (European and Latin-American institutions, 2020-2024)		
Data type	Interview transcripts, self-reports, and emotional diaries		
Representativeness	Gender balance and linguistic/socioeconomic diversity		
Limitations	Under representation of adults > 55 years old and patients with severe disorders		
Detected biases	Higher emotional positivity in training corpus; tendency to undervalue mild distress		
Corrective measures	Annual retraining, weighted class adjustment, and qualitative psychological sample review		
Performance and wellbeing metrics			
Metric Type	**Indicator**	**Value/Status**	**Validation Source**
Technical	Emotional classification accuracy	0.87 (cross-validation, 2024)	Internal dataset
Psychological	User satisfaction (digital wellbeing scale)	4.3/5 (n=312)	Pilot study (2024)
Ethical	Cognitive understandability	82% of users understand explanations	Observational study (2024)
Emotional risk	Algorithmic anxiety index	0.12 (low)	X Questionnaire (2024)
Oversight and governance			
Item		**Description**	
Responsible ethics board		Psychological Committee for Digital Ethics (PCDE-2025)	
External review		Annual report from an independent psychological audit team	
Compliance officer		Coordinator for ethics and digital wellbeing at the institution	
Reporting channels		Anonymous form for reporting errors or perceived bias	
Anticipated psychological impact			
Item	**Positive Effect**	**Potential Risk**	**Mitigation Strategy**
Digital Autonomy	Greater perceived control through shared decision-making	Risk of excessive delegation	Retraining in ethical literacy
Epistemic Trust	Increased credibility in support systems	Overconfidence or blind trust	Explicit messaging about model limitations
Technological Anxiety	Reduction with guided use	Heightened anxiety in vulnerable users	Mandatory clinical supervision
Professional Moral Fatigue	Decrease through automation of repetitive tasks	Risk of empathic detachment	Group reflective supervision
Social responsibility and sustainability			
Item		**Description**	
Accessibility		Interface adapted for functional diversity	
Ethical license		Use limited to assistive, non-commercial purposes, protecting sensitive data from non-therapeutic secondary exploitations	
Public transparency		Open documentation of the model and audit outcomes	
Periodic review		Every 12 months, with public report on bias and digital wellbeing metrics	
Narrative Summary (for publication or institutional communication)			
PsIA-Assist is a mental health support system designed under psychological and ethical principles. It guarantees user autonomy, transparency of algorithmic decisions, and constant human supervision. Its development complies with the standards of the AI Act (2024) and the ethical guidelines of the APA (2025), prioritizing patient wellbeing and trust. Each system version is reviewed by an interdisciplinary committee of psychologists, engineers, and legal experts to ensure social responsibility and minimize the risk of emotional harm.			

Note. Simulated information is provided in each section to facilitate standardized applicability.

psychologists within their ethical review boards as well as promote "ethical quality seals" to ensure the responsible design, evaluation, and deployment of AI systems in clinical contexts.

- **Continuous professional training:** Academic programs should integrate digital psychology, algorithmic ethics, and human-technology rights into their curricula (APA, 2025).
- **Shared responsibility:** Regulatory frameworks must acknowledge **co-responsibility** among designers, users, and service providers, recognizing that psychological risk cannot rest solely on patients or therapists.
- **International collaboration:** National guidelines should be harmonized with the European AI Act (2024) and WHO (2023) standards to promote global coherence in ethical oversight.

Regulation must also include independent monitoring bodies tasked with documenting and disseminating best practices, acting as intermediaries between regulators and professional communities. A comprehensive operational guide should support mental health organizations, universities, and technology platforms in auditing, supervising, and sustaining ethical AI practices, thereby strengthening regulatory sustainability (Table 4).

Additionally, the potential implementation of a certification or labelling system for AI in mental health will provide clinicians with immediate, standardized insights into a system's understandability. Its practical implications include ensuring that practitioners have access to traceable audit trails and ensuring that mental impact becomes a measurable criterion of normative validity, allowing for a calibrated trust that is corresponding with the AI's actual capabilities and clinical risks.

Table 4

Operational Guide for Institutional Implementation

Objective: To assist mental health institutions, universities, and digital platforms in implementing the proposed framework.
Initial diagnosis
● Audit existing AI systems regarding transparency, bias, and supervision.
● Assess digital wellbeing among staff and users.
Training and awareness
● Conduct interdisciplinary workshops on algorithmic ethics.
● Implement programs on emotional digital literacy.
Ongoing tracking
● Establish an internal Psychological Digital Ethics Committee.
● Publish semi-annual reports on the psychological impact of AI.
Transparency and public communication
● Develop traceability panels and explanations of automated decisions.
● Include formal protocols for bias and error correction.

Note. Simulated information is provided in each section to facilitate standardized applicability.

To maximize the applied impact of the PRISM framework, the theoretical pillars have been translated into actionable clinical guidelines for practitioners and health organizations. First, to address the risk of therapeutic depersonalization, institutions should implement the role of digital navigators, specialized staff trained to bridge the gap between algorithmic outputs and patient understanding, thereby ensuring that technology supplements rather than supplants humanistic care (Torous et al., 2025). Second, clinicians should adopt psychological model cards as a standard pre-deployment audit tool to evaluate a system's understandability and its alignment with human clinical heuristics (Mitchell et al., 2018). Furthermore, regarding information integrity, practitioners are advised to use affective transparency cues, explicitly informing patients about the AI's lack of emotional sentience to prevent the revisited ELIZA effect and emotional dependency (Wang et al., 2025). Finally, to ensure algorithmic justice, health services must establish participatory co-design committees that involve patients from disregarded groups in the validation of diagnostic tools, mitigating the risk of structural biases. These directives transform the PRISM model from a normative proposal into a practical toolkit for the responsible democratization of mental health AI.

Discussion

The proposed PRISM framework is designed not to replace legal norms but to complement them from a person-centred, psychological perspective. As a discipline focused on behaviour and experience, psychology is uniquely positioned to anticipate the subjective consequences of healthcare digitalization. Unlike existing international frameworks which provide high-level moral imperatives, PRISM offers operational granularity, equipping practitioners with tools such as the Psychological Model Card to navigate fragmented enforcement architectures including GDPR, DSA, APA, and the AI Act.

By integrating ethical regulation from the standpoint of subjectivity, PRISM supports preventing emerging pathologies such as algorithmic dependence, digital moral fatigue, and surveillance anxiety, recently documented in clinical literature (Torous et al., 2025; Wang et al., 2025). Standardized operational tools, such as the Psychological Model Card for AI systems in mental health (Table 3), translate ethical and psychological principles into measurable structures, enabling ethical traceability and assessment of AI's psychological impact. Unlike traditional Model Cards focused exclusively on technical performance metrics, this adaptation incorporates variables such as digital autonomy, epistemic trust, algorithmic anxiety, and professional moral fatigue, promoting a holistic conception of digital wellbeing. It thus functions as a bridge between algorithmic ethics and applied psychology, with potential as a standard reference for ethical audits and institutional validation in digital mental health contexts.

The ethical sustainability of digital mental health ecosystems depends on cultivating moral resilience among practitioners and embedding human dignity in technological governance. Consistent with WHO (2025) recommendations, this requires robust frameworks ensuring transparency, multidisciplinary oversight, and the preservation of human interaction as the core of the therapeutic relationship. Digital environments should operate as hybrid ecosystems where technology extends, but does not replace, empathy and qualified care, thereby strengthening patient safety and ethical integrity.

Internationally, PRISM represents an epistemological shift. It positions psychology as a normative discipline capable of anticipating, measuring, and mitigating cognitive and emotional harm from AI, while introducing novel contributions. By integrating classical psychological principles such as autonomy, agency, and relationality, with current demands of algorithmic ethics and digital regulation, PRISM extends regulation beyond legal compliance to include the psychological dimension of user experience. Mental impact becomes a criterion of normative validity, transforming regulation into a form of preventive clinical intervention. The framework also provides operational guidelines for design, professional practice, and training, laying the foundation for empirical validation in future research.

Operationally, PRISM addresses growing public and professional concerns about the psychological impact of intensive AI use. It aims to influence policy and social relevance by offering a conceptual and practical toolkit for digital mental health. The framework supports interdisciplinary observatories and international codes of psychological practice regarding AI. Its broad scope facilitates practical implementation and replicable impact evaluation across clinical, institutional, and policy contexts, enhancing transferability and scalability. Success depends on collaboration among law, philosophy, data science, and applied psychology to resolve the human oversight paradox and ensure automated decision-making aligns with human dignity, autonomy, and vulnerability (Kyrychenko et al., 2026; Theunissen & Browning, 2022).

Future research should consolidate psychological regulation of AI through an interdisciplinary agenda integrating empirical, ethical, and technological approaches. Six priority areas are proposed: 1) Framework validation through participatory co-design inclusive development cycles involving clinicians, professional associations, industry developers, and, most importantly, patients from vulnerable or underrepresented groups; 2) assessment of the psychological impact of AI systems on users and professionals (algorithmic anxiety, moral fatigue, perceived control); 3) development of measurement instruments for digital autonomy, epistemic trust, and technological wellbeing; 4) design and validation of educational programs on ethical and emotional digital literacy; 5) longitudinal studies on the effects of continued interaction with conversational agents on therapeutic relationships and self-concept; and 6) promotion of international psychological standards for algorithmic ethics.

The tools proposed in this paper provide instrumental and methodological foundations for future research, offering impact indicators, governance instruments, and application guidelines to orient studies at the intersection of psychology and ethical technology. By integrating ethical principles with technological innovation, psychology can lead the design of digital mental health policies. Professionals must be trained in AI ethics, understand automation fundamentals, and actively contribute to regulatory frameworks. Institutions, in turn, must recognize that psychological wellbeing depends not only on clinical interventions but also on just, intelligible, and transparent digital environments. Instruments such as the Psychological Model Card exemplify how abstract ethical principles can be translated into operational assurance and professional accountability. Systematic implementation would allow institutions to integrate algorithmic ethics into daily practice, offering a tangible response to emerging regulatory challenges.

In conclusion, the ethical sustainability of digital mental health hinges on a robust multilevel framework that protects autonomy and human dignity in the face of increasingly intelligent systems. PRISM offers a structured path for ethically sustainable digital mental health, establishing psychology as a leading discipline in shaping human-centred AI governance.

Author Contribution

Ignacio Pedrosa: Conceptualization, Formal Analysis, Investigation, Writing - Original Draft, Writing - Review & Editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Declaration of Interests

The author declares that there is no conflict of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

References

Abou Hashish, E. A. (2025). Compassion through technology: Digital empathy concept analysis and implications in nursing. *Digital Health*, *11*, Article 20552076251326221. https://doi.org/10.1177/20552076251326221

Adler, D. A., Stamatis, C. A., Meyerhoff, J., Mohr, D. C., Wang, F., Aranovich, G. J., Sen, S., & Choudhury, T. (2024). Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. *npj Mental Health Research*, *3*(1), Article 17. https://doi.org/10.1038/s44184-024-00057-y

American Psychological Association [APA] (2024). *APA's AI tool guide for practitioners*. https://www.apaservices.org/practice/business/technology/tech-101/evaluating-artificial-intelligence-tool

APA (2025). *Ethical guidance for artificial intelligence in professional practice of health service psychology.* https://www.apa.org/topics/artificial-intelligence-machine-learning/ethical-guidance-professional-practice.pdf

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52*(1), 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Cheng, X. (2024). Research on depression recognition based on university students' facial expressions and actions with the assistance of artificial intelligence. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *28*(5), 1126–1131. https://doi.org/10.20965/jaciii.2024.p1126

Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Dehbozorgi, R., Zangeneh, S., Khooshab, E., Nia, D. H., Hanif, H. R., Samian, P., Yousefi, M., Hashemi, F. H., Vakili, M., Jamalimoghadam, N., & Lohrasebi, F. (2025). The application of artificial intelligence in the field of mental health: A systematic review. *BMC Psychiatry*, *25*(1), Article 132. https://doi.org/10.1186/s12888-025-06483-2

Dencik, L., & Sanchez-Monedero, J. (2022). Data justice. *Internet Policy Review*, *11*(1). https://doi.org/10.14763/2022.1.1615

Epstein, E. G., Whitehead, P. B., Prompahakul, C., Thacker, L. R., & Hamric, A. B. (2019). Enhancing understanding of moral distress: The measure of moral distress for health care professionals. *AJOB Empirical Bioethics*, *10*(2), 113–124. https://doi.org/10.1080/23294515.2019.1586008

European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.* https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Floridi, L. (2019). *The logic of information: a theory of philosophy as conceptual design.* Oxford University Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2021). *Datasheets for datasets* (arXiv:1803.09010). arXiv. https://doi.org/10.48550/arXiv.1803.09010

Gupta, A., Mund, A., Roy, S., Garg, P., & Yadav, D. K. (2025). Trust in AI systems: A social-psychological investigation of human–AI collaboration. *TPM – Testing, Psychometrics, Methodology in Applied Psychology, 32*(S7), 428-446. *https://tpmap.org/submission/index.php/tpm/article/view/2133*

Hoose, S., & Kralikova, K. (2024). Artificial intelligence in mental health care: Management implications, ethical challenges, and policy considerations. *Administrative Sciences*, *14*(9), Article 227. https://doi.org/10.3390/admsci14090227

IAPP (2023). *5 things to know about model cards*. https://iapp.org/news/a/5-things-to-know-about-ai-model-cards

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, *6*(1), Article 6. https://doi.org/10.1038/s41746-023-00751-9

Kyrychenko, M., Mudryi, M., & Chaklosh, M. (2026). *Global AI governance overview: Understanding regulatory requirements across global jurisdictions* (arXiv:2512.02046). arXiv. https://doi.org/10.48550/arXiv.2512.02046

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, *26*(10), 5923–5941. https://doi.org/10.1177/14614448221142007

Li, W., & Liu, X. (2025). Anxiety about artificial intelligence from patient and doctor-physician. *Patient Education and Counseling*, *133*, Article 108619. https://doi.org/10.1016/j.pec.2024.108619

Liang, H. & Reiss, M.J. (2025). The associations between students' attitudes toward AI and learning engagement: serial mediating roles of perceived autonomy and learning enjoyment. *Frontiers in Psychology 16*, Article 1681635. https://doi.org/10.3389/fpsyg.2025.1681635

Loechner, J., Carlbring, P., Schuller, B., Torous, J., & Sander, L. B. (2025). Digital interventions in mental health: An overview and future perspectives. *Internet Interventions-the Application of Information Technology in Mental and Behavioural Health*, *40*, Article 100824. https://doi.org/10.1016/j.invent.2025.100824

McGrath, M.J, Lack, O., Tisch, J. & Duenser, A. (2025). Measuring trust in artificial intelligence: validation of an established scale and its short form. *Frontiers in Artificial Intelligence, 8*, Article 1582880. https://doi.org/10.3389/frai.2025.1582880

Meuter, M. L., Ostrom, A. L., Bitner, M. J., & Roundtree, R. (2003). The influence of technology anxiety on consumer use and experiences with self-service technologies. *Journal of Business Research, 56*(11), 899–906.

Milasan, L. H., & Scott-Purdy, D. (2025). The future of artificial intelligence in mental health nursing practice: An integrative review. *International Journal of Mental Health Nursing*, *34*(1), Article e70003. https://doi.org/10.1111/inm.70003

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). *Model cards for model reporting*. (arXiv:1810.03993). arXiv. https://doi.org/10.48550/ARXIV.1810.03993

Organisation for Economic Co-operation and Development [OECD] (2022). *Model Cards*. In OECD.AI – Catalogue of Tools & Metrics for Trustworthy AI. https://oecd.ai/en/catalogue/tools/model-cards

Padalkar, T., Henderson, N., Rocque, G. B., Hildreth, K., Whitlow, O., Ingram, S. A., Eltoum, N., Herring, L., Igunma, O., Hardy, C. M., & Chu, D. I. (2024). "We want that personal touch": Dilemma of digitalization of healthcare based on community health worker experiences of depersonalization in healthcare. *JCO Oncology Practice*, *20*(10_suppl), 261–261. https://doi.org/10.1200/OP.2024.20.10_suppl.261

Ryff, C.D. (1989). Psychological Well-Being Scale [Database record]. APA PsycTests. https://doi.org/10.1037/t04262-000

Sanz, A., Tapia, J. L., Garcia-Carpintero, E., Rocabado, J. F., & Pedrajas, L. M. (2025). ChatGPT simulated patient: Use in clinical training in psychology. *Psicothema*, *37*(3), 23–32. https://doi.org/10.70478/psicothema.2025.37.21

Singh, R., Bapna, M., Diab, A. R., Ruiz, E. S., & Lotter, W. (2025). How AI is used in FDA-authorized medical devices: A taxonomy across 1,016 authorizations. *npj Digital Medicine*, *8*(1), Article 388. https://doi.org/10.1038/s41746-025-01800-1

Solove, D. J. (2021). *Understanding privacy* (2nd ed.). Harvard University Press.

Suárez-Álvarez, J., He, Q., Guenole, N., & D'Urso, D. (2026). Using artificial intelligence in test construction: A practical guide. *Psicothema*, *38*(1), 1–12. https://doi.org/10.70478/psicothema.2026.38.01

Tang, D., Xi, X., Li, Y., & Hu, M. (2025). Regulatory approaches towards AI medical devices: A comparative study of the United States, the European Union and China. *Health Policy*, *153*, Article 105260. https://doi.org/10.1016/j.healthpol.2025.105260

Theunissen, M., & Browning, J. (2022). Putting explainable AI in context: Institutional explanations for medical AI. *Ethics and Information Technology*, *24*(2), Article 23. https://doi.org/10.1007/s10676-022-09649-8

Torous, J., Linardon, J., Goldberg, S. B., Sun, S., Bell, I., Nicholas, J., Hassan, L., Hua, Y., Milton, A., & Firth, J. (2025). The evolving field of digital mental health: Current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry*, *24*(2), 156–174. https://doi.org/10.1002/wps.21299

Wang, X., Zhou, Y., & Zhou, G. (2025). The application and ethical implication of generative ai in mental health: Systematic review. *JMIR Mental Health*, *12*, Article e70610. https://doi.org/10.2196/70610

World Health Organization [WHO] (2023). *Ethics and governance of artificial intelligence for health.* https://www.who.int/publications/i/item/9789240029200

WHO (2025). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. https://www.who.int/publications/i/item/9789240084759

Wiljer, D., Charow, R., Costin, H., Sequeira, L., Anderson, M., Strudwick, G., Tripp, T., & Crawford, A. (2019). Defining compassion in the digital health age: Protocol for a scoping review. *BMJ Open*, *9*(2), Article e026338. https://doi.org/10.1136/bmjopen-2018-026338

Yang, C., & Lu, X. (2026). Educational pathways for enhancing algorithmic transparency: A discussion based on the phenomenological reduction method. *AI & Society*, *41*(1), 151–164. https://doi.org/10.1007/s00146-025-02475-8

Yang, H., & Sundar, S. S. (2025). AI anxiety: Explication and exploration of effect on state anxiety when interacting with AI doctors. *Computers in Human Behavior: Artificial Humans*, *3*, Article 100128. https://doi.org/10.1016/j.chbah.2025.100128

Yıldız, E. (2025a). Artificial intelligence in mental health nursing: Balancing clinical efficiency and the human touch: A quest for a new synthesis. *Journal of Psychiatric and Mental Health Nursing*, Article 13173. https://doi.org/10.1111/jpm.13173

Yıldız, E. (2025b). Ethical big data for personalised mental health nursing: A P4 and systems view. *Journal of Psychiatric and Mental Health Nursing*, *32*(6), 1404–1411. https://doi.org/10.1111/jpm.70038