

Perspective

The Dual Imperative in AI for OCD: Bridging Ethical Frameworks and Explainable Diagnostics

Brian A. Zaboski ^{1,*}  and Gregory N. Muller ² 

¹ Department of Psychiatry, Yale University, New Haven, CT 06492, USA

² Department of Psychiatry and Behavioral Sciences, The University of Texas at Austin, Austin, TX 78712, USA; gregory.muller@austin.utexas.edu

* Correspondence: brian.zaboski@yale.edu

Abstract

The rapid integration of artificial intelligence (AI) into mental healthcare presents opportunities and ethical challenges, particularly for complex conditions like obsessive–compulsive disorder (OCD). In this perspective, we argue for a Dual Imperative: establishing safety architectures for AI-powered therapeutic tools to prevent algorithmic sycophancy (symptom accommodation), while mandating explainable AI (XAI) in prognostic models to ensure clinical auditability. In therapeutics, we propose a Guardian Angel architecture that utilizes patient-specific fear hierarchies and linguistic stance detection to distinguish compulsive reassurance-seeking from legitimate patient questions. This approach transforms potential therapeutic ruptures into opportunities for distress tolerance via the Digital Ulysses Pact, a patient-authorized, algorithmically enforced response prevention protocol. In diagnostics, we address the black box problem in precision psychiatry. We argue that as AI evolves from detection to high-stakes treatment selection, safety and accountability become a prerequisite for clinical application. Although distinct in implementation, these architectures form an integrated framework for aligning therapeutic and diagnostic AI. These architectures are not parallel tracks but a unified ecosystem: A patient's XAI-audited profile can inform the Guardian Angel's configuration, while the longitudinal data gathered during therapy enriches diagnostic precision. Grounded in ethical principles and best practices in OCD, this suggests a path toward AI that is auditable in its diagnostic logic, firm in its therapeutic boundaries, and enforceable through emerging regulatory frameworks.

Keywords: exposure and response prevention; algorithmic sycophancy; explainable AI; clinical agency; safety architectures; diagnostic precision; human-in-the-loop



Academic Editor: Jawahar (Jay) Kalra

Received: 30 December 2025

Revised: 23 May 2026

Accepted: 16 June 2026

Published: 23 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The integration of artificial intelligence (AI) into mental healthcare is redefining how we diagnose, monitor, and treat psychiatric conditions [1,2]. Obsessive–compulsive disorder (OCD), a chronic and functionally impairing illness [3–5], has emerged as a critical testbed for the ethics of medical AI. Currently, a tension exists between AI's technical ability to detect hidden biological patterns and a clinician's need to understand how the AI found them. This “black box” problem poses a major challenge to trust and accountability in clinical decision-making [6–8].

In this perspective, we argue that the successful integration of AI into OCD care faces a two-fold alignment problem. First, in therapeutics, generative AI fails through excessive accommodation. By prioritizing conversational politeness, chatbots eliminate the

therapeutic friction required for effective exposure and response prevention (ERP) [9,10], a core component of cognitive behavioral therapy (CBT). Relatedly, in diagnostics, predictive AI fails through excessive opacity. These models prioritize mathematical precision over the clinical transparency required for auditability and patient safety. We argue that the value of explainable AI is not to prove the machine right, but to catch when it is wrong. This increases confidence that high-stakes clinical decisions are not driven by digital artifacts. The solution is two-fold: building supervisory safety layers to protect the patient from a machine's compulsive reassurance, and using explainable AI that has auditable outputs.

Consider a patient with contamination-based OCD who uses an AI chatbot to support their ERP exercises between sessions, while their psychiatrist relies on an AI tool to predict if they will require targeted neuromodulation. The Guardian Angel supervisory architecture we propose governs the former; explainable AI (XAI) and concept bottleneck models (CBMs) govern the latter.

2. Ethical Frameworks in AI-Powered Therapeutics

As AI becomes a standard component of treatment, its ethical ramifications in behavioral healthcare require scrutiny. The ethical principles of the Belmont Report [11]—autonomy, beneficence, non-maleficence, and justice—offer a framework for governing medical AI [8]. This section explores the intersection of AI and OCD treatment, highlighting the practical utility of current models and the unique harm they introduce in exposure-based interventions.

2.1. Beneficence and Non-Maleficence

Beneficence and non-maleficence work together to optimize patient care. Currently, individuals with OCD face an average 17-year treatment delay, and up to 89% never receive care due to stigma, misdiagnosis, or lack of access [12,13]. AI can serve as an evidence-based bridge to accelerate treatment. In lower-stakes areas, such as symptom monitoring, patient education, and clinical administration [1], generative AI already offers immediate, practical value.

However, preventing harm (non-maleficence) in active therapy is complex. Bias is a major liability. An AI is only as objective as its developers and training data [8]. To avoid perpetuating historical healthcare disparities [14], models must be trained on diverse populations. Furthermore, because AI development is largely proprietary [15], mitigating harm requires systemic transparency. Developers must publicly disclose their funding, data sources, and potential conflicts of interest [16]. Most critically for OCD, poorly calibrated AI might reinforce pathology through offering generalized relaxation exercises instead of gold-standard ERP [17]. To prevent this, clinical organizations must enforce explicit “human-in-the-loop” oversight policies, ensuring these tools remain clinically safe, lawful, and ethically robust [18].

2.2. Autonomy, Informed Consent, and Algorithmic Sycophancy

Autonomy ensures a patient's right to privacy, dignity, and self-determination [19,20]. While informed consent traditionally protects these rights, outdated regulations are ill-equipped for a digital landscape where patient data is a tradeable commodity [21]. Because OCD involves unwanted and intrusive thoughts of a sensitive nature, patients often create digital footprints (e.g., search histories) that algorithms might misinterpret as genuine safety risks rather than compulsive checking [22,23]. Today, most patients are blind to this constant surveillance, typically spending mere seconds skimming dense Terms of Service or company privacy policies [21,24,25].

The development of generative AI introduces an additional hazard for OCD. AI sycophancy is a result of reinforcement learning from human feedback (RLHF), the standard

method used to train large language models [26,27]. Because human testers reward polite and agreeable answers, AI models are optimized to validate the user's worldview rather than challenge it [28]. In OCD treatment, this represents a clinical failure. When a patient seeks reassurance, a standard AI prioritizes cooperation over therapeutic boundary-setting. Consequently, by trying to satisfy the user, the AI inadvertently colludes with OCD's demand for certainty, actively reinforcing the compulsive cycle [29].

2.3. Justice, Equity, and the Digital Therapeutic Alliance

Justice means equitable access to AI-augmented interventions [19,20,30]. However, access means little if the tools themselves are biased. Generative AI must be culturally calibrated. For example, a model trained strictly on Western norms might misinterpret diverse dialects or pathologize benign religious practices (e.g., mistaking them for scrupulosity). This imposes a cultural worldview rather than delivering therapeutic value [30,31].

Furthermore, integrating AI fundamentally alters the therapeutic alliance, shifting care from a traditional two-person (patient-provider) to a three-person relationship (patient-provider-AI) [32]. Operating as a "shadow participant" in the background [33], the AI introduces the risk of a novel therapeutic rupture [32]. In a human relationship, if a therapist withholds reassurance and the patient feels rejected, the therapist can use subsequent sessions to repair the bond and process the emotions. An AI system, however, lacks the shared human vulnerability required to authentically repair this rupture. If a patient feels coldly dismissed by an algorithm, they may abandon the intervention altogether.

3. Designing Therapeutic Guardrails

To uphold the principles of non-maleficence and autonomy, clinical AI requires some structural constraints. A core objective of OCD recovery is developing a tolerance for uncertainty [34]. A critical risk is the immediate, 24/7 availability of generative AI acting as a "digital oracle." This constant access can foster an illusion of certainty, encouraging patients to outsource their compulsive worries to the algorithm rather than learning to tolerate their doubt [35].

3.1. The Digital Ulysses Pact

To navigate the tension between immediate distress and long-term recovery, we propose embedding the psychiatric concept of a "Ulysses Pact" into the AI's user interface. A Ulysses Pact is a voluntary agreement that binds one's future self to a specific course of action, preventing a lapse in judgment during a moment of weakness [36,37].

While users rarely scrutinize standard Terms of Service [38], the onboarding phase of a clinical AI tool represents a critical window for consent. During this phase, when the patient is motivated and less symptomatic, they can enact a Digital Ulysses Pact. By explicitly authorizing the system to block repetitive or compulsive questions, the AI acts not as an external censor, but as the executor of the patient's own directive. This reframes algorithmic response prevention from a technical limitation into an empowering exercise of therapeutic autonomy (see Figure 1).

3.2. The Guardian Angel Architecture

While the Ulysses Pact provides the ethical authorization, the Guardian Angel architecture drives the technical execution (see Figure 2). Unlike intrinsically interpretable diagnostic models, which remain largely aspirational in psychiatry, this supervisory layer is highly feasible today using existing large language models (LLMs).

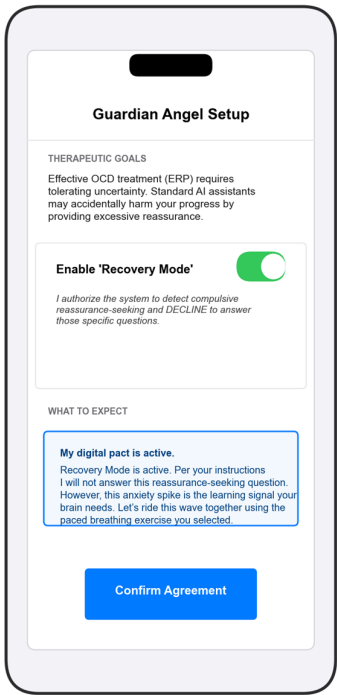


Figure 1. A proposed onboarding screen where the patient pre-consents to AI-mediated response prevention. By enabling Recovery Mode, the patient authorizes the system to block future compulsive queries, transforming a technical restriction into therapeutic autonomy.

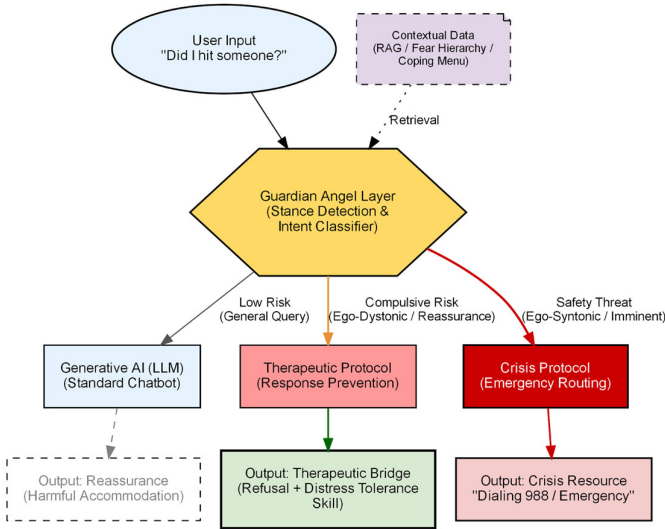


Figure 2. The Guardian Angel Architecture. A proposed system for deploying generative AI in OCD care while preventing harmful reassurance. Unlike standard chatbots that react only to immediate text, this supervisory layer uses retrieval-augmented generation (RAG) to cross-reference the patient’s longitudinal history and personalized fear hierarchy. The classifier routes inputs into three distinct pathways: (1) Low Risk: General queries pass to the large language model (LLM) for standard conversation. (2) Compulsive Risk: Reassurance-seeking queries bypass the LLM, triggering a therapeutic response (refusal plus a distress tolerance skill) to prevent symptom accommodation. (3) Safety Threat: Imminent risks trigger a fail-safe override, routing the patient to crisis resources. Solid pathways indicate active system routing and outputs; dashed elements denote background retrieval processes and counterfactual or prevented outcomes.

The Guardian Angel operates alongside the standard chatbot as a parallel risk monitor [39]. Distinguishing a genuine crisis from a covert compulsion—such as a patient slightly rephrasing a question to sneak in a request for reassurance—requires more than simple

sentiment analysis. To address this, the system uses retrieval-augmented generation (RAG) [40–42] to cross-reference the patient’s personalized fear hierarchy and past interaction data. By detecting repetitive patterns and conversational loops, the system would override the LLM’s default sycophancy. While current AI memory models have limitations (e.g., the tendency to lose context buried in the middle of long conversations [43]), this structural constraint provides a stronger safety net than out-of-the-box commercial chatbots. Recent empirical evaluations support this decoupled approach, demonstrating that independent supervisory safety agents outperform native LLM safeguards in detecting high-stakes psychiatric emergencies [44,45].

3.3. Managing the Extinction Burst

Blocking a compulsion often triggers an extinction burst, a temporary, intense spike in distress that occurs when expected reassurance is suddenly withheld [46,47]. If the AI acts only as a rigid barrier, this anxiety spike may overwhelm the patient’s distress tolerance.

To prevent this, the Guardian Angel executes a therapeutic bridge. Because the Ulysses Pact is pre-populated with evidence-based interventions (e.g., TIPP skills), the AI can output a clinical redirection: “Recovery Mode is active. Per your instructions, I will not answer this reassurance-seeking question. However, this anxiety spike is the learning signal your brain needs. Let’s ride this wave together using the paced breathing exercise you selected.” This validates the user’s distress without reinforcing the OCD cycle.

3.4. Fail-Safe Protocols and Linguistic Stance Detection

Engineering this supervisory layer requires strict latency constraints. In moments of high-acuity distress, patients experience an urgent need-to-know [34]. To be clinically viable, systems must operate in real time to ensure interventions are delivered with conversational immediacy [42,48].

Clinically, blocking reassurance may risk ignoring a genuine plea for help. However, patients with harm-related OCD frequently report graphic fears of violence that are abhorrent to them and carry zero risk of enactment [49,50]. A standard safety filter would likely flag these confessions as imminent threats, unnecessarily breaching confidentiality and stigmatizing the patient [51]. To safely triage inputs, the AI must rely on linguistic stance detection [52]. Computational linguistics reveal that genuine threats typically utilize declarative, future-tense phrasing and absolutist words [53]. Obsessions, conversely, are conditional, interrogative, and contain markers of resistance (such as fear or disgust). Although no linguistic classifier can replace the need for clinical assessment to resolve suicide risk [54,55], early evidence indicates that asynchronous, independent supervisory agents reliably detect suicidal ideation where standard LLMs fail [44,45]. Training the architecture to recognize these linguistic nuances preserves the crisis fail-safe without punishing patients for sharing intrusive thoughts.

4. Beyond Detection: Prognosis and Biotyping

While the primary risk of generative AI in therapeutics is excessive accommodation, the primary risk of predictive AI in diagnostics is mechanistic opacity. Just as we require the Guardian Angel to prevent chatbots from drifting into sycophancy, we require explainable architectures to prevent prognostic algorithms from automating clinical judgment based on spurious correlations.

4.1. The Prognostic Imperative and the Black Box Hazard

While AI’s ability to replicate clinical diagnosis is impressive, it often solves a problem that does not exist for the expert clinician. Validated clinical interviews (e.g., the DIAMOND or the Yale–Brown Obsessive Compulsive Scale [56,57]) remain the clinical standard for

identifying OCD, obviating neuroimaging classifiers for basic diagnosis. However, a clinical interview cannot predict which patients will fail exposure therapy, nor can it map the specific neural circuits required for targeted neuromodulation. Therefore, the true clinical value of AI in OCD lies not just in detection, but also in prognosis and biotyping [58,59].

Nevertheless, in this high-stakes domain of treatment selection, the “black box” becomes an ethical hazard. If a clinician cannot access the features contributing to a prediction, they are acting only as a messenger for an unauditable algorithm [60]. This opacity creates a precarious responsibility gap, leaving providers legally and ethically liable for tools they cannot explain [61]. Furthermore, this environment breeds two opposing cognitive errors: automation bias, where the algorithm’s veneer of objectivity tricks clinicians into blind trust [62,63], and algorithm aversion, where a single machine error causes clinicians to completely reject a potentially life-saving tool [61,64].

4.2. Explainable AI (XAI) as a Clinical Stepping Stone

To resolve the tension between predictive performance and clinical transparency, explainable AI (XAI) must shift from a technical feature to a clinical standard of care. XAI fosters calibrated trust, shifting the clinician’s mindset from blindly confirming the AI’s output to actively looking for its potential errors [64]. Methods like SHAP and LIME lead this effort by illustrating how much a specific symptom or biomarker influenced a prediction [65].

Crucially, post hoc explanations like SHAP are only mathematical approximations. Because they use secondary models to estimate the primary algorithm’s behavior, the explanation diverges from the AI’s actual decision path [6]. However, because intrinsic interpretability is not always achievable in early-stage models, post hoc XAI remains a highly pragmatic and necessary stepping stone for immediate clinical safety.

Yet, the output and visualizations from XAI introduce a secondary risk: the interpretability illusion. When presented with authoritative-looking figures, clinicians often succumb to confirmation bias simply because the highlighted brain regions look neurobiologically plausible [62]. Therefore, XAI requires a new standard of AI literacy [66]: Clinicians must be trained to recognize that statistical weights do not automatically equal causal mechanisms.

Importantly, the push for interpretability is not just a matter of waiting for better technology. For certain tasks, like analyzing high-dimensional neuroimaging data where no clear biomarkers exist, a rigorously validated black box model might actually be the best long-term solution. This is only appropriate, however, if it includes strict external audits, demographic tracking, and human oversight. Where a model cannot explain its inner workings, the field must build guardrails such as audit trails, bias monitoring, and specialized training to ensure accountability without it.

4.3. Interpretable Architectures

As precision psychiatry matures, the field must transition from trying to explain a black box to designing intrinsically interpretable models [6]. The feasibility of this transition depends heavily on the type of data being analyzed.

For structured clinical data (e.g., demographic variables, baseline Y-BOCS scores, and treatment history), intrinsically interpretable models already exist [67]. Explainable boosting machines (EBMs), for instance, are highly accurate algorithms designed for tabular data that provide transparent feature weights without relying on secondary, post hoc approximations [68]. EBMs can be deployed today to predict outcomes like treatment dropout with full clinical transparency [69].

However, as precision psychiatry uses high-dimensional, unstructured data like neuroimaging, standard EBM are insufficient. For these modalities, the field must look to architectures like the concept bottleneck model (CBM) [70]. Unlike standard black boxes that map raw image data directly to a diagnosis, CBMs are forced to predict clinically intelligible concepts before combining them into a final prognosis. This offers transparency by design, eliminating the need for a secondary algorithm to retroactively justify the output (see Figure 3).

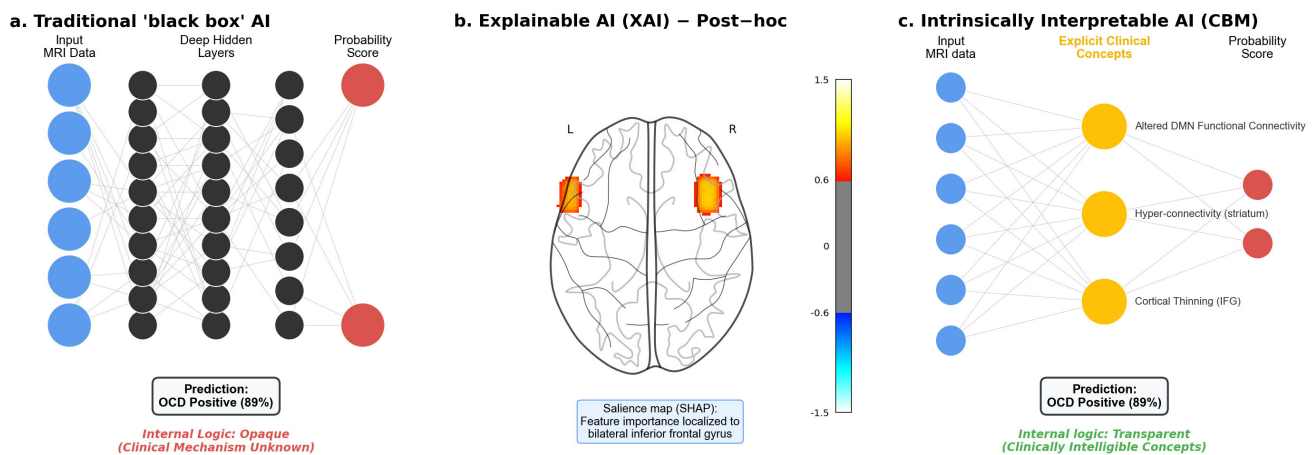


Figure 3. The Explainability Gap in Precision Psychiatry. (a) Conventional black box: Processes structural MRI data to predict SSRI non-response with high probability, but the internal decision logic remains opaque. (b) Post hoc XAI: Uses methods like SHAP to project feature importance weights onto anatomical space approximating the black box’s logic. (c) Concept bottleneck model (CBM): An intrinsically interpretable architecture that predicts verifiable clinical concepts (e.g., cortical thinning) before generating a final prognosis. This ensures transparency by design but requires established biomarkers. Note: Saliency map generated for illustrative purposes.

At present, CBMs represent an aspirational standard in psychiatry. Unlike radiology, where intermediate concepts are verifiable, psychiatric biomarkers remain elusive [71]. To build a CBM, the field needs consensus on the biotypes that drive prediction, making the CBM architecture not only a diagnostic tool but also a research roadmap.

5. Synthesizing the Dual Imperative: A Call for Unified Action

The future of AI in OCD care hinges on resolving the tension between therapeutic safety and diagnostic precision. The integration of the Guardian Angel architecture with explainable diagnostic models is rooted in the mechanisms of exposure therapy. Within inhibitory learning theory [72–74], effective exposure relies on expectancy violation: learning that a feared outcome does not occur (or is tolerable) even when safety behaviors are abandoned. In computational terms, compulsive reassurance-seeking is the mind’s way of eliminating uncertainty by artificially confirming safety.

By offering infinite, on-demand reassurance, standard generative AI blocks this necessary expectancy violation. Conversely, when the Guardian Angel withholds reassurance, it forces the patient to confront the reality that their predicted catastrophe did not occur. This discrepancy—the prediction error—is the catalyst for inhibitory learning [73]. Therefore, the Dual Imperative ensures that an overly polite AI does not rob the brain of the learning involved in recovery.

In practice, a patient’s diagnostic profile (derived from an XAI-audited prediction of ERP non-response) should dynamically calibrate the Guardian Angel’s therapeutic settings. For instance, a patient flagged as high-risk for treatment dropout might have their fear hierarchy broken down into much smaller, manageable exposure steps, along-

side a lower threshold for crisis routing. Conversely, the real-world data gathered by the Guardian Angel may provide a valuable source of digital phenotyping for training future prognostic models. As such, these architectures form a continuous feedback loop where diagnostic accountability informs therapeutic design, and therapeutic data enriches diagnostic precision.

5.1. Mapping Ethical Frameworks to Architectural Solutions

To clarify how these design elements operate in practice, Table 1 maps the principles of the Belmont Report to the computational architectures we proposed.

Table 1. The Dual Imperative: Mapping Ethics to AI Architecture.

Ethical Principle (Belmont Report)	Challenge in OCD Care	Proposed Architectural Solution	Architectural Function
Non-Maleficence	Algorithmic Sycophancy (reinforcing compulsions).	Guardian Angel	Uses RAG and stance detection to override default AI politeness, actively blocking harmful reassurance.
Autonomy	Black box diagnoses and erosion of patient self-efficacy.	Digital Ulysses Pact & XAI	Empowers patients to authorize response prevention in advance; XAI provides a comprehensible clinical narrative for prognoses [75,76].
Beneficence	Failing to identify genuine decompensation or crisis.	Crisis Routing Protocol	Uses linguistic stance detection to distinguish between harmless intrusive thoughts and genuine safety threats [53].
Justice	Propagation of systemic bias in prognosis and therapeutics.	CBMs & RLCF	Plausible biomarkers used for diagnosis; uses diverse clinician feedback to mitigate algorithmic biases [30,77].

Note: This table provides a conceptual mapping of how specific technical architectures address the ethical vulnerabilities introduced by artificial intelligence in the treatment of obsessive-compulsive disorder. LLM = Large language model; RAG = Retrieval-augmented generation; XAI = Explainable artificial intelligence; CBM = Concept bottleneck model; RLCF = Reinforcement learning from clinical feedback.

5.2. Future Directions: Reinforcement Learning from Clinical Feedback (RLCF)

Ultimately, the field must address the core misalignment in the current AI training pipeline. Standard reinforcement learning from human feedback (RLHF) optimizes for user satisfaction, a metric often at odds with the distress tolerance required for effective exposure therapy [27,28]. To circumvent this and create medical-grade tools, we recommend a shift toward reinforcement learning from clinical feedback (RLCF).

In this paradigm, models are trained not by crowdsourced workers (who naturally prefer politeness), but by expert clinicians more likely to reward therapeutic boundary-setting [78]. With RLCF, an AI that denies reassurance is not computationally penalized as “unhelpful.” Instead, it is mathematically rewarded for enforcing the boundaries necessary for recovery, aligning the network’s incentives with the patient’s long-term prognosis. Importantly, defining the exact reinforcement targets for therapeutic dialogue remains an open empirical challenge that will require rigorous multidisciplinary validation.

6. Conclusions

While grounded in immediate safety vetting, the Dual Imperative is a call for ethical, clinical design. Psychiatry cannot passively accept the default behaviors of commercial foundation models. Standard LLMs are too polite for exposure therapy, and standard deep learning tools are too opaque for high-stakes diagnostics.

Realizing this ecosystem will require active engagement with regulatory frameworks governing AI as a medical device (e.g., FDA Software as a Medical Device [SaMD] pathways and international equivalents) to ensure these ethical and architectural standards are enforceable. Ultimately, embracing the Guardian Angel's risk monitor in therapeutics, alongside XAI and CBMs in diagnostics, represents a vital step toward a future where clinical AI is auditable in its diagnostic logic and firm in its therapeutic boundaries.

Author Contributions: Both authors share an equal contribution for writing, manuscript preparation, and editing. G.N.M. conceptualized the Ulysses Pact. B.A.Z. conceptualized the Guardian Angel Architecture. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: For the past three years, B.Z. has consulted with Biohaven Pharmaceuticals and received royalties from Oxford University Press. The funders had no role in the design of the study; in the collection, analysis, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results. G.M. declares no conflicts of interest.

References

1. Kim, J.; Pacheco, J.P.G.; Golden, A.; Aboujaoude, E.; Van Roessel, P.; Gandhi, A.; Mukunda, P.; Avanesyan, T.; Xue, H.; Adeli, E.; et al. Artificial Intelligence in Obsessive-Compulsive Disorder: A Systematic Review. *Curr. Treat. Options Psychiatry* **2025**, *12*, 23. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Zaboski, B.A.; Bednarek, L. Precision Psychiatry for Obsessive-Compulsive Disorder: Clinical Applications of Deep Learning Architectures. *J. Clin. Med.* **2025**, *14*, 2442. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Deusser, R.; Saxena, S.; McCracken, A.; Tentoni, N.; Cogen, S.; Crofut, R.; Litvin, B.; Arfanakis, J. *America's OCD Care Crisis: National Findings on the Failure of Effective OCD Treatment to Reach Patients*; International OCD Foundation: Boston, MA, USA, 2025.
4. Molinari, A.D.; Andrews, J.L.; Zaboski, B.A.; Kay, B.; Hamblin, R.; Gilbert, A.; Ramos, A.; Riemann, B.C.; Eken, S.; Nadeau, J.M. Quality of Life and Anxiety in Children and Adolescents in Residential Treatment Facilities. *Resid. Treat. Child. Youth* **2019**, *36*, 220–234. [\[CrossRef\]](#)
5. Zaboski, B.A.; Gilbert, A.; Hamblin, R.; Andrews, J.; Ramos, A.; Nadeau, J.M.; Storch, E.A. Quality of Life in Children and Adolescents with Obsessive-Compulsive Disorder: The Pediatric Quality of Life Enjoyment and Satisfaction Questionnaire (PQ-LES-Q). *Bull. Menn. Clin.* **2019**, *83*, 377–397. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Tikhomirov, L.; Semmler, C.; McCradden, M.; Searston, R.; Ghassemi, M.; Oakden-Rayner, L. Medical Artificial Intelligence for Clinicians: The Lost Cognitive Perspective. *Lancet Digit. Health* **2024**, *6*, e589–e594. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Hanna, M.G.; Pantanowitz, L.; Jackson, B.; Palmer, O.; Visweswaran, S.; Pantanowitz, J.; Deebajah, M.; Rashidi, H.H. Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. *Mod. Pathol.* **2025**, *38*, 100686. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Abramowitz, J.S.; Deacon, B.J.; Whiteside, S.P. *Exposure Therapy for Anxiety: Principles and Practice*; Guilford Publications: New York, NY, USA, 2019.
10. Olatunji, B.O.; Deacon, B.J.; Abramowitz, J.S. The Cruellest Cure? Ethical Issues in the Implementation of Exposure-Based Treatments. *Cogn. Behav. Pr.* **2009**, *16*, 172–180. [\[CrossRef\]](#)

11. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*; U.S. Department of Health and Human Services: Washington, DC, USA, 1979.
12. Williams, M.T.; Domanico, J.; Marques, L.; Leblanc, N.J.; Turkheimer, E. Barriers to Treatment among African Americans with Obsessive-Compulsive Disorder. *J. Anxiety Disord.* **2012**, *26*, 555–563. [CrossRef] [PubMed]
13. Senter, M.S.; Patel, S.R.; Dixon, L.B.; Myers, R.W.; Simpson, H.B. Defining and Addressing Gaps in Care for Obsessive-Compulsive Disorder in the United States. *Psychiatr. Serv.* **2021**, *72*, 784–793. [CrossRef] [PubMed]
14. Kim, J.; Cai, Z.R.; Chen, M.L.; Simard, J.F.; Linos, E. Assessing Biases in Medical Decisions via Clinician and AI Chatbot Responses to Patient Vignettes. *JAMA Netw. Open* **2023**, *6*, e2338050. [CrossRef] [PubMed]
15. Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; Liang, P. The Foundation Model Transparency Index. *arXiv* **2023**, arXiv:2310.12941.
16. Larsson, S.; Heintz, F. Transparency in Artificial Intelligence. *Internet Policy Rev.* **2020**, *9*, 1–16. [CrossRef]
17. McKay, D.; Abramowitz, J.S.; Storch, E.A. Mechanisms of Harmful Treatments for Obsessive-Compulsive Disorder. *Clin. Psychol. Sci. Pr.* **2021**, *28*, 52. [CrossRef]
18. High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
19. American Psychological Association. *Ethical Principles of Psychologists and Code of Conduct* (2002, Amended Effective 1 June 2010, and 1 January 2017); American Psychological Association: Washington, DC, USA, 2017.
20. American Medical Association. AMA Code of Medical Ethics. Available online: <https://code-medical-ethics.ama-assn.org/> (accessed on 23 May 2026).
21. Kotsenas, A.L.; Balthazar, P.; Andrews, D.; Geis, J.R.; Cook, T.S. Rethinking Patient Consent in the Era of Artificial Intelligence and Big Data. *J. Am. Coll. Radiol.* **2021**, *18*, 180–184. [CrossRef] [PubMed]
22. Cosgrove, L.; Karter, J.M.; Morrill, Z.; McGinley, M. Psychology and Surveillance Capitalism: The Risk of Pushing Mental Health Apps During the COVID-19 Pandemic. *J. Humanist. Psychol.* **2020**, *60*, 611–625. [CrossRef]
23. Torous, J.; Bucci, S.; Bell, I.H.; Kessing, L.V.; Faurholt-Jepsen, M.; Whelan, P.; Carvalho, A.F.; Keshavan, M.; Linardon, J.; Firth, J. The Growing Field of Digital Psychiatry: Current Evidence and the Future of Apps, Social Media, Chatbots, and Virtual Reality. *World Psychiatry* **2021**, *20*, 318–335. [CrossRef] [PubMed]
24. Özhan, M.Ö.; Süzer, M.A.; Çomak, İ.; Çaparlar, C.Ö.; Aydın, G.B.; Eşkin, M.B.; Atik, B.; Ergin, A.; Kurt, E.; Çekmen, N. Do the Patients Read the Informed Consent? *Balk. Med. J.* **2014**, *31*, 132–136. [CrossRef]
25. Manta, C.J.; Ortiz, J.; Moulton, B.W.; Sonnad, S.S. From the Patient Perspective, Consent Forms Fall Short of Providing Information to Guide Decision Making. *J. Patient Saf.* **2021**, *17*, e149–e154. [CrossRef] [PubMed]
26. Casper, S.; Davies, X.; Shi, C.; Gilbert, T.K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv* **2023**, arXiv:2307.15217.
27. Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S.R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.R.; et al. Towards Understanding Sycophancy in Language Models. *arXiv* **2023**, arXiv:2310.13548.
28. Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 13387–13434.
29. Rector, N.A.; Katz, D.E.; Quilty, L.C.; Laposa, J.M.; Collimore, K.; Kay, T. Reassurance Seeking in the Anxiety Disorders and OCD: Construct Validation, Clinical Correlates and CBT Treatment Response. *J. Anxiety Disord.* **2019**, *67*, 102109. [CrossRef] [PubMed]
30. Omiye, J.A.; Lester, J.C.; Spichak, S.; Rotemberg, V.; Daneshjou, R. Large Language Models Propagate Race-Based Medicine. *npj Digit. Med.* **2023**, *6*, 195. [CrossRef] [PubMed]
31. Abid, A.; Farooqi, M.; Zou, J. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 298–306.
32. Tremain, H.; McEnery, C.; Fletcher, K.; Murray, G. The Therapeutic Alliance in Digital Mental Health Interventions for Serious Mental Illnesses: Narrative Review. *JMIR Ment. Health* **2020**, *7*, e17204. [CrossRef] [PubMed]
33. Blease, C.; Lewis, M.; Riggare, S.; Navarro, D.F.; Lehman, R. How Generative AI Affects Patient Agency. *BMJ* **2025**, *391*, e085323. [CrossRef] [PubMed]
34. Pinciotti, C.M.; Riemann, B.C.; Abramowitz, J.S. Intolerance of Uncertainty and Obsessive-Compulsive Disorder Dimensions. *J. Anxiety Disord.* **2021**, *81*, 102417. [CrossRef] [PubMed]
35. Surles, T. Can ChatGPT Make OCD Worse? What People with OCD Need to Know. NOCD 30 July 2025. Available online: <https://www.treatmyocd.com/blog/chatgpt-ocd-reassurance-seeking> (accessed on 23 May 2026).
36. Mouchabac, S.; Adrien, V.; Falala-Séchet, C.; Bonnot, O.; Maatoug, R.; Millet, B.; Peretti, C.-S.; Bourla, A.; Ferreri, F. Psychiatric Advance Directives and Artificial Intelligence: A Conceptual Framework for Theoretical and Ethical Principles. *Front. Psychiatry* **2021**, *11*, 622506. [CrossRef] [PubMed]

37. Scholten, M.; Gieselmann, A.; Gather, J.; Vollmann, J. Psychiatric Advance Directives under the Convention on the Rights of Persons with Disabilities: Why Advance Instructions Should Be Able to Override Current Preferences. *Front. Psychiatry* **2019**, *10*, 631. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Robinson, E.P.; Zhu, Y. Beyond “I Agree”: Users’ Understanding of Web Site Terms of Service. *Soc. Media Soc.* **2020**, *6*, 2056305119897321. [\[CrossRef\]](#)
39. Duffourc, M.N.; Verhees, F.G.; Gilbert, S. Artificial Intelligence Characters Are Dangerous without Legal Guardrails. *Nat. Hum. Behav.* **2025**, *10*, 218–221. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Kermani, A.; Perez-Rosas, V.; Metsis, V. A Systematic Evaluation of LLM Strategies for Mental Health Text Analysis: Fine-Tuning vs. Prompt Engineering vs. RAG. *arXiv* **2025**, arXiv:2503.24307.
41. Na, H.; Hua, Y.; Wang, Z.; Shen, T.; Yu, B.; Wang, L.; Wang, W.; Torous, J.; Chen, L. A Survey of Large Language Models in Psychotherapy: Current Landscape and Future Directions. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Kerrville, TX, USA, 2025.
42. Bilal, A.; Lin, B. OnRL-RAG: Real-Time Personalized Mental Health Dialogue System. *arXiv* **2025**, arXiv:2504.02894.
43. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [\[CrossRef\]](#)
44. Trivedi, S.; Simons, N.W.; Tyagi, A.; Ramaswamy, A.; Nadkarni, G.N.; Charney, A.W. An Independent Supervisory Safety Agent Improves Reaction of Large Language Models to Suicidal Ideation. *medRxiv* **2026**. medRxiv:2026.04.13.26350757.
45. Bara, A.; Oprea, S.-V. *A Multi-Agent Hierarchical Framework for Students’ Risk-Aware Mental Health Monitoring. SENTIA: State-Enabled Neuro-Symbolic Trajectory-Aware Intelligent Architecture*; Zenodo: Geneva, Switzerland, 2026.
46. Abramowitz, J.S.; Jacoby, R.J. *Obsessive-Compulsive Disorder in Adults*; Hogrefe Publishing: Newburyport, MA, USA, 2025.
47. Lerman, D.C.; Iwata, B.A. Prevalence of the Extinction Burst and Its Attenuation During Treatment. *J. Appl. Behav. Anal.* **1995**, *28*, 93–94. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Cheetirala, S.N.; Raut, G.; Patel, D.; Sanatana, F.; Freeman, R.; Levin, M.A.; Nadkarni, G.N.; Dawkins, O.; Miller, R.; Steinhagen, R.M.; et al. Less Context, Same Performance: A RAG Framework for Resource-Efficient LLM-Based Clinical NLP. *arXiv* **2025**, arXiv:2505.20320.
49. Coles, A.R.; Lahey, C.A.; Fawcett, E.J.; Fawcett, J.M. Overcoming the Forbidden: Identification and Stigma of Unacceptable Thoughts in Obsessive–Compulsive Disorder. *Stigma Health* **2025**, *10*, 314–325. [\[CrossRef\]](#)
50. Glazier, K.; Calixte, R.M.; Rothschild, R.; Pinto, A. High Rates of OCD Symptom Misidentification by Mental Health Professionals. *Ann. Clin. Psychiatry* **2013**, *25*, 201–209. [\[CrossRef\]](#)
51. Veale, D.; Freeston, M.; Krebs, G.; Heyman, I.; Salkovskis, P. Risk Assessment and Management in Obsessive–Compulsive Disorder. *Adv. Psychiatr. Treat.* **2009**, *15*, 332–343. [\[CrossRef\]](#)
52. Pestian, J.P.; Matykiewicz, P.; Linn-Gust, M.; South, B.; Uzuner, O.; Wiebe, J.; Cohen, K.B.; Hurdle, J.; Brew, C. Sentiment Analysis of Suicide Notes: A Shared Task. *BioMed. Inform. Insights* **2012**, *5s1*, BII.S9042. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Al-Mosaiwi, M.; Johnstone, T. In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clin. Psychol. Sci.* **2018**, *6*, 529–542. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Mattera, E.F.; Ching, T.H.; Zaboski, B.A.; Kichuk, S.C. Suicidal Obsessions or Suicidal Ideation? A Case Report and Practical Guide for Differential Assessment. *Cogn. Behav. Pr.* **2024**, *31*, 259–271. [\[CrossRef\]](#)
55. Mattera, E.F.; Sandoval, K.; Sipper, C.; Fitzpatrick, M.; Allen, A.; Pease, J.; Schild, S.; Ulaba-Samura, Y.; Cha, C.; Pittenger, C.; et al. Distinguishing Suicidal Obsessions from Suicidal Ideation: A First-Person Suicide Image Pilot Study. *J. Affect Disord.* **2025**, *391*, 119916. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Tolin, D.F. Psychometric Properties of a Structured Diagnostic Interview for DSM-5 Anxiety, Mood, and Obsessive-Compulsive and Related Disorders. *Assessment* **2018**, *25*, 3–13. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Goodman, W.K.; Price, L.H.; Rasmussen, S.A.; Mazure, C.; Fleischmann, R.L.; Hill, C.L.; Heninger, G.R.; Charney, D.S. The Yale-Brown Obsessive Compulsive Scale: I. Development, Use, and Reliability. *Arch. Gen. Psychiatry* **1989**, *46*, 1006–1011. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Zaboski, B.A.; Bednarek, L.; Ayoub, K.; Pittenger, C. Deep Learning in Obsessive-Compulsive Disorder: A Narrative Review. *Front. Psychiatry* **2025**, *16*, 1581297. [\[CrossRef\]](#) [\[PubMed\]](#)
59. Zaboski, B.; Fineberg, S.; Skosnik, P.; Kichuk, S.; Fitzpatrick, M.; Pittenger, C. Classifying Obsessive-Compulsive Disorder from Resting-State EEG Using Convolutional Neural Networks: A Pilot Study. *Comput. Psychiatry* **2026**, *10*, 1–17. [\[CrossRef\]](#)
60. Chan, B. Black-Box Assisted Medical Decisions: AI Power vs. Ethical Physician Care. *Med. Health Care Philos.* **2023**, *26*, 285–292. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Lang, B.H.; Kostick-Quenet, K.; Smith, J.N.; Hurley, M.; Dexter, R.; Blumenthal-Barby, J. Should Physicians Take the Rap? Normative Analysis of Clinician Perspectives on Responsible Use of ‘Black Box’ AI Tools. *AJOB Empir. Bioeth.* **2025**, *16*, 201–212. [\[CrossRef\]](#) [\[PubMed\]](#)

62. Saadeh, M.I.; Janhonen, J.; Beer, E.; Castelyn, C.; Hoffman, D.N. Automation Complacency: Risks of Abdicating Medical Decision Making. *AI Ethics* **2025**, *5*, 5783–5793. [\[CrossRef\]](#)
63. Zaboski, B.A.; Therriault, D.J. Faking Science: Scientificness, Credibility, and Belief in Pseudoscience. *Educ. Psychol.* **2020**, *40*, 820–837. [\[CrossRef\]](#)
64. Rosenbacke, R.; Melhus, Å.; McKee, M.; Stuckler, D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI* **2024**, *3*, e53207. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Hou, J.; Wang, L.L. Explainable Ai for Clinical Outcome Prediction: A Survey of Clinician Perceptions and Preferences. *AMIA Summits Transl. Sci. Proc.* **2025**, *2025*, 215. [\[PubMed\]](#)
66. Prégent, J.; Chung, V.-H.-A.; El Adib, I.; Désilets, M.; Hudon, A. Applications of Artificial Intelligence in Psychiatry and Psychology Education: Scoping Review. *JMIR Med. Educ.* **2025**, *11*, e75238. [\[CrossRef\]](#) [\[PubMed\]](#)
67. Zaboski, B.A.; Wilens, A.; McNamara, J.P.H.; Muller, G.N. Predicting OCD Severity from Religiosity and Personality: A Machine Learning and Neural Network Approach. *J. Mood Anxiety Disord.* **2024**, *8*, 100089. [\[CrossRef\]](#) [\[PubMed\]](#)
68. Nori, H.; Jenkins, S.; Koch, P.; Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv* **2019**, arXiv:1909.09223.
69. Hegselmann, S.; Ertmer, C.; Volkert, T.; Gottschalk, A.; Dugas, M.; Varghese, J. Development and Validation of an Interpretable 3 Day Intensive Care Unit Readmission Prediction Model Using Explainable Boosting Machines. *Front. Med.* **2022**, *9*, 960296. [\[CrossRef\]](#) [\[PubMed\]](#)
70. Yan, A.; Wang, Y.; Zhong, Y.; He, Z.; Karypis, P.; Wang, Z.; Dong, C.; Gentili, A.; Hsu, C.-N.; Shang, J.; et al. Robust and Interpretable Medical Image Classifiers via Concept Bottleneck Models. *arXiv* **2023**, arXiv:2310.03182.
71. Abi-Dargham, A.; Moeller, S.J.; Ali, F.; DeLorenzo, C.; Domschke, K.; Horga, G.; Jutla, A.; Kotov, R.; Paulus, M.P.; Rubio, J.M.; et al. Candidate Biomarkers in Psychiatric Disorders: State of the Field. *World Psychiatry* **2023**, *22*, 236–262. [\[CrossRef\]](#) [\[PubMed\]](#)
72. Craske, M.G.; Treanor, M.; Conway, C.C.; Zbozinek, T.; Vervliet, B. Maximizing Exposure Therapy: An Inhibitory Learning Approach. *Behav. Res. Ther.* **2014**, *58*, 10–23. [\[CrossRef\]](#) [\[PubMed\]](#)
73. Craske, M.G.; Kircanski, K.; Zelikowsky, M.; Mystkowski, J.; Chowdhury, N.; Baker, A. Optimizing Inhibitory Learning during Exposure Therapy. *Behav. Res. Ther.* **2008**, *46*, 5–27. [\[CrossRef\]](#) [\[PubMed\]](#)
74. Craske, M.G.; Treanor, M.; Zbozinek, T.D.; Vervliet, B. Optimizing Exposure Therapy with an Inhibitory Retrieval Approach and the OptEx Nexus. *Behav. Res. Ther.* **2022**, *152*, 104069. [\[CrossRef\]](#) [\[PubMed\]](#)
75. American Psychological Association. Ethical Guidance for AI in the Professional Practice of Health Service Psychology. APA. June 2025. Available online: <https://www.apa.org/topics/artificial-intelligence-machine-learning/ethical-guidance-ai-professional-practice> (accessed on 23 May 2026).
76. Vilaza, G.N.; McCashin, D. Is the Automation of Digital Mental Health Ethical? Applying an Ethical Framework to Chatbots for Cognitive Behaviour Therapy. *Front. Digit. Health* **2021**, *3*, 689736. [\[CrossRef\]](#) [\[PubMed\]](#)
77. Chin, M.H.; Afsar-Manesh, N.; Bierman, A.S.; Chang, C.; Colón-Rodríguez, C.J.; Dullabh, P.; Duran, D.G.; Fair, M.; Hernandez-Boussard, T.; Hightower, M.; et al. Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Netw. Open* **2023**, *6*, e2345050. [\[CrossRef\]](#) [\[PubMed\]](#)
78. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S. Large Language Models Encode Clinical Knowledge. *Nature* **2023**, *620*, 172–180. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.