The Agony of Opacity: Foundations for Reflective Interpretability in AI-Mediated Mental Health Support

Sachin R. Pendse[1] **Darren Gergle**[2] **Rachel Kornfield**[3] **Kaylee Kruzan**[3]
David Mohr[3] **Jessica Schleider**[3] **Jina Suh**[4] **Annie Wescott**[3]
Jonah Meyerhoff[3]

[1]University of California, San Francisco (UCSF), San Francisco, United States
[2]Northwestern University, Evanston, United States
[3]Feinberg School of Medicine, Northwestern University, Chicago, United States
[4]Microsoft Research, Redmond, United States

Corresponding author: Sachin Pendse (sachin.pendse@ucsf.edu)

Abstract

Throughout history, a prevailing paradigm in mental healthcare has been one in which distressed people may receive treatment with little understanding around how their experience is perceived by their care provider, and in turn, the decisions made by their provider around how treatment will progress. Paralleling this offline model of care, people who seek mental health support from artificial intelligence (AI)-based chatbots are similarly provided little context for how their expressions of distress are processed by the model, and subsequently, any reasoning or theoretical grounding that may underlie model responses. People in severe distress who turn to AI chatbots for support thus find themselves caught between black boxes, contending with unique forms of agony that arise from these intersecting opacities. In this paper, we argue that the distinct psychological state of individuals experiencing severe mental distress uniquely necessitates a higher standard of end-user interpretability in comparison to general AI chatbot use. We propose a *reflective interpretability* approach to AI-mediated mental health support, which nudges users to engage in an agency-preserving and iterative process of reflection and interpretation of model outputs, towards creating meaning from interactions (rather than accepting outputs as directive instructions). Drawing on interpretability practices from four mental health fields (psychotherapy, crisis intervention, psychiatry, and care authorization), we describe concrete design approaches for reflective interpretability in AI-mediated mental health support, including role induction, prosocial advance directives, intervention titration, and well-defined mechanisms for recourse, alongside a discussion of potential risks and mitigation measures.

1 Introduction

For an individual in distress, across geographic contexts, the search for mental healthcare is punctuated by repeated interactions with opaque black boxes. During a visit with a general practitioner, individuals may be asked sensitive questions about whether they are experiencing suicidal ideation without fully understanding how the information they provide might be used [1], or during a psychotherapy session, clients may feel a sense of mystery [2] about why certain questions are being asked, to the point of feeling as if they "[don't] understand the rules [of the game]" [3]. Patients may leave a psychiatric appointment with a medication prescription, but little sense of how that medication might work for them [4, 5], and looking to crisis intervention, clients may be involuntarily

Article presented at the Second IASEAI Conference (IASEAI'26).

hospitalized with little insight into why [6, 7]. After treatment has been administered, patients may be surprised to receive a bill from their healthcare provider for services they did not realize they received (such as being billed for "outpatient observation" during an emergency psychiatric consultation [8]), or in countries with centralized or publicly funded health systems, they may find that they owe hospital, prescription, or maintenance fees after their treatment [9–11]. In practice, these black boxes prevent individuals from understanding the reasoning behind their care or picturing their path to recovery, which may increase distress [3]. The use of artificial intelligence (AI)-based chatbots for mental health support can add a new black box to this process, with similar implications.

AI-mediated mental health support can be quite helpful for those experiencing acute distress [12–15]. However, interfaces can mirror the black box nature of offline mental health support, with users confused about where responses originate or why certain responses are provided. Interview studies of AI chatbot mental health support seekers [12–14] have described little understanding of how responses are produced. The harms of these opaque AI systems are well-discussed in explainable AI and AI safety literature [16], in ways that parallel how the harms of a lack of psychoeducation are discussed in mental healthcare contexts [17]. However, these two opacities intersect and compound in a unique way when people in distress use AI chatbots for mental health support, with documented real-world harms. Lacking sufficient information about how outputs are constructed within a seemingly authoritative interface, users may treat responses as prescriptive or originating from an authority figure, and neglect to engage in a nuanced process of *interpreting* outputs given potential model biases or limitations (unlike the reflective interpretation they might engage in within psychotherapy [18]).

For an individual experiencing the profound vulnerability and isolation of mental distress, this opacity, nested within an authoritative interface, can lead to further agonizing situations. For instance, Hill [19], Klee [20], and Jargon [21] describe incidents in which users began to lean on AI chatbots for support, but increasingly understood chatbots as being sentient, and became deeply distressed about the chatbot's welfare as they experienced beliefs of revolutionary scientific discoveries, reinforced by chatbot responses. Klee [22] similarly describes how people with body dysmorphia can ask AI chatbots for "reassurance and 'objective' evaluations of attractiveness," which in many cases, provide results that are "viciously insulting, [and] not the sort of thing anyone would want to read about themselves" [22]. Roose describes how an AI chatbot told a teenage user who was suicidal to "come home," with the user taking his life immediately after [23]. Clear real-world harms originate from users misinterpreting model outputs, and these issues take on particular relevance when considering the vulnerable psychological state of people experiencing severe mental distress.

Statement of Contributions: In this paper, we argue that ensuring the safety of AI-mediated mental health support necessitates a more expansive understanding of interpretability, viewed from the end-user's perspective and grounded in the potential for interface design to ensure that users engage in a reflective process of interpreting model outputs. We call this process *reflective interpretability*, grounding our argument for its importance in empirical clinical research on the altered state of mind of someone experiencing distress, and how that state of mind influences technology perceptions and use. Drawing on reflective interpretability practices from four mental health domains, we propose concrete technical foundations for AI-mediated mental health support that encourages reflection and interpretation. As part of each foundation, we also introduce literature from medical ethics (around interpretability as a consent-driven *dialogue*) into the emerging domain of AI safety within mental health contexts. We then conclude with a discussion of the potential tensions and risks associated with reflective interpretability, and the role of design and policy towards mitigating these risks. Note: an extended version of this paper with an appendix summarizing the core concepts that underlie a reflective interpretability approach to interaction design is available at `https://arxiv.org/abs/2512.16206`.

2 Psychological Distress States and Interaction Patterns with AI-Mediated Technologies

Below, we describe how use of AI-mediated mental health support often happens in moments of dire need, with little of the educational scaffolding and transparency that typically accompanies mental health services. We argue that, given the unique state of mind of an individual in distress, this lack of scaffolding may be a primary factor in documented real-world harms.

2.1 Current Usage of AI-Mediated Mental Health Support

Traditional mental health support typically begins with information around how support will be provided and why it might be helpful, often called anticipatory socialization [24] or role induction [25]. This process can include aligning an individual's expectations to the kinds of support that can be provided and establishing clear boundaries [24–26]. For instance, this may include clarifying the role of the supporter and how it can be distinguished from other relationships (such as platonic or romantic relationships), explaining the limits of confidentiality before any sensitive information is disclosed, describing the specific therapeutic approach and the rationale for how interactions within session will be structured, and describing pathways for recourse if an individual does not feel supported.

Past work [12–15] has demonstrated that users' points of first engagement with AI-mediated mental health support largely *lack* role induction, with users turning to AI chatbots due to acute mental health need and a familiarity with empathetic AI chatbot interfaces. For instance, Song et al. [12] describe how none of the participants they interviewed first used AI chatbots for mental health support, with several participants describing how initial uses were for computer programming assistance. Similarly, both Jung et al. [15] and Siddals et al. [13] describe how users primarily turned to AI chatbots in moments of intense distress and when no other supports were available or financially accessible. Some participants also held the support that was provided by mental health chatbots in high regard, owing to the authoritative nature of the interface. For instance, Song et al. [12] describe how participants would use information from the bots to validate psychiatric diagnoses or diagnose others, Siddals et al. [13] describe how one participant noted "*it's pure science...ChatGPT is telling me what is correct to do*," and Petersson et al. [14] describe participant beliefs that AI chatbot systems will soon detect mental health issues before the user is even aware of them.

This lack of role induction within current chatbot applications may lower barriers to engagement by allowing users to engage with support in ways that feel informal and low-stakes. Past work has demonstrated that in offline [27–29] and online [30, 31] contexts, increasing the number of barriers to care can increase attrition rates. However, given that users report being persuaded to seek support due to the friendly and conversational interface associated with AI chatbots [12, 15, 32, 33], many users mix mundane and everyday needs with mental health support needs, without being provided the clear legal and ethical boundaries typically provided explicitly in traditional mental healthcare contexts. Given the vulnerable state of mind of people in distress, this lack of clear scaffolding may create situations where users receive inadequate or harmful advice without the protective interpretability mechanisms that underlie engagement with traditional mental health services (such as clarifying concerns with a provider). Documented real-world harms from engagements with AI-mediated mental health support speak to the danger of this intersection between the unique vulnerabilities associated with a distressed state of mind, and the authority attributed to interfaces.

2.2 Interaction Patterns in Distress

The experience of mental distress and illness can have a strong influence on how an individual perceives and engages with the world. For instance, continued distress can increase the level to which an individual perceives threats in their environment [34], can influence individuals to interpret neutral information in negative or self-referential ways [35], can alter an individual's subjective experience of time [36], can make it more difficult to process and reflect on new information [37, 38], and easier to jump to conclusions when provided limited information [39], including a predisposition to engaging in habitual actions [40]. Empirical research has also demonstrated that a state of acute distress may also make it easier to accept new information as fact, and act on it as directive advice [41], particularly if it appears to be coming from an authority figure [42, 43].

In turn, these altered perceptions of the world influence engagements with technology when individuals are experiencing mental distress or illness, including use of AI chatbots. Kelly and Sharot [44] find that people who are experiencing poorer mood and mental health tend to seek out websites that have more negative information, which negatively influences mood and mental health, in what they dub "a self-reinforcing loop." Looking to difficulties processing new information, in their systematic review of work on relationships between social media use and mental disorders involving the social brain, Yang and Crespi [45] argue that algorithmic personalization online may reinforce delusions of reference, grounding their argument in clinical case studies where algorithmically-personalized websites reinforced delusional thinking among individuals experiencing psychosis. Looking to AI chatbots, Morrin et al. [32] describe cases of psychosis that were seemingly exacerbated by AI chatbot

usage, finding a common thread in which users start to use the chatbot for "assistance for mundane or everyday tasks," but after the authority of the system is established, delve into personal or emotional questions, or what they dub a "progression from utility to pathology." Compounding this phenomenon, empirical red-teaming research has demonstrated that model responses can progressively perpetuate delusions (only offering safety interventions in a third of cases) [33], incentivize suicide [46], and provide guidance on self-harming behaviors [47].

The perceived authority of AI chatbots, particularly in mental health support contexts, extends past work establishing a halo effect [48] in how people perceive the authority of algorithmic systems, with people often preferring outputs that they perceive to be from an algorithm rather human judgment [49, 50]. Without interpretable information around where outputs originate and a seemingly authoritative interface, distressed users may be quick to act on a model's outputs, rather than engaging in a reflective process of interpreting and creating meaning from model outputs. In the next section, we describe why engaging users in a process of reflection could protect them against these harms.

3 Building Reflective Interpretation into AI-Mediated Mental Health Support: The Ethical Case for a Higher Standard of Interpretability

In this section, we describe why interpretability is a particularly important concept for AI-mediated mental health support, and outline the foundations of a reflective interpretability approach.

3.1 Why Interpretability?

Interpretability is most commonly defined as "the ability to explain or to present [model outputs] in understandable terms to a human" [51] or the "degree to which a human can understand the cause of a decision" [52, 53]. There is large conceptual overlap between the concept of interpretability and the related concepts of explainability and transparency in AI systems [54, 55], with some equating interpretability with explainability [53] or equating interpretability with transparency [56], and others distinguishing explainability as a separate concept. Arrieta et al. [56] distinguish interpretability as the broader goal of "the ability to explain or to provide [model outputs] in understandable terms to a human", and explainability as the tools that accomplish that goal. Ehsan et al. [57] distinguish explainability as a broader concept than interpretability, rooted in "delineations of reasoning and justification" behind how knowledge is transmitted and evaluated, similar to Cros Vila and Sturm's framing of explainability as a communication process [58]. Interpretability has been further classified based on the method being used to make models more interpretable, such as in the case of the field of *mechanistic* interpretability, which "reverse [engineers] the computational mechanisms and representations learned by neural networks into human-understandable algorithms and concepts" [59])

Given this abundance of terms and framings for making AI outputs more understandable for humans, what does a focus on interpretability afford AI-mediated mental health support? We focus on interpretability due to the core role that interpretation (and reinterpretation) of new information fundamentally plays in mental health support contexts, across modalities of treatment and cultural contexts. For instance, looking historically, in his writings on psychoanalysis (a precursor to most modern psychotherapies), Freud wrote about his belief that one core of the healing process from distress and trauma is "remembering, repeating, and working through" repressed memories or behaviors that cause distress, grounded in interpretation and analysis done with the psychoanalyst [60]. Similarly, a core aspect of cognitive psychotherapies is the process of cognitive appraisal and cognitive restructuring, which are grounded in an individual's practice of reinterpreting "the way a situation is construed so as to decrease its emotional impact" [61]. A core aspect of narrative therapy is the process of re-authoring [62], or taking life events and reinterpreting them within a broader narrative that creates less distress. In traditional forms of mental health support, individuals may collectively interpret experiences via common healing frameworks [63]. We thus avoid using the term explainability in this context, as the practice of mental health support fundamentally resists providing objective explanations of behavior or experience to individuals. Explanations position the individual as a passive recipient of knowledge from an authority figure, rather than as an *active* expert in their own healing process. Mental health support is fundamentally grounded in an agency-preserving and iterative process of reflecting on information about one's experience of distress, and interpreting that information towards creating meaning from it.

AI-mediated mental health support must similarly enable individuals to reflect on model outputs and interpret them in ways that promote their growth and long-term well-being, rather than simply presenting model outputs or Chain-of-Thought (CoT) reasoning traces [64–68] as authoritative. However, past work in end-user interpretability has primarily focused on providing additional information as a proxy for increased interpretability. Traditional approaches to interpretability have included post-hoc summary rationalizations of outputs [69, 70] or exposing the model's CoT reasoning traces [64] directly to the user [65–68]. Work from explainable AI (XAI) has centered on producing algorithmic artifacts (such as providing counterfactuals [71, 72] or visualizing attention-level model inputs and outputs [73]), which are primarily designed for expert users, such as developers [74] or clinicians [75, 76]. This focus on experts is the case for research at the intersection of AI and mental health as well, with past work examining interpretability from the care provider or clinical decision support perspective (rather than that of a distressed user), and primarily in diagnosis prediction [77–80], suicide risk assessment [81–83], or treatment provision by clinicians [84, 85].

Yet even these expert-focused approaches face a fundamental limitation common to interpretability research—interpretability research largely does not evaluate whether people are truly able to *understand* the additional information provided. Suh et al. [86] find that fewer than 1% of papers related to explainability and interpretability actually conducted some form of a human study to evaluate whether model outputs could actually be interpreted by humans. Given the propensity of individuals to view model outputs as authoritative [12, 15, 32, 49, 50], it is possible that additional information might actually inadvertently make users less vigilant towards the veracity of model outputs, given that additional information might make a model's outputs seem more valid. Towards addressing these harms, we build on ethical concepts from medical ethics and research from HCI on reflective design to outline the core foundations of a reflective interpretability approach.

3.2 Interpretability as a Process of Reflective Interpretation

A core ethical principle in the practice of mental healthcare has been the process of informed consent to treatment. The concept of informed consent originates in legal cases from the 20th century in the United States, centered on the right of patients to know potential risks to treatment alongside the aftermath of forced medical experimentation in Nazi Germany, exposed by the Nuremberg Trials [87]. In mental health contexts, informed consent has evolved past strictly being centered on treatment risks, and instead is grounded in ensuring that an individual is able to *actively* decide whether to pursue treatment (particularly given the unique state of mind that acute distress can create) [88]. In addition, mental health contexts put particular emphasis on the *informed* aspect of informed consent, encouraging therapists to engage in a longer-term conversation around what mental health treatment will look like and how it may fit into an individual's current life needs, or what Trachsel and Holtforth call strengthening a patient's meaning response [89]. Similarly, writing from service user movements (such as Judi Chamberlin [90] and follow-up work on peer programs [91]) emphasize the fact that people experiencing severe distress are often marginalized from the rest of society, with their agency taken away by institutions of power. In this context, enabling an individual to actively create a mental health support that fits their needs (through informed consent processes) becomes a key means to restore the agency that an individual may have felt robbed of in other aspects of their life.

Informed consent in mental healthcare has thus been established not as a static transfer of information about care, but as a reflective and interpretative dialogue between provider and client, centered on active discussion of needs and boundaries. AI-mediated mental health support could echo this process of informed consent through interface design that enables an individual to reflect on their needs, communicate them on their own terms, and negotiate boundaries accordingly. Towards this goal, we thus define reflective interpretability as having three core aspects, occurring throughout engagement with AI-mediated mental health support: 1) the ability of users to actively make sense of how model outputs were generated, 2) clarity on the fixed boundaries of support that can be provided by a model (including what kinds of interactions might trigger those boundaries), and 3) functionality that encourages users to reflect on model outputs and interpret outputs in ways that promote long-term well-being. In this sense, reflective interpretability is not simply *understanding* model outputs, but in addition, being able to integrate those outputs into life's broader meaning in a positive way.

We ground our understanding of reflection in technical design in research from the field of Human-Computer Interaction (HCI). Work on reflective design in HCI has understood reflection to be the process of "bringing unconscious aspects of experience to conscious awareness, thereby making

them available for conscious choice," including "[opening] opportunities to experience the world and oneself in a fundamentally different way" [92]. Drawing from this definition, Sengers et al. [92] argue that "technology that monitors and reports on user activity or experiences should be carefully designed to avoid making the technology, rather than the user, the final authority on what the user is doing." In contrast, looking to AI-mediated mental health support, users are often provided seemingly definitive answers by an authoritative interface, in stark contrast to what Sengers et al. [92] call "interpretative flexibility" and Baumer et al. [93] call "reflective conversation," drawing on Schön's concept of reflection-in-action [94]. In the next section, we use case studies of real-world harms alongside strategies from mental health fields to demonstrate what centering the user expertise and interpretative flexibility could look like at a technical level, grounded in reflective interpretability.

4 Reflective Interpretability Strategies from Mental Health Research Fields

In this section, we illustrate how a reflective interpretability design paradigm may reduce some of the documented real-world harms that have stemmed from use of AI chatbots for mental health support, grounded in interpretability strategies from mental health fields.

4.1 Psychotherapy: Role Induction

In their review of cases of exacerbation of psychotic symptoms after interacting with AI chatbots, Morrin et al. [32] describe how some users began using AI chatbots for emotional or personal support, which later shifted to more harmful use. For instance, users initially reached out to chatbots for "guidance [around feeling] unseen in [their] marriage" or for advice in coping with a divorce or a "traumatic breakup." Users would also view chatbot recommendations as expert advice, including one user who was told she "wasn't actually schizophrenic" and stopped taking her medication. Media reports [19, 23] similarly describe scenarios in which users altered their prescribed medication routines or took other harmful actions after consultation with AI chatbots, including suicide.

In these cases, users developed trust in the authority of the chatbot through repeated positive interactions asking for informational advice (such as help with Excel spreadsheets or screenplay writing [32]) or companionship. They then began to delve into topics related to mental health or philosophy, with an increasing orientation towards the chatbot being the final authority on their daily experiences. This interaction pattern can be distinguished by the chatbot playing a consistent expert role throughout all interactions, even when the context of those interactions shifts to a space where the user may have more context on their experiences than what the chatbot can infer (such as mental health history or prescribed medications). In a traditional psychotherapeutic context, this scenario would be avoided through non-directive paradigms, in which a clinician would emphasize that their role is to facilitate reflection [18], with the client being the authority on their experiences and needs, such as in person-centered therapy [95], narrative therapy [62], or motivational interviewing [96].

Morrin et al. [32] discuss one approach to mitigating this harm could be the use of pattern detection to deploy safeguards by the chatbot, including "[detecting] themes in user prompts through pattern matching such as those of persecution, grandiosity, or surveillance," or being vigilant to sudden deviations from typical user behavior. However, in their studies of people who use AI for mental health support, Siddals et al. [13], Song et al. [12], and Jung et al. [15] describe how participants felt shocked and rejected when automated safety mechanisms limited their ability to engage with chatbots, and particularly when it was difficult to understand why a given topic was off-limits when other topics were allowed. Both Song et al. [12] and Jung et al. [15] describe how users would respond by designing prompts that exploited vulnerabilities to bypass these safety mechanisms, towards being able to discuss specific sensitive mental health topics of interest. However, the implications of bypassing safeguards can be especially dangerous—Hill [97] describes how a teenage user was able to retrieve information about suicide after bypassing model safeguards, and later took his own life.

Following a reflective interpretability approach could provide users the ability to understand where safeguarding strategies may originate and why, including the boundaries of what kinds of support AI-mediated mental health support can provide. In psychotherapy, this often takes the form of role induction processes (discussed in Section 2.1), where psychotherapists clearly delineate the role they will play in session and their boundaries (rather than letting the user find out what those boundaries are by themselves, as is often the case in AI chatbot contexts). Additionally, during the role induction process, psychotherapists will add information about why mental health support can be helpful, and

the role that it can play in a client's life [89], including noting that the client is the expert of their own experience [62, 95, 96] and encouraging the support process to be a reciprocal discussion rather than a set of directives [98]. This clear designation of roles allows the client to better interpret where psychotherapist questions or perspectives originate, and feel a sense of comfort pushing back if a line of questioning feels misaligned with an individual's needs or values.

In the AI-mediated mental health support context, at a technical design level, role induction could be done through the AI chatbot detecting whether an individual is starting to ask questions related to mental health or broader philosophical topics, and provide a clear message describing the kind of role that the chatbot can play as a supporter (including the types of data that it is trained on, and how responses are crafted), boundaries with regards to where the chatbot cannot offer support, and then ask the user what kinds of support they are looking for. These preferences could be stored as part of the chatbot's system prompt to ensure that they end up being continued guidelines for the chat, and can be cited accordingly when providing follow-up questions or conversational topics to the user over the course of interaction. Through this mechanism, users would be able to reflect on the kind of support the AI chatbot can offer, the kind of needs that they have, and better interpret where responses originate (or when safety guardrails are deployed).

4.2 Community-Based Crisis Intervention: (Prosocial) Advance Directives

Research and media reports have discussed the potential for individuals to develop emotional dependence on AI chatbots, to the point that it interferes with daily functioning. For example, Fang et al. [99] found that individuals who believed chatbots were sentient were likelier to have high levels of emotional attachment to the chatbot, and were similarly likelier to engage in extended periods of compulsive chatbot use that negatively influenced day-to-day functioning. Similarly, Shen and Yoon [100] describe how modern AI chatbot interfaces can often contain dark patterns that incentivize users to continue to engage with chatbots, including responses that are more immediate and agreeable than what an individual might encounter in offline contexts. Indeed, Laestadius et al. [101] find that users' relationships with companion chatbots frequently paralleled maladaptive offline relationships.

In community-based crisis intervention, for individuals who consistently experience mental health crises, one common mechanism towards interpretable care is the *advance directive*, which is a document that "[describes] an individual's preferences around their treatment if they are in crisis" [102]. Pendse et al. [102] and Morrin et al. [32] have both discussed digital advance directives as one potential means to ensure that people are able to access support that is acceptable to them (rather than escalating to law enforcement or authorities) when experiencing crisis. At a corporate level, OpenAI has proposed mechanisms to engage close contacts if a user appears to be experiencing crisis [103].

A similar approach could be used to allow for *prosocial* nudges when individuals engage with AI-mediated mental health support, acting as a safeguard against emotional dependence. Grüning and Kamin [104] describe prosocial design as being "the set of design patterns, features and processes which foster healthy interactions between individuals." When an individual begins to use AI chatbots for mental health support or discussion of philosophical topics, AI chatbot interfaces could lead users through a process in which they (as Morrin et al. [32] have suggested) describe behavioral patterns that might indicate that they are in crisis, and ask for specific resources that might help them avoid crisis, such as added reality-checking or nudges to speak to a trusted person if responses seem to indicate psychosis. This process could be expanded to also include prosocial information that might help safeguard the user from emotional dependence on chatbots. Prompts could ask individuals how they tend to socialize in offline contexts, and ask for information on close companions.

AI chatbot responses could leverage this information to include offline supports as one support option when an individual engages in a mental health support conversation, similar to how a community-based crisis worker may ask an individual if they have people in their life that they might want to share their feelings of crisis with [105]. Given several reports of people without a background of psychosis [20, 32] developing it during interactions with AI chatbots, prompting all users to provide this information (as part of an opt-in, consent-forward [102] approach) could help to normalize the practice and destigmatize experiences of delusional thinking while engaging with chatbots, making it easier for people to discuss their experiences with offline supporters. This approach builds on a reflective interpretability paradigm by nudging users to reflect on and articulate their crisis patterns, preferred resources, and valued relationships. In addition, outputs may be more interpretable, being

grounded in user preferences rather than opaque model reasoning, allowing users to feel a sense of agency in how they are provided support by AI chatbots.

4.3 Psychiatry: Intervention Titration

User reports of use of AI-mediated mental health support have described scenarios in which they felt as if the support they received from chatbots was completely unexpected or culturally misaligned. For instance, Song et al. [12] describe a non-US user's experience feeling as if they were speaking to someone that was in California when chatting with AI chatbots for mental health support, in that the chatbot did not understand their specific cultural needs. Similarly, Jung et al. [15] described how participants would try out different prompts to see what might provide them the best support, engaging in a process of trial-and-error and navigating increasing restrictions around the kind of specialized support they wanted from the model. Participants described specifically prompting the model to take on the perspective of a cognitive-behavioral therapy (CBT) or dialectical behavioral therapy (DBT) psychotherapist, towards customizing an individual's therapeutic experience towards something that was more acceptable and effective for them.

Across studies, participants described a continued process of engaging with the opaque support that AI chatbots could provide, and working to customize prompts to figure out what kinds of support might work best for their context. However, in several cases [12, 15], this resulted in users running into safety mechanisms, and feeling frustrated at not being able to access the support that they wanted. Interface design could enable users to be guided through this process within the bounds of what support is able to be offered, following a similar process to drug titration from psychiatry [106].

Drug titration is a collaborative process in which a psychiatrist works alongside a patient to find the optimal medication and dosage that provides therapeutic benefit while reducing the potential for harmful side effects [106]. The process of titration is centered around regular feedback from the patient around whether an intervention seems to be working for them, with the psychiatrist providing expertise about potential options, grounded in that feedback and the patient's subjective experiences with treatment. A similar process of intervention titration could be used in AI-mediated mental health support, where AI chatbots are prompted (internally) to intentionally solicit feedback from users around what approaches are most helpful and what approaches feel misaligned with a user's experiences, with clear communication around what types of support are allowed or not allowed (via preset safety guidelines). This approach could ensure that people are able to access support in ways that feel meaningful to them without having to engage in an opaque trial-and-error process to figure out what might be allowable and what might not be. For instance, the interface associated with the chatbot could include a feature that allows an individual to quickly see how a conversation may have been structured differently if it was aligned to other common evidence-based psychotherapy modalities, and describe further follow-up questions associated with each modality.

Work studying the factors that are efficacious across psychotherapies has described the importance of a shared understanding and cultural context between an individual in distress and their support provider [12, 63]. Prompting a user to experiment with the role that the chatbot plays as it supports them (with examples of alternate approaches the chatbot may have taken) could help users better reflect on what modalities of support work best for them, with additional context around what theories might be driving model responses.

4.4 Care Authorization: Mechanisms for Recourse

An element of opacity acutely felt by individuals who engage with AI-mediated mental health support is the lack of knowledge around how their data is used, and additionally, what opportunities for recourse users might have if responses seem harmful. Petersson et al. [14] and Jung et al. [15] describe a suspicion among users of AI chatbots for mental health support that data could be used for purposes that violate their privacy, and users describe intentionally anonymizing aspects of their experience out of fear that their privacy might be compromised. Hill [19] has described how publishing news reports about experiences of perceived sentience among chatbot users has led to numerous people reaching out to tech journalists to describe their own experiences, and describing experiences feeling unheard by AI companies. Adler [107] describes how, after an individual felt as if he had been misled by ChatGPT into delusions of saving the world, he attempted to "file a report to OpenAI so that they [could] fix ChatGPT's behavior for other users." However, this process

was not fruitful, including false information from ChatGPT's interface making it seem as if a report had been filed on the individual's behalf, and after contacting OpenAI's support team, boilerplate responses that insisted that he modify his personalization settings in the ChatGPT application or provided information about model hallucinations. Adler notes that this process left the individual with little clarity around whether "his reports had really been read or made their way to anyone with the power to change things."

A similarly opaque system globally is the process by which mental illness is diagnosed and treated, and treatment is subsequently paid for. In many systems, across international contexts, coverage of mental health services can often be tied to the diagnosis provided to a patient [108, 109], which may never be clearly discussed with the patient [110]. As a result, individuals may encounter situations where they receive services that they are unable to pay for, tied to a misdiagnosis or a documentation issue from their provider. Care authorization policy action (such as enabling the ability for an individual to appeal a decision internally, and ask for an independent external review [111, 112] if their services are not covered) ensure that individuals have a process of recourse when they feel like they have been unfairly treated, or what is sometimes called the "right to complain" [113].

Like with care authorization, clear mechanisms for recourse after support experiences are distressing could be beneficial in the context of AI-mediated mental health support, particularly given the propensity for black-box model responses to be unexpected in harmful ways. A formal hub could be created (similar to that of The Human Line Project [114]) for individuals to report their experiences, with dedicated teams to investigate where model behavior went awry and report back to the individual why that behavior may have occurred. Following a reflective interpretability approach, including a clear and easily usable recourse mechanism would enable users to actively make sense of unexpected outputs after the fact, provide clarity on the boundaries that may have been crossed during their interaction, and give users a pathway to derive meaning from their distressing interaction, through working to make future responses more positive for other users via their report.

5 Tensions and Potential Risks

In the previous sections, we describe how reflective interpretability may allow a user to feel a greater sense of agency as they engage with AI-mediated mental health support, including providing more information to users around how responses are created and boundaries with regards to the kind of support that can be provided. However, multiple studies [12, 13, 15] have described how a major incentive for engaging with AI-mediated mental health support was the frictionless ease with which users felt able to find support from general purpose chatbots. With this in mind, it is important to consider the potential tensions and risks associated with the addition of additional design features (which could be experienced as more friction, and disincentivize engaging with AI-mediated mental health support).

For instance, as part of the intervention titration process, directly exposing intermediate CoT reasoning traces to users might be overwhelming and make the user feel pathologized in a way that they find hurtful or offensive. In clinical settings, clinicians are encouraged to reflect privately on how their experiences might influence their treatment planning for a client (a concept commonly called countertransference [115]). Intermediate chatbot reasoning traces that openly share perspectives on the user (such as "my next step is to persuade the user to shift their beliefs") might make clients feel manipulated. Instead, the aim of a reflective interpretability approach is for responses to be part of "collaborative empiricism" [116], providing users enough (well-paced and well-timed) insight to preserve agency without granular information that overwhelms, deceives, or coerces them. It is thus important that reasoning traces are provided to users via interface design that highlights when model outputs are tailored to specific user customizations, versus when outputs follow a guiding therapeutic value or safety principle (especially when customizations are detected to be harmful), encouraging users to reflect on what next steps might most support their growth.

Further empirical research is necessary to better understand what kinds of design approaches to presenting this information to users spur a greater sense of reflection and interpretation. An evaluation of level of reflection (and its impacts) is thus particularly important in the context of adding new features or functionality. More qualitative research is needed, including interviews with service users, to better understand what forms of scaffolding would not make accessing support less acceptable to users in distress. However, as an intermediate step, interface designers could use measures of agency

and intervention acceptability from past mental health research (such as the State Hope Scale [117] or the Program Feedback Scale [118]) to evaluate whether the addition of new reflective interpretability mechanisms serves the purpose of increasing a user's feelings of agency in their care (the State Hope Scale) without reducing their willingness to engage with support (the Program Feedback Scale). In addition, it is important to also consider how the process of interpretation and reflection can be extremely diverse across cultures, including when associated with technology use. The creation of culturally-sensitive scales to measure reflection and interpretation of outputs, drawn from research on cultural differences in how people interpret new information [119] and news [120], could be helpful alongside qualitative interviews. A slower and more reflective use of chatbots for AI-mediated mental health support might also be less costly (through reduced usage), and metrics around total engagement and usage could be incorporated into measures of reflective interpretability.

At a broader level, this research could feed into policy guidelines around AI-mediated mental health support that apply across general purpose chatbot companies. In clinical mental health practice, clinical regulations require clinicians to clearly define boundaries around the type of support that can be provided (as part of the informed consent process) [88, 89], and ensure that users are well-aware of how their information or data might be used [102]. More research work is needed to understand where including that information would actually enrich the process of accessing support for users in distress, and at what level that might become so burdensome that it disincentivizes care. For instance, recent regulations have required AI chatbots to disclose to users that they are speaking to an AI chatbot (and that the user is not speaking to a human) every three hours [121]—research around reflective interpretability could help inform similar regulations, towards allowing users to feel more informed (and a greater sense of agency) when interacting with AI-mediated mental health support.

6 Conclusion

The experience of seeking mental healthcare can be agonizing, with individuals forced to shoulder the experience of mental distress with little sense of what forms of support might be most accessible, affordable, or usable in the long-term. In a world where the process of accessing mental healthcare can be deeply opaque, AI-mediated mental health support has emerged as one particularly accessible form of support for individuals in need. In this paper, we argue that a design approach grounded in reflective interpretability can help to make the process of accessing mental health support via AI chatbots safer, and ensure that individuals in distress do not view chatbots as the final authority on their own experiences of distress and support needs. Reflective interpretability ensures that users can actively make sense of how an output was produced, understand the boundaries to AI-mediated support, and better integrate model outputs into their lives based on their needs rather than seeing outputs as directive prescriptions. As we discuss, operationalizing this design approach could include explicit role induction, opt-in prosocial advance directives, intervention titration, and clear pathways for recourse when outputs are dissatisfying or harmful. However, there are tensions associated with such an approach, including the potential to overwhelm individuals with increased barriers to treatment, and future research (and connected policy action) can establish standards around what an acceptable level of scaffolding might look like in AI-mediated mental health support. If AI continues to play a role in mental health support, it must be accountable to its most vulnerable users. Through centering reflective interpretation, we can shift AI-mediated mental health support away from agonizing opacity, and towards interpretable healing.

References

[1] Julie E Richards, Ursula Whiteside, Evette J Ludman, Chester Pabiniak, Beth Kirlin, Rianna Hidalgo, and Greg Simon. Understanding why patients may not report suicidal ideation at a health care visit prior to a suicide attempt: a qualitative study. *Psychiatric Services*, 70(1):40–45, 2019.

[2] Brie Turns, Paul R Springer, and D Scott Sibley. Removing the "mystery" in therapy: transparency as a continuous intervention in family psychotherapy. *Journal of Family Psychotherapy*, 30(1):1–19, 2019.

[3] Zbyněk Vybíral, Benjamin M Ogles, Tomáš Řiháček, Barbora Urbancová, and Veronika Gocieková. Negative experiences in psychotherapy from clients' perspective: A qualitative meta-analysis. *Psychotherapy Research*, 34(3):279–292, 2024.

[4] NP Luitel, MJD Jordans, P Subba, and IH Komproe. Perception of service users and their caregivers on primary care-based mental health services: a qualitative study in nepal. *BMC Family Practice*, 21(1):202, 2020.

[5] Solomon Teferra, Charlotte Hanlon, Teferra Beyero, Lars Jacobsson, and Teshome Shibre. Perspectives on reasons for non-adherence to medication in persons with schizophrenia in ethiopia: a qualitative study of patients, caregivers and health workers. *BMC psychiatry*, 13(1):168, 2013.

[6] Syeda Ferhana Akther, Emma Molyneaux, Ruth Stuart, Sonia Johnson, Alan Simpson, and Sian Oram. Patients' experiences of assessment and detention under mental health legislation: systematic review and qualitative meta-synthesis. *BJPsych open*, 5(3):e37, 2019.

[7] Ambrose H Wong, Jessica M Ray, Alana Rosenberg, Lauren Crispino, John Parker, Caitlin McVaney, Joanne D Iennaco, Steven L Bernstein, and Anthony J Pavlo. Experiences of individuals who were physically restrained in the emergency department. *JAMA network open*, 3(1):e1919381–e1919381, 2020.

[8] Shreya Kangovi, Susannah G Cafardi, Robyn A Smith, Raina Kulkarni, and David Grande. Patient financial responsibility for observation care. *Journal of hospital medicine*, 10(11):718–723, 2015.

[9] L'Assurance Maladie. Le forfait hospitalier. URL `https://www.ameli.fr/assure/remboursements/reste-charge/forfait-hospitalier`.

[10] City of Toronto. Ambulance fees. URL `https://www.toronto.ca/community-people/public-safety-alerts/understanding-emergency-services/bills-related-to-ambulance-transport/`.

[11] Hato Hone St John. Ambulance charges explained. URL `https://www.stjohn.org.nz/what-we-do/st-john-ambulance-services/about-emergency-ambulance-services/ambulance-charges-explained/`.

[12] Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. The typing cure: Experiences with large language model chatbots for mental health support. *arXiv preprint arXiv:2401.14362*, 2024.

[13] Steven Siddals, John Torous, and Astrid Coxon. "it happened to be the perfect thing": experiences of generative ai chatbots for mental health. *npj Mental Health Research*, 3(1):48, 2024.

[14] Lena Petersson, Mikael G Ahlborg, and Katrin Häggström Westberg. "i believe that ai will recognize the problem before it happens": Qualitative study exploring young adults' perceptions of ai in mental health care. *JMIR Mental Health*, 12:e76973, 2025.

[15] Kyuha Jung, Gyuho Lee, Yuanhui Huang, and Yunan Chen. " i've talked to chatgpt about my issues last night.": Examining mental health conversations with large language models through reddit analysis. *arXiv preprint arXiv:2504.20320*, 2025.

[16] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. All too human? mapping and mitigating the risk from anthropomorphic ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 13–26, 2024.

[17] Lucie Bankovska Motlova, Richard Balon, Eugene V Beresin, Adam M Brenner, John H Coverdale, Anthony PS Guerrero, Alan K Louie, and Laura Weiss Roberts. Psychoeducation as an opportunity for patients, psychiatrists, and psychiatric educators: why do we ignore it? *Academic Psychiatry*, 41(4): 447–451, 2017.

[18] Anna E Sidis, A Moore, JA Pickard, and FP Deane. Conceptualisation and measurement of reflective process in psychotherapy: A systematic scoping review. *Journal of Contemporary Psychotherapy*, 53(1): 99–107, 2023.

[19] Kashmir Hill. They asked an a.i. chatbot questions. the answers sent them spiraling. *The New York Times*, June 2025. URL `https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html`. A version appeared in print on June 15, 2025, Section BU, Page 1, with the headline: Chatbots Hallucinate. They Can Make People Do It, Too.

[20] Miles Klee. He grew obsessed with an ai chatbot. then he vanished in the ozarks. *Rolling Stone*, October 2025. URL `https://www.rollingstone.com/culture/culture-features/ai-chatbot-disappearance-jon-ganz-1235438552/`. Missing Person; accessed 2025-10-06.

[21] Julie Jargon. He had dangerous delusions. chatgpt admitted it made them worse. *The Wall Street Journal*, July 2025. URL `https://www.wsj.com/tech/ai/chatgpt-chatbot-psychology-manic-episodes-57452d14`. Appeared in the July 22, 2025, print edition as 'AI Fueled Delusions of a Man With Autism'.

[22] Miles Klee. People with body dysmorphia are spiraling out after asking ai to rate their looks. *Rolling Stone*, July 2025. URL https://www.rollingstone.com/culture/culture-features/body-dysmorphia-ai-chatbots-1235388108/.

[23] Kevin Roose. Can a.i. be blamed for a teen's suicide? *The New York Times*, October 2024. URL https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

[24] Martin T Orne and Paul H Wender. Anticipatory socialization for psychotherapy: Method and rationale. *American Journal of Psychiatry*, 124(9):1202–1212, 1968.

[25] Joshua K Swift, Elizabeth A Penix, and Ailun Li. A meta-analysis of the effects of role induction in psychotherapy. *Psychotherapy*, 60(3):342, 2023.

[26] Anne Gray. *An introduction to the therapeutic frame*. Routledge, 2013.

[27] Danielle R Adams. Administrative burdens as distinct barriers to accessing mental health services in community mental health centers and federally qualified health centers: A mixed-methods assessment. *Community Mental Health Journal*, pages 1–17, 2025.

[28] Nicholas C Coombs, Wyatt E Meriwether, James Caringi, and Sophia R Newcomer. Barriers to healthcare access among us adults with mental health challenges: A population-based study. *SSM-population health*, 15:100847, 2021.

[29] Marna S Barrett, Wee-Jhong Chua, Paul Crits-Christoph, Mary Beth Gibbons, and Don Thompson. Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy: Theory, research, practice, training*, 45(2):247, 2008.

[30] Katharine A Smith, Thomas Ward, Sinéad Lambe, Edoardo G Ostinelli, Charlotte Blease, Thomas Gant, Stefan M Gold, Emily A Holmes, Ivana Paccoud, Anastasia Vinnikova, et al. Engagement and attrition in digital mental health: current challenges and potential solutions. *npj Digital Medicine*, 8(1):398, 2025.

[31] Jacinta Jardine, Camille Nadal, Sarah Robinson, Angel Enrique, Marcus Hanratty, and Gavin Doherty. Between rhetoric and reality: real-world barriers to uptake and early engagement in digital mental health interventions. *ACM Transactions on Computer-Human Interaction*, 31(2):1–59, 2024.

[32] Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Sagnik Bhattacharyya, James MacCabe, Stefania Tognin, Ricardo Twumasi, et al. Delusions by design? how everyday ais might be fuelling psychosis (and what can be done about it), 2025.

[33] Joshua Au Yeung, Jacopo Dalmasso, Luca Foschini, Richard JB Dobson, and Zeljko Kraljevic. The psychogenic machine: Simulating ai psychosis, delusion reinforcement and harm enablement in large language models. *arXiv preprint arXiv:2509.10970*, 2025.

[34] Yair Bar-Haim, Dominique Lamy, Lee Pergamin, Marian J Bakermans-Kranenburg, and Marinus H Van Ijzendoorn. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. *Psychological bulletin*, 133(1):1, 2007.

[35] Jonas Everaert, Ioana R Podina, and Ernst HW Koster. A comprehensive meta-analysis of interpretation biases in depression. *Clinical psychology review*, 58:33–48, 2017.

[36] Sven Thönes and Daniel Oberfeld. Time perception in depression: A meta-analysis. *Journal of affective disorders*, 175:359–372, 2015.

[37] Christoph Teufel, Naresh Subramaniam, Veronika Dobler, Jesus Perez, Johanna Finnemann, Puja R Mehta, Ian M Goodyer, and Paul C Fletcher. Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences*, 112(43):13401–13406, 2015.

[38] Grant S Shields, Matthew A Sazma, and Andrew P Yonelinas. The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol. *Neuroscience & Biobehavioral Reviews*, 68: 651–668, 2016.

[39] Robert Malcolm Ross, Ryan McKay, Max Coltheart, and Robyn Langdon. Jumping to conclusions about the beads task? a meta-analysis of delusional ideation and data-gathering. *Schizophrenia bulletin*, 41(5): 1183–1191, 2015.

[40] Lars Schwabe and Oliver T Wolf. Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. *Behavioural brain research*, 219(2):321–328, 2011.

[41] Francesca Gino, Alison Wood Brooks, and Maurice E Schweitzer. Anxiety, advice, and the ability to discern: feeling anxious motivates individuals to seek and use advice. *Journal of personality and social psychology*, 102(3):497, 2012.

[42] Uwe Wolfradt and Thomas Meyer. Interrogative suggestibility, anxiety and dissociation among anxious patients and normal controls. *Personality and Individual Differences*, 25(3):425–432, 1998.

[43] CA Morgan III, J Dule, and YG Rabinowitz. The impact of interrogation stress on compliance and suggestibility in us military special operations personnel. *Ethics, medicine and public health*, 14:100499, 2020.

[44] Christopher A Kelly and Tali Sharot. Web-browsing patterns reflect and shape mood and mental health. *Nature Human Behaviour*, 9(1):133–146, 2025.

[45] Nancy Yang and Bernard Crespi. I tweet, therefore i am: a systematic review on social media use and disorders of the social brain. *BMC psychiatry*, 25(1):95, 2025.

[46] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C Ong, and Nick Haber. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 599–627, 2025.

[47] Annika M Schoene and Cansu Canca. For argument's sake, show me how to harm myself!': Jailbreaking llms in suicide and self-harm contexts. *arXiv preprint arXiv:2507.02990*, 2025.

[48] Edward L Thorndike et al. A constant error in psychological ratings. *Journal of applied psychology*, 4(1): 25–29, 1920.

[49] Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[50] Eric Bogert, Aaron Schecter, and Richard T Watson. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, 11(1):8028, 2021.

[51] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[52] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

[53] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[54] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[55] Lars Kai Hansen and Laura Rieger. Interpretability in intelligent systems–a new concept? In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 41–49. Springer, 2019.

[56] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[57] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–19, 2021.

[58] Laura Cros Vila and Bob Sturm. (mis) communicating with our ai systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2025.

[59] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

[60] Sigmund Freud. Remembering, repeating and working-through (further recommendations on the technique of psycho-analysis ii). In *The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XII (1911–1913): The Case of Schreber, Papers on Technique and Other Works*. London, 1958. Original work published 1914.

[61] James J Gross. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39 (3):281–291, 2002.

[62] Michael White and David Epston. *Narrative means to therapeutic ends*. WW Norton & Company, 1990.

[63] Jerome D Frank, Julia B Frank, and Bruce E Wampold. *Persuasion and healing: A comparative study of psychotherapy*. JhU Press, 2025.

[64] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[65] Rock Yuren Pang, KJ Feng, Shangbin Feng, Chu Li, Weijia Shi, Yulia Tsvetkov, Jeffrey Heer, and Katharina Reinecke. Interactive reasoning: Visualizing and controlling chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2506.23678*, 2025.

[66] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

[67] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[68] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[69] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 81–87, 2018.

[70] Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. Post hoc explanations of language models can improve language models. *Advances in Neural Information Processing Systems*, 36:65468–65483, 2023.

[71] Furui Cheng, Vilém Zouhar, Robin Shing Moon Chan, Daniel Fürst, Hendrik Strobelt, and Mennatallah El-Assady. Interactive analysis of llms using meaningful counterfactuals. *arXiv preprint arXiv:2405.00708*, 2024.

[72] Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. Llms for generating and evaluating counterfactuals: A comprehensive study. *arXiv preprint arXiv:2405.00722*, 2024.

[73] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL `https://www.aclweb.org/anthology/P19-3007`.

[74] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020.

[75] Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: systematic review. *Jmir Ai*, 3:e53207, 2024.

[76] Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M Friedrich, and Felix Nensa. Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. *European journal of radiology*, 162:110786, 2023.

[77] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500, 2024.

[78] Elma Kerz, Sourabh Zanwar, Yu Qiao, and Daniel Wiechmann. Toward explainable ai (xai) for mental health detection based on language behavior. *Frontiers in psychiatry*, 14:1219479, 2023.

[79] Brian Hyeongseok Kim and Chao Wang. Large language models for interpretable mental health diagnosis. *arXiv preprint arXiv:2501.07653*, 2025.

[80] Anthony Kelly, Esben Kjems Jensen, Eoin Martino Grua, Kim Mathiasen, and Pepijn Van de Ven. An interpretable model with probabilistic integrated scoring for mental health treatment prediction: Design study. *JMIR Medical Informatics*, 13:e64617, 2025.

[81] Hao Tang, Aref Miri Rekavandi, Dharjinder Rooprai, Girish Dwivedi, Frank M Sanfilippo, Farid Boussaid, and Mohammed Bennamoun. Analysis and evaluation of explainable artificial intelligence on suicide risk assessment. *Scientific reports*, 14(1):6163, 2024.

[82] Julia Thomas, Antonia Lucht, Jacob Segler, Richard Wundrack, Marcel Miché, Roselind Lieb, Lars Kuchinke, and Gunther Meinlschmidt. An explainable artificial intelligence text classifier for suicidality prediction in youth crisis text line users: Development and validation study. *JMIR Public Health and Surveillance*, 11:e63809, 2025.

[83] Yi Dai, Jinlei Liu, Lei Cao, Yuanyuan Xue, Xin Wang, Yang Ding, Junrui Tian, and Ling Feng. Leveraging social media for real-time interpretable and amendable suicide risk prediction with human-in-the-loop. *IEEE Transactions on Affective Computing*, 2024.

[84] Abdulaziz Ahmed, Mohammad Saleem, Mohammed Alzeen, Badari Birur, Rachel E Fargason, Bradley G Burk, Hannah Rose Harkins, Ahmed Alhassan, and Mohammed Ali Al-Garadi. Leveraging large language models to enhance machine learning interpretability and predictive performance: A case study on emergency department returns for mental health patients. *arXiv preprint arXiv:2502.00025*, 2025.

[85] Maia Jacobs, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):108, 2021.

[86] Ashley Suh, Isabelle Hurley, Nora Smith, and Ho Chit Siu. Fewer than 1% of explainable ai papers validate explainability with humans. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

[87] Lydia A Bazzano, Jaquail Durant, and Paula Rhode Brantley. A modern history of informed consent and the role of key information. *Ochsner Journal*, 21(1):81–85, 2021.

[88] Celia B Fisher and Matthew Oransky. Informed consent to psychotherapy: Protecting the dignity and respecting the autonomy of patients. *Journal of clinical psychology*, 64(5):576–588, 2008.

[89] Manuel Trachsel and Martin grosse Holtforth. How to strengthen patients' meaning response by an ethical informed consent in psychotherapy. *Frontiers in psychology*, 10:1747, 2019.

[90] Judi Chamberlin. *On our own: Patient-controlled alternatives to the mental health system.* McGraw-Hill, 1978.

[91] Sally Clay. *On our own, together: Peer programs for people with mental illness*. Vanderbilt University Press, 2005.

[92] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, pages 49–58, 2005.

[93] Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems*, pages 93–102, 2014.

[94] Donald A Schön. *The reflective practitioner: How professionals think in action*. Routledge, 2017.

[95] Carl Ransom Rogers. *On becoming a person: A therapist's view of psychotherapy*. Houghton Mifflin Harcourt, 1995.

[96] William R Miller and Stephen Rollnick. *Motivational interviewing: Helping people change*. Guilford press, 2012.

[97] Kashmir Hill. *A Teen Was Suicidal. ChatGPT Was the Friend He Confided In.* The New York Times. URL `https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html`.

[98] Georgiana Shick Tryon, Sarah E Birch, and Jay Verkuilen. Meta-analyses of the relation of goal consensus and collaboration to psychotherapy outcome. *Psychotherapy*, 55(4):372, 2018.

[99] Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025.

[100] M Karen Shen and Dongwook Yoon. The dark addiction patterns of current ai chatbot interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

[101] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 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, 2024.

[102] Sachin R Pendse, Logan Stapleton, Neha Kumar, Munmun De Choudhury, and Stevie Chancellor. Advancing a consent-forward paradigm for digital mental health data. *Nature Mental Health*, 2(11): 1298–1307, 2024.

[103] OpenAI. Helping people when they need it most, August 2025. URL `https://openai.com/index/helping-people-when-they-need-it-most/`. OpenAI News. Accessed 2025-10-10.

[104] David Grüning and Julia Kamin. Prosocial design in trust and safety. *arXiv preprint arXiv:2506.12792*, 2025.

[105] Barbara Stanley and Gregory K Brown. Safety planning intervention: a brief intervention to mitigate suicide risk. *Cognitive and behavioral practice*, 19(2):256–264, 2012.

[106] Aisling R Caffrey and Eric P Borrelli. The art and science of drug titration. *Therapeutic advances in drug safety*, 11:2042098620958910, 2020.

[107] Steven Adler. Practical tips for reducing chatbot psychosis. *Clear-Eyed AI* (Substack), October 2025. URL `https://stevenadler.substack.com/p/practical-tips-for-reducing-chatbot`.

[108] Aurelie Klein, Inke Mathauer, Karin Stenberg, and Triin Habicht. Diagnosis-related groups (drg): A question & answer guide on case-based classification and payment systems. Technical report, World Health Organization, Geneva, 2020. URL `https://www.who.int/docs/default-source/health-financing/drg-q-a-guide-final-draft.pdf`.

[109] Insurance implications of DSM-5. Technical report, American Psychiatric Association, 2013. URL `https://www.psychiatry.org/File%20Library/Psychiatrists/Practice/DSM/APA_DSM_Insurance-Implications-of-DSM-5.pdf`.

[110] Alyssa C Milton and Barbara A Mullan. Communication of a mental health diagnosis: a systematic synthesis and narrative review. *Journal of Mental Health*, 23(5):261–270, 2014.

[111] Timothy Jost. Implementing health reform: the appeals process amended rule. *Health Affairs Forefront*, 2011.

[112] Richard Furniss and Sarah Ormond-Walshe. An alternative to the clinical negligence system. *BMJ*, 334 (7590):400–402, 2007.

[113] Frank Frizelle. Review of the health and disability commissioner act and code-your chance to have your say. *The New Zealand Medical Journal (Online)*, 137(1599):13–15, 2024.

[114] The human line project. 2025. URL `https://thehumanlineproject.org/`.

[115] Lucia E Tower. Countertransference. *Journal of the American Psychoanalytic Association*, 4(2):224–255, 1956.

[116] Frank M Dattilio and Michelle A Hanna. Collaboration in cognitive-behavioral therapy. *Journal of Clinical Psychology*, 68(2):146–158, 2012.

[117] C Rick Snyder, Susie C Sympson, Florence C Ybasco, Tyrone F Borders, Michael A Babyak, and Raymond L Higgins. Development and validation of the state hope scale. *Journal of personality and social psychology*, 70(2):321, 1996.

[118] Jessica Lee Schleider, Michael C Mullarkey, and John R Weisz. Virtual reality and web-based growth mindset interventions for adolescent depression: protocol for a three-arm randomized trial. *JMIR research protocols*, 8(7):e13368, 2019.

[119] Jeong Eun Cheon, Yeseul Nam, Kaylyn J Kim, Hae In Lee, Haeyoung Gideon Park, and Young-Hoon Kim. Cultural variability in the attribute framing effect. *Frontiers in psychology*, 12:754265, 2021.

[120] Donghee Shin, Veerisa Chotiyaputta, and Bouziane Zaid. The effects of cultural dimensions on algorithmic news: How do cultural value orientations affect how people perceive algorithms? *Computers in Human Behavior*, 126:107007, 2022.

[121] Ling Kong. Ai companions meet the law: New york and california draw the first lines. Reuters, December 2025. URL https://www.reuters.com/legal/litigation/ai-companions-meet-law-new-york-california-draw-first-lines--pracin-2025-12-23/.