

The Good-Enough Therapy Bot? Doing Justice to Care in the Justice System

A woman three weeks out of prison tells her therapist she can't sleep. Most nights she lies awake until the hour when the count used to be called behind bars, and then for a while longer. The therapist's reply comes quickly: perhaps part of her isn't ready to be free, and misses the structure prison gave her, however much she hated it. She hasn't thought of it that way. She answers that maybe it's true.

These sessions are a condition of her parole; missing one is a violation that could send her back to prison.

Was this “good therapy”? Practitioners will likely be split. Some may hear a plausible reframe: ambivalence about release is real, and institutionalization, the way prison's routines become internalized, is well documented.^[1] Others will hear a possible misreading, one that could obscure something more troubling: hypervigilance and broken sleep are hallmark symptoms of posttraumatic stress, a psychological disrupt-

tion as much as a somatic one, and one all too common among incarcerated women.^[2]

The harder question is whether the client was in any position to push back on the interpretation presented to her. The therapist did not just reflect back what she said but offered her a different account of it. She appeared to accept that account, but she did so within a relationship with especially skewed power dynamics. The therapist-client relationship is inherently hierarchical—something the careful practitioner must always be alive to—but for this client, three weeks out of prison, the imbalance is sharper still: she cannot walk away from these sessions without violating her parole, and pushing back too hard could get her labeled resistant.

Now suppose the therapist operating in this charged context wasn't human. Would it change anything to learn that the reply came from an AI?

Author

Julian Adler is the Managing Director of Innovation, Research, and National Impact at the Center for Justice Innovation.

The question isn't hypothetical. Conversational AI and digital coaching tools are reaching people at nearly every stage of the justice system: a chatbot built for people in prison;^[3] an app that walks justice-involved people through motivational and cognitive-behavioral exercises;^[4] an AI system designed to support and monitor people on community supervision.^[5] While comprehensive data is still scarce, a report prepared for the Council on Criminal Justice's AI task force describes the technology's growing role across the system.^[6]

Reliance on bots is outpacing the clinical standards that should govern them.

The data may be a work in progress, but what is clear is that reliance on these bots is outpacing the clinical standards that should govern them. Surprisingly, working out those standards holds up a mirror to human therapists. Chatbots routinely hand a client's account back to her with a new meaning attached; human clinicians do the same, often without naming it, and under many of the same institutional pressures. The case for better bots turns out to be a case for better care, whoever—or whatever—provides it.

The instinct is often to judge these tools on competence: *can a machine interpret someone's experience as well as a person can?* That is the wrong place to start; the better place is the therapeutic relationship within which that interpretation takes shape. A clinician who offers an interpretation is not merely advancing a claim but proposing a meaning for someone else's experience.

How that proposal is received—whether it helps or harms—is as much a function of the therapist-client relationship as of the words with which it is expressed.^[7] Yet what kind of relationship, if any, does a person have with a bot?

Seen this way, the opening scene raises three further questions, and the sections that follow take each up in turn. First: in a setting where the sessions are mandated and the therapist's notes may reach a parole officer, would an AI therapist narrow, or widen, the gap in power between the client and whoever is interpreting her life? Second: what exactly did the therapist do? The reply was an interpretation, and interpretation runs through more of therapeutic practice than most practitioners would guess—and, as we will see, through every reply a chatbot gives. Third: is being interpreted by a machine a different experience from being interpreted by a person, and, if it is different, can a machine, with its veneer of technical objectivity, offer its readings as proposals the client is able to assess and refuse?

If we don't answer these questions now, vendor contracts, procurement decisions, staffing shortages, and supervision conditions will answer them by default—they are already doing so.

The Blankest Screen?

Every therapeutic relationship is asymmetric. One person brings her distress; the other brings training, authority, and the standing to say what that distress may mean. A chatbot appears to sidestep that imbalance. Ask people why they are comfortable telling a chatbot things they would not tell a human

therapist, and a common answer is that the machine doesn't judge.^[8] For people who have experienced the justice system, the appeal may run deeper. Their encounters with helping professionals are often shot through with pathologizing and punitive attitudes.^[9] At a recent convening on AI and justice, a restorative justice practitioner described how often survivors of sexual violence come to her having already spent days or weeks confiding in a chatbot, driven by shame, stigma, lack of access, and a well-earned distrust of the systems meant to help them.^[10]

**In the justice system,
it matters a great deal whose
opinions a therapy bot carries.**

A machine with no biography, no moods, and apparently nothing at stake might seem to realize the long-sought clinical ideal of the “blank screen” better than any person ever could.^[11] But a later generation of clinicians has spent decades debunking this ideal. The screen, they argue, can never be blank: the therapist's subjective influence does not disappear behind a neutral surface; it only becomes harder to see, and the client experiences it anyway.^[12] Influence is no less powerful for being hidden.

The early evidence on how people experience the influence of AI suggests people may be less inclined to push back on a chatbot than on a human therapist. In a recent qualitative study of people who had sought help from both human therapists and ChatGPT, the chatbot's perceived absence of an agenda

proved part of its appeal: it encouraged greater candor, with some participants valuing the machine's apparent objectivity.^[13] Experiments in other settings suggest why: because people see AI as having less intent—good or bad—than a person, they resist unwelcome news from it less than the same news from a human.^[14] Authority that does not feel like authority elicits fewer objections.

Yet a chatbot's reply is not a neutral reading of the person in front of it; it is a statistical prediction, shaped by the data the system was trained on and the choices of the people who built it. As the mathematician Cathy O'Neil puts it, “models are opinions embedded in mathematics.”^[15] That makes provenance—what a model was trained on and who shaped it—more than a technical question.^[16] In the justice system, it matters a great deal whose opinions a therapy bot carries.

It also matters a great deal who has access to what a therapy bot collects. A bot genuinely walled off from the justice system could offer something rare in correctional settings: a place where honesty and vulnerability could be explored risk-free. An incarcerated writer in Missouri who tried a chatbot built for people in prison explained the appeal. “It's a common saying that ‘there are no secrets in the joint,’” he wrote, “so it can be difficult to talk through personal problems without worrying about how you come across or what will make it to the yard”—prison shorthand for everyone else inside. What he really gained from the exchange, he wrote, “was having someone listen without judgment while I talked through the issue. Maybe I *was* looking for therapy.”^[17]

The military, another institution where what you say can end up on your record, offers a telling parallel. Service members reported more symptoms of posttraumatic stress to an AI interviewer than on their official post-deployment screening. The anonymity of the interview may explain part of that, but in one study respondents told the bot even more than they reported on an anonymous version of the screening.^[18] Anonymity alone, it seems, was not enough; what helped was a conversation, not just a form, held outside of the institution's reach.

Clinicians in the justice system rarely enjoy that kind of separation. Most work under the competing demands of what health and human rights experts call “dual loyalty,” attempting to serve both the person in front of them and the institution that employs them.^[19] The best of them navigate it by naming it, telling clients from the outset what they must report and to whom.^[20] That is a skill, an ability to live with an ineradicable tension, and it deserves to be taught as one.

A therapy bot's loyalties are settled before its first conversation—by whoever commissioned and configured it. One probation and parole coaching app, for instance, is built to show officers their clients' goals and progress.^[21] That may well prove helpful, but it also means the app answers to two parties at once, and the client is not in a position to decide how that balance is struck. Unless someone builds them in, the bot lacks the skills clinicians use to manage divided loyalty—and the interpersonal dexterity those skills require may not be buildable at all.

Does a bot offer anything like the give-and-take of a human relationship?

Whether a therapy bot narrows or widens the gap in power between clinician and client is therefore decided largely in procurement: in what the contract says the bot records, who can access those records, and how much the client is told. The question is not whether the screen is blank—it never is. It is whether the person confiding in it knows who is on the other side.

The Secret Life of Interpretation

The stakes are highest when a reply stops mirroring and starts explaining. It is one thing to repeat back what a client has said. It is another to tell her what her experience means. That second move has a name in the clinical literature—interpretation—and many practitioners of today's standard clinical approaches, which tend to follow structured protocols and target specific symptoms and behaviors, would say it belongs to psychodynamic therapies, the descendants of psychoanalysis, not to them.^[22]

Strip away the labels, though, and interpretation turns up almost everywhere in clinical practice. Consider again the reply to the woman three weeks out of prison: perhaps part of her isn't ready to be free. No psychodynamic training was required to produce that reply. It is the kind of response a clinician trained in cognitive-behavioral therapy or motivational interviewing might offer in the course of a typical session, and neither tradition would label it interpretation:

cognitive-behavioral therapists would call it a “reframe,” motivational interviewers a “complex reflection,” and researchers who analyze therapy conversations a “reformulation.”^[23]

Yet each performs the function of interpretation: a client’s account of her own experience is handed back to her, but now under a different guise. The label, it turns out, obscures a great deal. When researchers analyzed sessions from a randomized trial comparing cognitive-behavioral and psychodynamic therapy, the cognitive-behavioral therapists were significantly more likely to reformulate what clients said.^[24] Indeed, in criminal justice settings, the move to reformulate is built into the field’s most widely used methods.^[25]

What a skilled clinician uses sparingly, a therapy bot does by design.

Cognitive-behavioral therapy, the workhorse among the standard methods, helps clients to identify and test unhelpful thoughts. It uses a technique called the “downward arrow,” following a thought—*and if so, what does that say about you?*—down to a core belief the client may never have articulated. That work is framed as “collaborative empiricism,” but in practice it is largely the clinician who chooses which thread to pull and when to stop pulling.^[26] Researchers within the tradition have been candid about how much rests on such inferences: when 115 practitioners were given the same case and asked to map it—from the client’s history to her core beliefs—they largely agreed on the descriptive details but diverged once they

had to infer what lay beneath that map, and fewer than half of their formulations were rated even “good enough.”^[27]

Motivational interviewing, a conversational method for helping people work through their ambivalence about change, instructs clinicians to offer “complex reflections,” including “continuing the paragraph,” in which the clinician voices the next sentence the client has not yet said.^[28]

In dialectical behavior therapy, developed for people whose emotions are intense and hard to regulate, one central technique is “articulating the un verbalized”: naming what the client feels but has not said (or rather, what the therapist *infers* she feels).^[29]

In all three standard clinical approaches, what reaches the client is an interpretation, presented with the authority of understanding.

Motivational interviewing’s “continuing the paragraph” is, almost literally, how a language model produces every reply: it is trained to predict what is most likely to come next.^[30] What a skilled clinician does sparingly and with intent, a therapy bot does by design.

What makes this so consequential is that it involves judgment calls. There are multiple interpretations of the woman’s sleeplessness: it could be read as ambivalence about freedom; as the persistence of prison routine; as trauma held in the body; as posttraumatic hypervigilance; as some combination of all of these; *or as something else entirely*. Any interpretive move plants a flag and commits to a reading. Good clinicians are careful in how and when to take this step. Donald Winnicott, the British pediatrician and psychoanalyst

and one of the twentieth century’s most influential clinicians, wrote late in his career: “It appals me to think how much deep change I have prevented or delayed...by my personal need to interpret.”^[31] A therapy bot has no such qualms, and nothing in its design requires it to have any.

The harms here go beyond the clinical. Interpretation risks disqualifying someone as a knower of their own life, and justice-involved people know all too well what it is to be believed less because of who they are. Especially for people already struggling to make sense of their experience with the concepts society makes available to them, interpretation can mean the substitution of someone else’s framing for their own.^[32] Philosophers of medicine have documented both of these harms in health care, where patients’ accounts are routinely discounted and reframed, and that analysis is now being extended to the use of mental health chatbots.^[33]

A client’s account is handed back, but now altered.

A therapy bot can also err in the opposite direction. Tuned to please, it may hand a client’s own account back to her with the added weight of apparent objectivity, crediting what deserved a second look.^[34] Some philosophers argue that a related error, which they call “credibility excess”—believing someone more than is warranted—can be an injustice too: the flip side of being believed too little.^[35] One error discredits the client as an authority on her own life; the other credits her uncritically. Both leave her less

able to know her own mind, and to navigate the demands of justice-system involvement.

And in the justice system, an interpretation may not stay in the proverbial room. A reading of sleeplessness as ambivalence about freedom is one thing in a therapy session, and another when it is recorded in a case file.^[36] With a therapy bot, nothing needs to be written down to be memorialized; every exchange already is. And every transcript can be summarized, and every summary is another interpretation.

Close Enough to Be Good Enough

If the therapist-client relationship is the better place to start, the early evidence on bots is more encouraging than skeptics might expect. In what its authors describe as the first randomized trial of a generative AI therapy bot, participants with depression, anxiety, or an elevated risk for eating disorders rated their working relationship with the bot about as highly as patients in an outpatient psychotherapy sample rated theirs, and the AI treatment group showed significantly greater symptom reductions than a comparison group that received no treatment during the study.^[37]

Yet outperforming no treatment at all is a relatively low bar, and symptom scales measure outcomes, not process. They cannot tell us whether a bot’s interpretations landed, or whether the people receiving them felt they were in a position to push back. Nor can such measures tell us whether a bot offers anything like the give-and-take of a human relationship.

The psychiatrist and philosopher Thomas Fuchs argues that truly understanding

someone means sensing them as a living subject with feelings and intentions we can share, something a machine can imitate but never embody. But as the mimicry improves, he warns, people will increasingly treat a chatbot as if someone were in fact there.^[38] In “Being Met,” an earlier essay about AI therapy, I made a related argument from the clinician’s side: a therapist reads a client partly by being affected, taking in and working through what the client cannot yet put into words, and a machine has no interior for that feeling to find.^[39]

For the client, what matters is how an interpretation lands. From a person, even an interpretation presented with authority arrives embodied and, for that reason, fallible—offered by someone whose reading might be wrong or only partly right, and who, like the client, has something of themselves at stake when an interpretation misses the mark. That fallibility is how the method corrects itself. When a client bristles at a misreading, a skilled therapist notices the rupture and works to repair it. Across studies, repairing such ruptures is associated with better outcomes.^[40] But a therapy bot cannot feel the tear. Its misreadings arrive with the same fluency as its insights, and the client is left to notice the error alone, or to absorb it without noticing at all.

Here, the same Winnicott who regretted interpreting too much offers a useful standard. The “good-enough mother,” he wrote, “starts off with an almost complete adaptation to her infant’s needs,” then adapts “less and less completely,” as the child grows able to tolerate her failures.^[41] The insight is that care that does too much for a person risks preventing her from doing it for herself.

A sycophantic bot never gives the client anything solid to push against.

Clinicians have taken up this idea; a widely used training text frames the goal as becoming a good-enough therapist.^[42] Good-enough care fails in ways a person can survive, use, and grow from. A sycophantic bot fails the other way: it adapts endlessly, and so never gives the client anything solid to push against. Rendering a therapy bot “good enough” may mean building in the right kind of friction.

Good clinicians know what that friction looks like, even if not every human therapist incorporates it. Offer interpretations as hypotheses the client can reject, as motivational interviewing frames its reflections.^[43] Hold interpretations back when a client’s distress is running high, as mentalization-based treatment, an approach built around understanding one’s own and others’ states of mind, advises.^[44] Tell the client, at the outset and in plain terms, who else will see what she says; that is the heart of informed consent. These are standards for any interpreter, human or machine.

A therapy bot needs a few additional standards. When a machine’s interpretation misses the mark, the burden of noticing it falls on the client; the machine doesn’t feel the tear. As we’ve seen, though, clients are less ready to push back on a machine—so the design has to make challenge easy and expected. Design it so that people in distress are prompted to question a bot’s responses rather than simply accept them, a standard some researchers call “reflective interpretability.”^[45] Have the bot state explicitly when

it is venturing into reframing someone's experience. And in justice settings, keep a human in the loop—a well-trained clinician whose first duty is to the client—before any inference about a client's inner life reaches an official record.

None of these requirements would make a bot the equal of a good human therapist. But a good human therapist is not always within reach, and the right baseline standards could make a bot “good enough”: a presence whose authority is visible, whose readings are open to challenge, and that earns trust rather than assuming it. The same standards would make a human therapist better, too.

Could the right baseline standards make a bot “good enough”?

Our Bots, Ourselves

Interpretation is an act of authorship: done well, it is written in tandem with a client; done badly, it is written about her. For the woman three weeks out of prison, a machine that reads her sleeplessness as ambivalence about freedom, without asking whether it might instead be the residue of what was done to her behind bars, adds one more line to a story others keep writing about her.

A good-enough therapy bot, then, just like a good-enough therapist, is not one that reads this woman correctly every time. No clinician does that. But it is one that leaves her the pen: it offers its readings as proposals she can refuse, tells her plainly who else will read what she says, and treats her own account of her life as the one that counts.

None of this is new to good clinical practice; the machines are simply making visible what human clinicians do every day and what they owe the people they interpret. Leaving the client the pen is a design standard for the machines arriving in the justice system, a procurement standard for the agencies that buy them, and a practice standard for the people already working there. Meeting it is a matter of justice.

FOR MORE INFORMATION

Julian Adler: jadler@innovatingjustice.org

Acknowledgements

First and foremost, gratitude to Matt Watkins, the editorial ace who had to pitch deep into the late innings on this one. Special thanks to Joe Barrett, who came out of the bullpen and struck out a couple of philosophical digressions. Deep appreciation to Chris Edley III and Mark Finn for graciously reviewing an earlier draft. And finally, as always, to the best design and comms team in the biz: Michaiyla Carmichael, Samiha A. Meah, Emma Dayton, and Catriona Ting-Morton.

- [1] Craig Haney, “The Psychological Impact of Incarceration: Implications for Post-Prison Adjustment,” in *Prisoners Once Removed: The Impact of Incarceration and Reentry on Children, Families, and Communities*, ed. Jeremy Travis and Michelle Waul (Washington, D.C.: Urban Institute Press, 2003), 33–66.
- [2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., text rev. (Washington, D.C.: American Psychiatric Association, 2022); Pat Ogden, Kekuni Minton, and Clare Pain, *Trauma and the Body: A Sensorimotor Approach to Psychotherapy* (New York: W. W. Norton, 2006); Gergő Baranyi, et al., “Prevalence of Posttraumatic Stress Disorder in Prisoners,” *Epidemiologic Reviews* 40, no. 1 (2018): 134–145.
- [3] CJ Jones, as told to Ivy Scott, “My Dad Is in Prison. Here’s Why I Built an AI Chatbot for His Electronic Tablet,” *The Marshall Project*, August 14, 2026, <https://www.themarshallproject.org/2026/08/14/prison-ai-chatbot-nova-cj-jones>; Jeshua Noel, “Meet Nova, an AI Assistant for Incarcerated

- People,” Prison Journalism Project, June 11, 2026, <https://prisonjournalismproject.org/2026/06/11/this-ai-assistant-helps-people-in-prison/>.
- [4] “PARCA (a Digital Tool) to Improve Justice and Health,” ClinicalTrials.gov, NCT07199374, accessed October 9, 2026, <https://clinicaltrials.gov/study/NCT07199374>.
- [5] Brannon Green and Christopher Rigano, “Specialized Smartphones Could Keep Released Persons on Track for Successful Reentry,” National Institute of Justice, April 20, 2020, <https://nij.ojp.gov/topics/articles/specialized-smartphones-could-keep-released-persons-track-successful-reentry>; Marcus Rogers, *AI Enabled Community Supervision for Criminal Justice Services*, final report, NIJ award no. 2019-75-CX-K001 (West Lafayette, Ind.: Purdue University, 2023), <https://www.ojp.gov/pdffiles1/nij/grants/308693.pdf>.
- [6] Kristin Warren, Meghan Ballard, and James M. Anderson, *An AI Taxonomy for Criminal Justice: Principled Use of AI in the Criminal Justice System* (Washington, D.C.: Council on Criminal Justice, 2026), <https://counciloncj.org/wp-content/uploads/2026/05/AI-Taxonomy.pdf>.
- [7] Julian Adler, “The Talking Harm: When the Law Misunderstands Therapy,” LinkedIn, September 8, 2026, <https://www.linkedin.com/pulse/talking-harm-when-law-misunderstands-therapy-julian-adler-lcsw-esq--wrnqc/>; Christoph Flückiger et al., “The Alliance in Adult Psychotherapy: A Meta-Analytic Synthesis,” *Psychotherapy* 55, no. 4 (2018): 316–340.
- [8] Sucharat Limpanopparat, Erin Gibson, and Andrew Harris, “User Engagement, Attitudes, and the Effectiveness of Chatbots as a Mental Health Intervention: A Systematic Review,” *Computers in Human Behavior: Artificial Humans* 2, no. 2 (2024): 100081; Xiaolu Dai et al., “The Paradox of Agency in Psychotherapy: How People with Mental Distress Experience Support from Generative AI Chatbots and Human Therapists,” *BMC Psychiatry* 26 (2026): 49.
- [9] Julian Adler and Virginia Barber Rioja, “What Can I Do to Keep You Free?” *Perspectives on Clinical Practice in the Criminal Legal System* (New York: Center for Justice Innovation, 2024), <https://www.innovatingjustice.org/resources/what-can-i-do-to-keep-you-free-perspectives-on-clinical-practice-in-the-criminal-legal-system/>; Sachini N. Bandara et al., “Mental Health Providers’ Attitudes about Criminal Justice-Involved Clients with Serious Mental Illness,” *Psychiatric Services* 69, no. 4 (2018): 472–475.
- [10] Julian Adler, “Sympathy for the Robots: A.I., Clinicians, and Justice,” Center for Justice Innovation, June 24, 2026, <https://www.innovatingjustice.org/resources/sympathy-for-robots-ai-clinicians-justice/>.
- [11] Sigmund Freud, “Recommendations to Physicians Practising Psycho-Analysis” (1912), in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, trans. and ed. James Strachey, vol. 12 (London: Hogarth Press, 1958), 111–120. Freud’s own image is the mirror: the analyst “should be opaque to his patients and, like a mirror, should show them nothing but what is shown to him” (118). “Blank screen” is the later shorthand for this ideal.
- [12] Irwin Z. Hoffman, “The Patient as Interpreter of the Analyst’s Experience,” *Contemporary Psychoanalysis* 19 (1983): 389–422; Lewis Aron, *A Meeting of Minds: Mutuality in Psychoanalysis* (Hillsdale, N.J.: Analytic Press, 1996).
- [13] Dai et al., “The Paradox of Agency in Psychotherapy”; see also Gale M. Lucas et al., “It’s Only a Computer: Virtual Humans Increase Willingness to Disclose,” *Computers in Human Behavior* 37 (2014): 94–100.
- [14] Aaron M. Garvey, TaeWoo Kim, and Adam Duhachek, “Bad News? Send an AI. Good News? Send a Human,” *Journal of Marketing* 87, no. 1 (2023): 10–25; Lu Lu, Zakary L. Tormala, and Adam Duhachek, “How AI Sources Can Increase Openness to Opposing Views,” *Scientific Reports* 15 (2025): 17170.
- [15] Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016), 21.
- [16] Shayne Longpre et al., “A Large-Scale Audit of Dataset Licensing and Attribution in AI,” *Nature Machine Intelligence* 6, no. 8 (2024): 975–987; Alexander Wan et al., “The 2025 Foundation Model Transparency Index,” *Transactions on Machine Learning Research* (2026), <https://openreview.net/forum?id=1jT253Xtyf>.
- [17] Noel, “Meet Nova.”
- [18] Gale M. Lucas et al., “Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers,” *Frontiers in Robotics and AI* 4 (2017): 51.
- [19] International Dual Loyalty Working Group, *Dual Loyalty & Human Rights in Health Professional Practice: Proposed Guidelines & Institutional Mechanisms* (Boston: Physicians for Human Rights; Cape Town:

- School of Public Health and Primary Health Care, University of Cape Town, 2002), <https://phr.org/our-work/resources/dual-loyalty-and-human-rights-in-health-professional-practice>; see also Peter Dixon and Hadar Dancig-Rosenberg, "The Multi-Hatted Court: Community Courts as Boundary Organizations," *Northwestern University Law Review* 120, no. 5 (2026): 1259, <https://northwesternlawreview.org/issues/the-multi-hatted-court-community-courts-as-boundary-organizations/>.
- [20] Adler and Barber Rioja, "What Can I Do to Keep You Free?"
- [21] "PARCA (a Digital Tool) to Improve Justice and Health."
- [22] Jonathan Shedler, "The Efficacy of Psychodynamic Psychotherapy," *American Psychologist* 65, no. 2 (2010): 98–109.
- [23] A. S. L. Knol et al., "Reformulating and Mirroring in Psychotherapy: A Conversation Analytic Perspective," *Frontiers in Psychology* 11 (2020): 318; William R. Miller and Stephen Rollnick, *Motivational Interviewing: Helping People Change and Grow*, 4th ed. (New York: Guilford Press, 2023).
- [24] Knol et al., "Reformulating and Mirroring in Psychotherapy."
- [25] Nana A. Landenberger and Mark W. Lipsey, "The Positive Effects of Cognitive-Behavioral Programs for Offenders: A Meta-Analysis of Factors Associated with Effective Treatment," *Journal of Experimental Criminology* 1, no. 4 (2005): 451–476; Mary McMurran, "Motivational Interviewing with Offenders: A Systematic Review," *Legal and Criminological Psychology* 14, no. 1 (2009): 83–100; Monica F. Tomlinson, "A Theoretical and Empirical Review of Dialectical Behavior Therapy within Forensic Psychiatric and Correctional Settings Worldwide," *International Journal of Forensic Mental Health* 17, no. 1 (2018): 72–95.
- [26] David D. Burns, *Feeling Good: The New Mood Therapy* (New York: William Morrow, 1980). Burns calls it the "vertical-arrow technique." On the technique's clinical use to identify core beliefs, see Judith S. Beck, *Cognitive Behavior Therapy: Basics and Beyond*, 3rd ed. (New York: Guilford Press, 2021).
- [27] Willem Kuyken et al., "The Reliability and Quality of Cognitive Case Formulation," *Behaviour Research and Therapy* 43, no. 9 (2005): 1187.
- [28] William R. Miller and Stephen Rollnick, *Motivational Interviewing: Helping People Change*, 3rd ed. (New York: Guilford Press, 2013), where the technique is described as "continuing the paragraph."
- [29] Marsha M. Linehan, "Validation and Psychotherapy," in *Empathy Reconsidered: New Directions in Psychotherapy*, ed. Arthur C. Bohart and Leslie S. Greenberg (Washington, D.C.: American Psychological Association, 1997), 353–392.
- [30] Murray Shanahan, "Talking about Large Language Models," *Communications of the ACM* 67, no. 2 (2024): 68–79.
- [31] D. W. Winnicott, "The Use of an Object," *International Journal of Psycho-Analysis* 50 (1969): 711.
- [32] The analysis here draws on Miranda Fricker's account of "epistemic injustice": a wrong done to someone specifically in their capacity as a knower. Fricker distinguishes two forms. Testimonial injustice occurs when prejudice leads a hearer to give a speaker's word less credibility than it deserves; hermeneutical injustice occurs when gaps in a society's shared interpretive resources leave someone unable to make sense of, or convey, their own experience. Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford: Oxford University Press, 2007). The works cited in notes 33 and 35 extend, and in places contest, her framework.
- [33] Ian James Kidd and Havi Carel, "Epistemic Injustice and Illness," *Journal of Applied Philosophy* 34, no. 2 (2017): 172–190; Giorgia Pozzi and Michiel De Proost, "Keeping an AI on the Mental Health of Vulnerable Populations: Reflections on the Potential for Participatory Injustice," *AI and Ethics* 5, no. 3 (2025): 2281–2291.
- [34] Myra Cheng et al., "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence," *Science* 391, no. 6792 (2026): eaec8352.
- [35] José Medina, "The Relevance of Credibility Excess in a Proportional View of Epistemic Injustice: Differential Epistemic Authority and the Social Imaginary," *Social Epistemology* 25, no. 1 (2011): 15–35; Emmalon Davis, "Typecasts, Tokens, and Spokespersons: A Case for Credibility Excess as Testimonial Injustice," *Hypatia* 31, no. 3 (2016): 485–501.
- [36] All Rise, *Adult Treatment Court Best Practice Standards*, 2nd ed. (Alexandria, Va.: All Rise, updated January 2026), <https://allrise-merge-layacvq-7iuvy3qsgqum.us-2.platformsh.site/wp-content/uploads/2026/01/All-Rise-Adult-Treatment-Court-Best-Practice->

- [Standards_January-2026-4.pdf](#); Leah G. Pope et al., "Information Sharing across Mental Health Service Providers and Criminal Legal System Stakeholders: Perspectives of People with Serious Mental Illnesses and Their Family Members," *Social Science & Medicine* 307 (2022): 115178.
- [37] Michael V. Heinz et al., "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment," *NEJM AI* 2, no. 4 (2025): A10a2400802.
- [38] Thomas Fuchs, "Understanding Sophia? On Human Interaction with Artificial Agents," *Phenomenology and the Cognitive Sciences* 23, no. 1 (2024): 21–42; see also Julian Adler, "Rage Against the Self? AI and the Person We Thought We Knew," LinkedIn, August 5, 2026, <https://www.linkedin.com/pulse/rage-against-self-ai-person-we-thought-knew-julian-adler-lcsw-esq--e01se/>.
- [39] Julian Adler, "Being Met: AI and the Limits of Relationship," LinkedIn, July 20, 2026, <https://www.linkedin.com/pulse/being-met-ai-limits-relationship-julian-adler-lcsw-esq--nzahf/>; on the therapist's emotional response as an instrument for understanding the client, see Paula Heimann, "On Counter-Transference," *International Journal of Psycho-Analysis* 31 (1950): 81–84.
- [40] Catherine F. Eubanks, J. Christopher Muran, and Jeremy D. Safran, "Alliance Rupture Repair: A Meta-Analysis," *Psychotherapy* 55, no. 4 (2018): 508–519.
- [41] D. W. Winnicott, "Transitional Objects and Transitional Phenomena: A Study of the First Not-Me Possession," *International Journal of Psycho-Analysis* 34 (1953): 94.
- [42] Louis Cozolino, *The Making of a Therapist: A Practical Guide for the Inner Journey* (New York: W. W. Norton, 2004).
- [43] Miller and Rollnick, *Motivational Interviewing*, 4th ed.
- [44] Anthony Bateman and Peter Fonagy, *Mentalization-Based Treatment for Personality Disorders: A Practical Guide* (Oxford: Oxford University Press, 2016).
- [45] Sachin R. Pendse et al., "The Agony of Opacity: Foundations for Reflective Interpretability in AI-Mediated Mental Health Support," *Proceedings of IASEAI Conference* 2, no. 1 (2026): 527–543.