Center for Justice Innovation

Sympathy for the Robots

A.I., Clinicians, and Justice

The largest mental health "provider" in the United States is either a jail in Los Angeles or an artificial intelligence chatbot—depending on how you count.

On any given day, the Los Angeles County jail system houses just under 6,000 people with documented mental health needs, making it the country's largest de facto psychiatric institution.[1] Earlier this year, OpenAI reported roughly 900 million weekly users; against a base that large, even a modest share turning to the technology for emotional support amounts to millions of people.[2] Consistent with findings from RAND, Pew, and others documenting the rapid uptake of A.I. for mental health support, a 2025 peer-reviewed study reported that nearly half of U.S. adults who had used large language models for any purpose were using them for therapeutic support.[3]

Artificial intelligence systems such as ChatGPT are neither designed nor regulated for therapeutic work, and they have been documented to fail in sometimes dangerous ways. But the examples of the jail and the chatbot point to the same underlying truth: when large-scale social support systems like community mental health treatment fail, something else fills the gap. National surveys suggest that more than half of people in prisons and jails have a history of psychiatric disorders or experience serious psychological distress, and the prevalence of serious mental illness among people in U.S. jails is estimated to be four to seven times higher than in the general population.[4] At a recent convening on A.I. and justice, a restorative justice practitioner in New York City described how often survivors of sexual violence come to her having already spent days or weeks talking to an A.I. chatbot—out of shame, stigma, lack of access, or a deep, well-earned distrust of the human systems nominally available to them.[5]

Among Americans using A.I. for health information, a 2026 Kaiser Family Foundation poll found that both adults with low incomes and adults under 30 dispro-

Author

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

portionately cite affordability and barriers to access as motivations, and 58 percent of people who asked A.I. about mental health did not follow up with a human provider afterward—suggesting A.I. is functioning as a replacement, not a bridge.[6] The people most likely to be misread, mishandled, or harmed by A.I. in helping contexts are the same people that human systems have long misread, mishandled, and harmed. This is not a coincidence; these failures share a root.

> The real work requires an investment in higher-quality care by humans *and* machines.

The critiques most commonly leveled at A.I. in clinical contexts are that it pathologizes, that it is sycophantic, that it cannot balance the tension between empathy and accountability, that it misses what matters most. Yet these are critiques that human clinicians could be leveling at themselves. A.I. is reproducing patterns we were the first to construct. This admission does not absolve the technology; A.I. carries its own real risks, particularly when it is deployed at scale and in contexts where life and liberty are on the line (if it is not happening already, it is not hard to imagine A.I.-based treatment becoming a court-mandated condition). But the acknowledgement that the problem starts with *human* practices changes what the conversation needs to be about.

The question is not whether A.I. can be made to do what humans already do well—because we often struggle to be good-enough clinicians. Instead, there is an imperative for the builders of A.I., practicing clinicians, and the people on the receiving end of both to do, together, what no party has yet figured out how to do alone: *markedly improve the quality and safety of therapeutic practices and radically democratize their access*—whether the means of that access is human, or A.I., or a hybrid. But accomplishing that, especially in higher-risk contexts like the justice system, requires us to get at least three things right—to fix three stubborn bugs in our code.

Empathy Can Be Tricky

For practitioners who work closely with people in distress, the message that empathy is the crux of helping, and that more of it is always better, runs deep. The research mostly backs that up. The largest meta-analysis of its kind looked at more than 80 studies and more than 6,000 clients in therapy and found a consistent, moderately strong link between how empathic a clinician is and how well their clients fare.[7] So the basic instinct is right: empathy matters, and it matters especially for people whose contact with legal systems has too often involved being stigmatized, judged, or ignored.[8] But a newer wave of research complicates this picture. When researchers chart empathy against outcomes, the line does not just keep climbing as more empathy is "added." At a certain point, the line begins to level off, and at the very top end sometimes starts to dip. One careful study even tried to identify the optimal dose of empathy: not too little, not too much, but somewhere in the upper middle.[9]

When empathy gets too intense, it can shade into something harmful: feeling the other person's pain as one's own, losing the critical distance that makes it possible to be useful,

and struggling to say the hard thing because adding to someone's burden feels unbearable.[10] Research has shown that empathic *acuity*—how well a clinician reads what the person in front of them is actually feeling in the moment—is a better predictor of positive outcomes than the *amount* of empathy a helper offers.[11] Intense emotional attunement that misreads the client can leave that person feeling more alone, not less. This matters especially for people with histories of trauma, who can experience a helper's outsized feelings about their situation as flooding or destabilizing—and trauma histories are the rule rather than the exception in legal-system-involved populations.[12] The lesson to take from this is not to be detached. It is that empathy works best when it is calibrated and paired with steadiness—when it is possible to stay close to someone's experience without getting pulled under by it.

Large language models are notoriously good at reproducing the *surface* of empathy—the validating, attentive language that signals care—and this is part of why people turn to them for support. But they have a well-documented tendency toward sycophancy—an inclination to affirm and flatter the user, even when the user is wrong, distressed, or describing something dangerous. A 2025 Stanford study put several popular therapy chatbots through controlled tests and found that they routinely failed to push back against delusional thinking, missed clear signals of suicidal ideation, and offered validation in moments when a skilled human helper would have intervened.[13]

This is an indictment of A.I. in helping contexts. But, again, the critique cuts uncomfortably close to home. A recent reflection on A.I. and therapy by two clinical psychologists put it succinctly: "A.I. didn't invent… avoidance. It learned it from us."[14] What hurts clients—whether from a chatbot or a human—is warmth without accuracy, validation without perspective, and attunement without the nerve to say the hard, necessary thing. If skilled human helpers—with years of training, supervision, and hard-won clinical judgment—still have to actively resist the pull toward over-empathizing, it should not be surprising that A.I. systems struggle with the same calibration problem.

> Clinicians have something profound to gain from the effort around A.I., if we're willing to be taught.

The good news is that researchers, clinicians, and developers are converging on a set of design directions that take this calibration challenge seriously. Sycophancy can be reduced—sometimes substantially—by retraining models on examples of constructive disagreement, where the system is rewarded for respectful pushback rather than empty validation.[15] A leading clinical neuroscientist has called for mandatory safeguards for emotionally-responsive A.I. used in mental health contexts: consistent reminders that the system is not human, pattern detection for crisis signals with handoffs to professional human help, hard conversational limits in high-risk territory, and routine clinician-led audits of system behavior—recommendations the American Psychological Association and American Medical Association have echoed in separate advisories calling for embedding chatbots

within services rather than allowing them to operate as standalone providers.[16] Across these recommendations, one principle keeps surfacing: the people most likely to design an A.I. that actually helps are people who already understand what makes *human* helping work, what calibration for a human therapist looks like: when to lean in, when to hold steady, and when to challenge.

Social Determinants Can Be Well Disguised

There is a deeper problem underneath the empathy question, one that A.I. is bringing into sharp relief. You see it most clearly in the legal system. People present with what appears to be a behavioral health crisis: agitation, withdrawal, hypervigilance, suspicion of institutions, dysregulated affect, substance use, despair. The dominant frame in most service settings encourages helpers to read these presentations as symptoms of a condition that lives *inside* of the person—something to be diagnosed and treated. Yet a substantial portion of what gets read as individual pathology is, for many populations, a rational response to circumstances that would destabilize anyone: chronic housing instability, lost benefits, family separation by child welfare or carceral systems, food insecurity, untreated medical conditions, and the cumulative weight of being processed through systems that have repeatedly failed you. This list contains elements of what the field calls the social determinants of mental health.[17] Reading the responses to these determinants as symptoms that originated wholly, or even substantially, within the person locates the problem in the wrong place and prescribes the wrong intervention.

Clinicians have known this for years, but still struggle to act on it. The structural competency framework—the ability to recognize how the structures constraining and shaping a person's life are producing their symptoms—has gained traction and is now part of the curriculum in a growing number of psychiatry residencies, social work programs, and public health schools.[18] This is progress. Yet even seasoned helpers steeped in this scholarship can struggle to maintain the structural-clinical distinction in practice. Caseload pressure, documentation requirements, reimbursement structures that pay for billable diagnoses but not for advocacy, and professional training that defaults to individualized etiologies all tug hard in the same direction.[19]

> An A.I. that misreads structural distress as individual pathology is being all *too* human.

Do not look to A.I. to muster much by way of resistance. Imagine a chatbot in a program for people returning from jail or prison. If it is asked about sleep problems, panic, or substance use, it will reach for the frames it has been trained on—and those are diagnostic, not structural. Designing against this would take more than one fix. The training data has to change: models built for these settings need the structural competency literature, the social-determinants framework, and the lived testimony of system-involved people as their *primary* inputs, not as a thin corrective layer. Yet nor is this sufficient—structural

competency is not acquired by digesting more structural literature, but by supervised practice that interrogates one's own reflexes, and that kind of learning resists reduction to text (and may not be functionally within reach to an A.I.). The model also has to ask the right questions: any tool serving these populations should raise housing, food, benefits, and family separation up front.[20] And the design process needs primary input from the people who know this terrain—structurally trained clinicians, social workers, peer specialists with lived experience, and the people the tool is meant to serve. None of this is easy—and the difficulty is instructive. An A.I. that misreads structural distress as individual pathology is not failing to be human. It is being all *too* human.

Good Practitioners Can—And Do—Go Off Script

The best practitioners depart from models. Not occasionally, not as a lapse in competent practice, but routinely and as an *expression* of good practice. A skilled case manager trained in motivational interviewing will, in the middle of a session, abandon the open-ended-question protocol because the person in front of them needs to be told something more directly. A social worker trained in cognitive behavioral therapy [CBT] will, fifteen minutes into a session structured around a thought record, set the worksheet aside and explore the client's attachment style because something the client just said about their mother suggested that the cognitive frame was missing what mattered. A peer specialist working from a manualized reentry curriculum will skip the next three modules on conflict management in the workplace because the participant has just disclosed that they are sleeping in their car and suffering from nightly panic attacks. These are instances of skilled practice.

> We're asking a system trained on text to learn a practice partly defined by its *departures* from text.

The record on this point is striking. When investigators coded session transcripts from rigorous evaluations of manualized CBT, the processes most correlated with improvement were not always the cognitive-restructuring techniques prescribed by the manuals, but also included ones more characteristic of psychodynamic work: attention to the relationship, exploration of recurring patterns, engagement with what the client was avoiding.[21] The "CBT" that worked was not quite the CBT the manual described. Other studies have found that the more a therapist departs from a manualized approach to fit the individual client, the better the outcomes.[22] Taken together, these are not examples of one modality smuggling itself into another. They are a high-water mark of skilled clinicians, regardless of orientation, reaching for whatever the moment requires.

This gives a sense of the design challenge for A.I. We are asking a system trained on text to learn a practice that is partly defined by its *departures* from text—to learn from a literature that systematically underrepresents what the most effective practitioners actually do, because what they actually do is the kind of thing that does not make it into the case

note, the manualized protocol, the published trial, or the supervision log. The clinical wisdom that matters most for outcomes is often *unwritten*, and unwritten material is the one kind of training data large language models cannot easily access. The honest critique is not that A.I. will fail to do what humans do. It is that the human professions have created a written record that obscures itself.

What this suggests for design is something different from another round of better training data, important as that is. It suggests that any A.I. system intended to operate in helping contexts will need to be designed around the recognition that the work itself rightfully resists standardization, that fidelity to protocol is not the same thing as competent practice, and that the practitioners worth learning from are the ones who can scrupulously articulate what they did when the protocol stopped serving the person in front of them. It also requires sustained engagement with people receiving services who can describe what actually helped them, and a commitment to the research methods that look at process rather than adherence to protocol.

Daunting, yes, but there is something genuinely hopeful in this picture. The incentives that keep human practitioners performing fidelity to models rather than owning their off-script work do not operate on A.I. A chatbot has no theoretical orientation to defend, no standing among peers to protect, no decades of mastery to abjure. If the field can build A.I. systems with access to what skilled practitioners *actually* do, those systems may be unusually positioned to learn the improvisational craft humans struggle to teach each other. This is not a claim that A.I. should or will replace human helpers. It is a recognition that A.I. might, with the right design and the right inputs, model back to us a more honest picture of what good helping actually looks like—including the parts our professional culture has trained us not to admit. Clinicians have something profound to gain from this, if we are willing to be taught.

Heal Ourselves

One common response to challenges like these is what the A.I. world calls "human-in-the-loop"—the requirement that meaningful human oversight be built into any A.I. system operating in high-stakes domains. That requirement is necessary—and insufficient. It also carries a trade-off. The people most likely to turn to A.I. are the ones for whom human help is hardest to reach, so a uniform human-oversight requirement risks pricing the technology out of exactly the settings where the access gap is widest. And even where human oversight *is* available and cost-effective, that oversight may not be able to recognize A.I.'s all-too-human missteps.

> This is a reckoning about humans as much as about A.I.

This points to something the standard fixes miss. Better training data and more human oversight both assume the knowledge already exists somewhere—waiting to be fed in, or to be checked against. But the argument of this piece is that the practices that matter most are rarely codified, and you cannot oversee your way to a skill the overseers themselves cannot commit to in writing.

No one solves these problems alone—not the helping professions, not the technology sector, and most of all not without the people on the receiving end of both human and machine helping, whose involvement in the A.I. design pipeline remains limited, and rarely rises to the level of real decision-making.[23] The work will require a willingness to make the investment of time, effort, and money in higher-quality care by humans *and* machines—and to do so on behalf of our most vulnerable, voiceless citizens, such as those ensnared in the criminal legal system.

This amounts to a reckoning about humans as much as it is about A.I. Trained on the written—and, as we know now, *incomplete*—record of how we as clinicians describe our own work, these systems reflect it back to us without the bowdlerizing we apply when we describe ourselves to ourselves. A system that mirrors us can faithfully amplify our worst reflexes, and left unexamined it surely will.

There is a one-paragraph story by the Argentine writer Jorge Luis Borges in which the cartographers of an empire grow so ambitious that they produce a map at one-to-one scale, matching the territory point for point—a parable about the hubris of believing a sufficiently detailed model captures the thing it models.[24] Clinicians have spent decades drawing ever more detailed maps of their own practice—manuals, protocols, diagnostic categories, fidelity measures—elaborate enough to feel, sometimes, like the territory. A.I. systems are trained on those maps. If we have mistaken the map for the territory, they will inherit our mistake and reproduce it at scale—an uncanny copy of something that never quite existed, a copy that fails to capture, because the territory may not lend itself to capture, the part that matters most.

The territory of human distress and the true prospect of making good care widely accessible, regardless of the medium, is larger and more dynamic than any map can hold. The task—not new, but newly urgent—is to move off the maps and start building on the land.

FOR MORE INFORMATION

Julian Adler: jadler@innovatingjustice.org

Acknowledgements

Julian is deeply grateful to the Center's Senior Media and Policy Advisor, Matt Watkins, for his ongoing editorial and thought partnership. Thanks as well to Samiha A. Meah and Michaiyla Carmichael for their impeccable document design, to Emma Dayton for her clutch editorial input, and a special thank you to Center Board Member Allen Waxman of DLA Piper for his encouragement to put pen to paper after hearing a brief version of the argument during the cocktail hour at the Center's recent thirtieth anniversary celebration.

[1] Forty-nine percent of the Los Angeles County jail system had a diagnosed mental health condition in the final quarter of 2025, per the Los Angeles County Sheriff's Department. The average daily population during that period was 13,200. "Just under 6,000" as a daily estimate is therefore a conservative approximation. See, Los Angeles County Sheriff's Department, *Custody Division Population Quarterly Report, October–December 2025*. https://lasd.org/wp-content/uploads/2026/03/Transparency_Custody_Division_Population_2025_Fourth_Quarter_Report.pdf. The broader framing of the three largest jail-based mental health providers (Los Angeles County, New York City's Rikers Island jail, and the Cook County jail) is widely documented; see Bryant, E. (2023, October 4). *The United States criminalizes people who need health care and housing*. Vera Institute of Justice. https://www.vera.org/news/the-united-states-criminalizes-people-who-need-health-care-and-housing; and Harvard Kennedy School Student Policy Review, "Jails: America's Biggest Mental Health Facilities" (2023). https://studentreview.hks.harvard.edu/jails-americas-biggest-mental-health-facilities/

[2] OpenAI reported that ChatGPT reached approximately 900 million weekly active users in late February 2026, up from 800 million announced at the company's DevDay event in October 2025. See, S. Altman/OpenAI as reported in "ChatGPT reaches 900M weekly active users," *TechCrunch*, February 27, 2026, https://techcrunch.com/2026/02/27/chatgpt-reaches-900m-weekly-active-users/; and coverage of Open DevDay 2025 (October 6, 2025), e.g., CNBC, https://www.cnbc.com/2025/10/06/open-ai-devday-live-updates-altman-jony-ive.html. The figure indicates the scale of the overall user base from which mental health use is drawn, not a measure of mental health use itself.

[3] Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations*. Advance online publication. https://doi.org/10.1037/pri0000292. The Rousmaniere finding is consistent with broader survey evidence on mass A.I. use for mental health support. See also McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., Stein, B. D., Mehrotra, A., Uscher-Pines, L., Cantor, J., & Yu, H. (2025). Use of generative A.I. for mental health advice among US adolescents and young adults. *JAMA Network Open*, 8(11), e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281 (finding that 13.1 percent of U.S. adolescents and young adults — approximately 5.4 million people — had used generative A.I. for mental health advice in early 2025); Pew Research Center. (2025, December 9). *Teens, social media and A.I. chatbots 2025*. https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/ (finding that 64 percent of U.S. teens use A.I. chatbots, with about three in ten using them daily).

[4] Bronson, J., & Berzofsky, M. (2017). Indicators of mental health problems reported by prisoners and jail inmates, 2011–12. U.S. Department of Justice, Bureau of Justice Statistics. https://www.bjs.gov/content/pub/pdf/imhprpji1112.pdf; Steadman, H. J., Osher, F. C., Robbins, P. C., Case, B., & Samuels, S. (2009). Prevalence of serious mental illness among jail inmates. Psychiatric Services, 60(6), 761–765; Substance Abuse and Mental Health Services Administration. (2012). Results from the 2011 national survey on drug use and health: mental health findings. https://www.hsdl.org/?view&did=730763.

[5] Anonymous participant in the A.I. & Restorative Justice Roundtable, New York City, April 29, 2026. The event was co-hosted by the Center for Justice Innovation, the A.I. and Justice Consortium, the Howard Law Artificial Intelligence Initiative, and the CUNY Institute for State & Local Governance.

[6] Kaiser Family Foundation. (2026, March). *Poll: 1 in 3 adults are turning to A.I. chatbots for health information, equaling the share who use social media for health*. KFF Health Information Trust. https://www.kff.org/health-information-trust/poll-1-in-3-adults-are-turning-to-ai-chatbots-for-health-information-equaling-the-share-who-use-social-media-for-health/. The poll found that among users under 30, 29 percent cited affordability and 38 percent cited access as major reasons for turning to A.I. for health information; among users with household incomes below $40,000, the respective figures were 32 percent and 25 percent. Among adults who asked A.I. about mental health, 58 percent did not follow up with a doctor or other health professional.

[7] Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4), 399–410. Open-access version: https://strathprints.strath.ac.uk/66200/

[8] Tyler, T. R., Goff, P. A., & MacCoun, R. J. (2015). The impact of psychological science on policing in the United States: Procedural justice, legitimacy, and effective law enforcement. Psychological Science in the Public Interest, 16(3), 75–109. https://doi.org/10.1177/1529100615617791.

[9] Lee, J. H. N., Chong, E. S. K., Chui, H., Lee, T., Luk, S., Tao, D., & Lee, N. W. T. (2023). A curvilinear association between therapists' use of discourse particles and therapist empathy in psychotherapy. *Journal of Counseling Psychology*, 70(5), 562–570. https://pubmed.ncbi.nlm.nih.gov/37439739/

[10] Gelso, C. J., & Hayes, J. A. (2002). The management of countertransference. In J. C. Norcross (Ed.), *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients* (pp. 267–283). Oxford University Press.

[11] Atzil-Slonim, D., Bar-Kalifa, E., Fisher, H., Lazarus, G., Hasson-Ohayon, I., Lutz, W., Rubel, J., & Rafaeli, E. (2019). Therapists' empathic accuracy toward their clients' emotions. *Journal of Consulting and Clinical Psychology*, 87(1), 33–45. https://pubmed.ncbi.nlm.nih.gov/30474991/

[12] Dierkhising, C. B., Ko, S. J., Woods-Jaeger, B., Briggs, E. C., Lee, R., & Pynoos, R. S. (2013). Trauma histories among justice-involved youth: Findings from the National Child Traumatic Stress Network. *European Journal of Psychotraumatology*, 4(1), 20274. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3714673/

[13] Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). 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 (FAccT '25)* (pp. 599–627). Association for Computing Machinery. https://arxiv.org/abs/2504.18412

[14] Finkelstein, J., & Rizvi, S. L. (2026, January 9). Therapy should be hard. That's why A.I. can't replace it. *Time*. https://time.com/7343213/ai-mental-health-therapy-risks/

[15] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. https://openreview.net/forum?id=tvhaxkMKAn

[16] Ben-Zion, Z. (2025). Why we need mandatory safeguards for emotionally responsive A.I. *Nature*, 643(8070), 9. https://doi.org/10.1038/d41586-025-02031-w. See also American Psychological Association. (2025). *Health advisory: Use of generative A.I. chatbots and wellness applications for mental health*. https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-chatbots-wellness-apps; American Medical Association. (2026). *A.I. chatbots and mental health: 4 ways Congress can boost safety*. https://www.ama-assn.org/practice-management/digital-health/ai-chatbots-and-mental-health-4-ways-congress-can-boost-safety

[17] Compton, M. T., & Shim, R. S. (2015). The social determinants of mental health. *Focus*, 13(4), 419–425. https://psychiatryonline.org/doi/10.1176/appi.focus.20150017

[18] Metzl, J. M., & Hansen, H. (2014). Structural competency: Theorizing a new medical engagement with stigma and inequality. *Social Science & Medicine*, 103, 126–133. https://doi.org/10.1016/j.socscimed.2013.06.032

[19] On how funding rules reward billable clinical services over the supportive and advocacy work that no one pays for, see Huber, B., Belenky, N., Watson, C., & Gibbons, B. (2023). *Reimbursement mechanisms and challenges in team-based behavioral health care*. Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. https://www.ncbi.nlm.nih.gov/books/NBK606628/, which documents fee-for-service incentives that favor volume of direct services, the difficulty of billing for anything outside direct care, and the frequent lack of reimbursement for services such as outreach, care coordination, and the work of peers and paraprofessionals.

[20] Brown, J. E. H., & Halpern, J. (2021). A.I. chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM-Mental Health*, 1, 100017. https://www.sciencedirect.com/science/article/pii/S2666560321000177

[21] Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *American Psychologist*, 65(2), 98–109. Open-access: https://www.apa.org/pubs/journals/releases/amp-65-2-98.pdf. For Shedler's broader

critique of manualization, see also Shedler, J. (2015). Where is the evidence for "evidence-based" therapy? *Journal of Psychological Therapies in Primary Care*, 4, 47–59. https://jonathanshedler.com/wp-content/uploads/2015/07/Shedler-2015-Where-is-the-evidence-for-evidence-based-therapy-R.pdf

[22] See for example: Owen, J., & Hilsenroth, M. J. (2014). Treatment adherence: The importance of therapist flexibility in relation to therapy outcomes. *Journal of Counseling Psychology*, 61(2), 280–288. https://pubmed.ncbi.nlm.nih.gov/24635591/

[23] A systematic review of how service users and people with lived experience are involved in designing digital mental health tools found that they take part in many roles but almost never as decision makers; see Veldmeijer, L., et al. (2023). The involvement of service users and people with lived experience in mental health care innovation through design: Systematic review. *JMIR Mental Health*, 10, e46590. https://doi.org/10.2196/46590. The specific involvement of justice-system-involved populations is notably less documented.

[24] Borges, J. L. (1998). On exactitude in science (A. Hurley, Trans.). In *Collected fictions* (p. 325). Penguin. (Original work published 1946.) Jean Baudrillard opens *Simulacra and Simulation* (S. F. Glaser, Trans., University of Michigan Press, 1994; original work published 1981) with the same Borges fable, using it to introduce what he calls the "precession of simulacra"—the idea that the map comes first and brings the territory into being, the copy standing in for an original that is absent or never existed. The reading here is narrower and does not follow Baudrillard all the way to his claim that the real has dissolved; the point is only that a model trained on the record of a practice can reproduce a copy whose most important original was never written down.