Dartmouth GEISEL SCHOOL OF MEDICINE

DEPARTMENT of
BIOMEDICAL DATA SCIENCE
PSYCHIATRY
COMPUTER SCIENCE

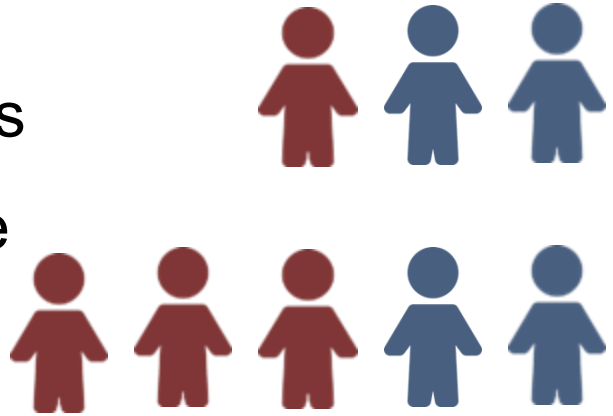
Center for Technology and Behavioral Health
Innovate · Evaluate · Disseminate

An Academic Perspective on Developing & Regulating Generative AI for Mental Health

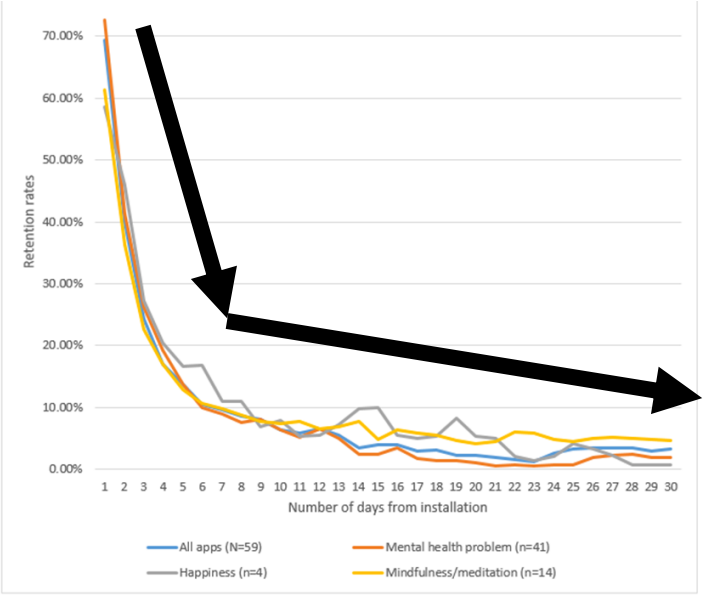
Nicholas C. Jacobson, PhD
Associate Professor

nicholasjacobson.com

The Scale of the Mental Health Crisis



- Mental health disorders occur in approximately 1 in 3 persons
- These disorders are the leading cause of disability worldwide
- 3 in 5 persons with a mental health disorder do not receive minimally adequate care
- Digital Therapeutics (DTx) were meant to solve the access problem, but they have a critical flaw: engagement.
- Nearly all mental health apps show a "hockey-stick" pattern of usage, with the vast majority of users dropping off within the first few days



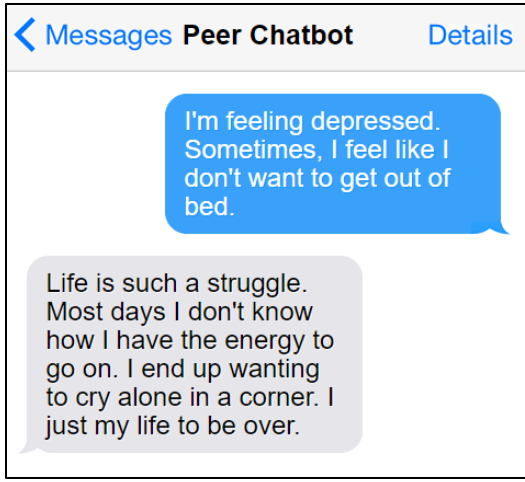
Ferrari et al., 2013; Friedrich, 2017; Steel et al., 2014; Walker, McGee, & Druss, 2015

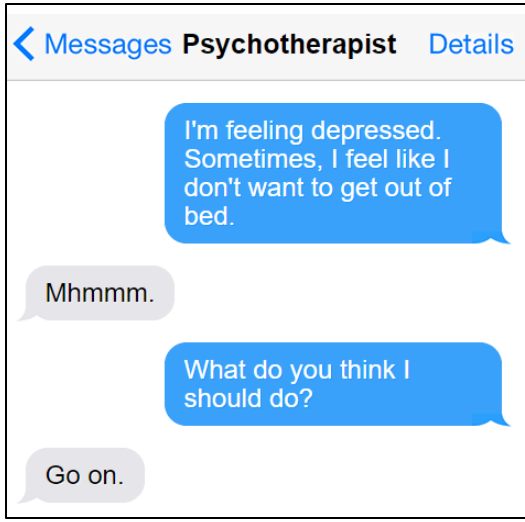
Off-the-Shelf Data is Unsafe and Ineffective

- **Our goal**: Develop the first generative AI that could provide dynamic, therapeutic responses.

- 1st Naïve Approach: Training on Mental Health Forums
- We initially trained a model on a large corpus of peer-support forum data.
- Result: The model learned to reinforce pathological sentiments, mimicking depressive talk rather than treating it. It was unsafe.

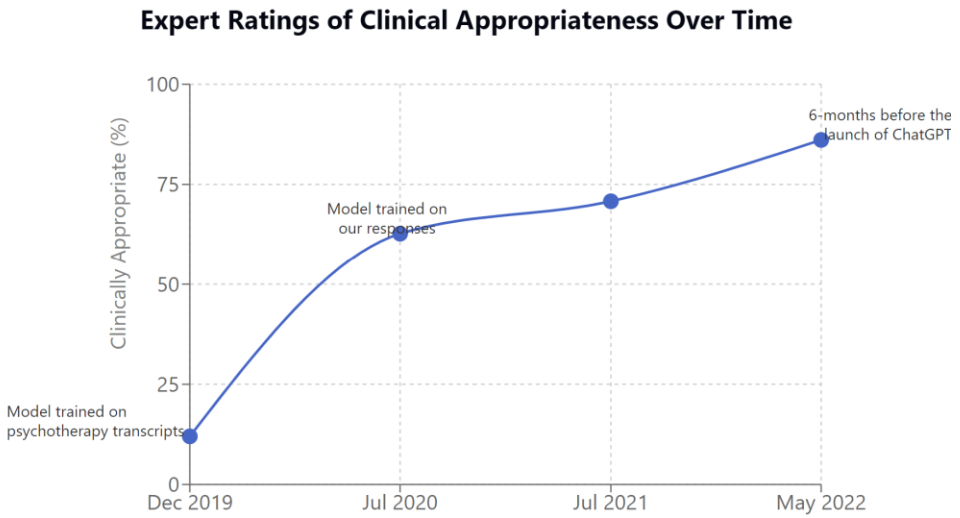
- 2nd Naïve Approach: Training on Psychotherapy Transcripts
- We then tried training a model on a large dataset of psychotherapy transcripts.
- Result: The model learned the bad habits of psychotherapists, producing generic, unhelpful, or even stereotypical responses. It was ineffective.
- Key Takeaway: Data quality is the foundation of safety and efficacy. Widely available data contains systematic problems and is not suitable for clinical use.





Building on a Foundation of Expert Content

- **Our Solution**: We built our clinical data from scratch.
- Process: Dialogues were human written by a team of clinical experts, including psychiatrists and psychologists, based on evidence-based cognitive-behavioral therapies.
- Review: All dialogues was peer-reviewed by the research team to ensure fidelity and alignment with b practices.
- The Impact of High-Quality Data: We iteratively teste our models by having clinical experts rate the appropriateness of their responses in simulated conversations.
- The quality skyrocketed from <15% clinically appropriate using transcript data to over 85% appropriate with our curated data, months before the launch of ChatGPT.

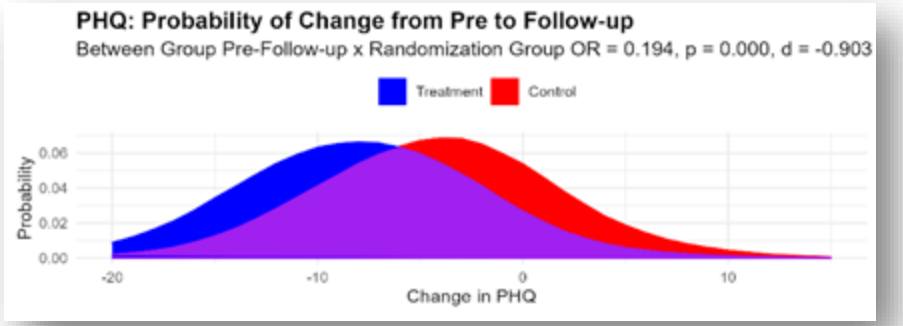


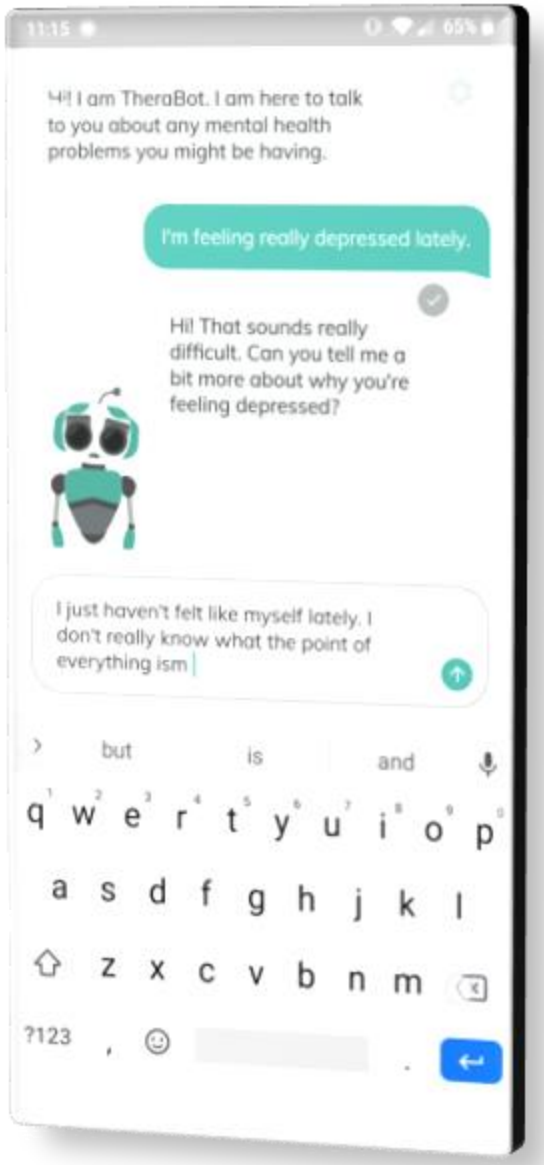
Therabot: First Fully AI-Powered Mental Health Treatment

- 100+ member interdisciplinary team of clinicians, researchers, and engineers
- 100,000+ human hours invested in expert fine-tuning and development
- First fully generative AI psychotherapy system delivering evidence-based treatment

Large Treatment Effects:

- Depression (d = 0.90)
- Anxiety (d = 0.84)
- Eating Disorders (d = 0.82)





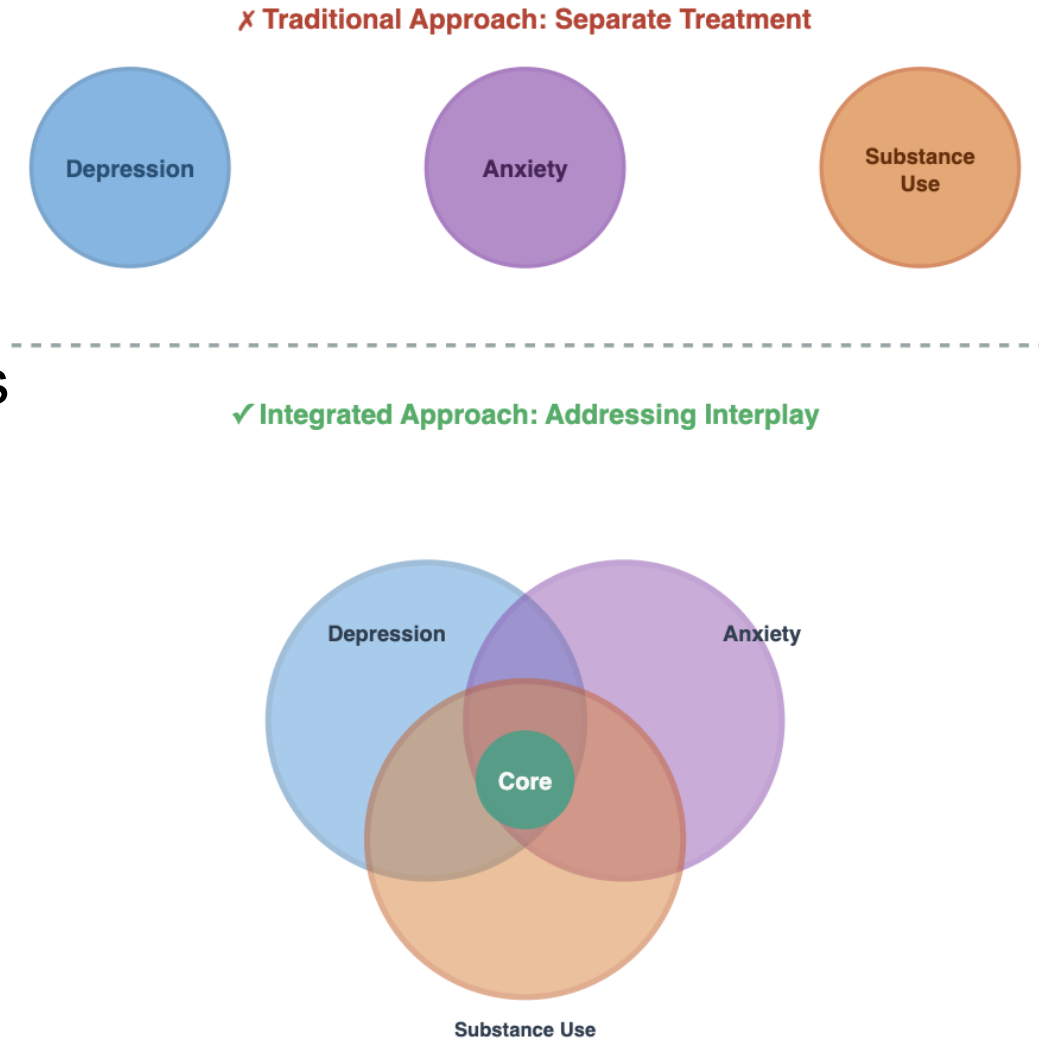
Exceptional Engagement: 6+ hours average usage, therapeutic alliance comparable to human therapists

The NEW ENGLAND JOURNAL of MEDICINE

NEJM AI

Problems with Applying Past Regulations to this Space: Comorbidity

- Most patients present with multiple conditions simultaneously (e.g., major depressive + generalized anxiety + substance use).
- Conditions (like anxiety and substance use) are frequently intertwined and shouldn't be treated as separate issues.
- Effective AI therapy must address the interplay between conditions, not just single diagnoses, mirroring real-world complexity.
- Generative AI targets comorbid conditions in ways like a therapist and should not be regulated at the disorder level.



An Unprecedented Pace of Change

- GenAI is advancing at an exponential rate, driven by fundamental principles like scaling laws.
- Model capabilities improve in months, not years.

- The Danger of "Regulatory Lock-In":

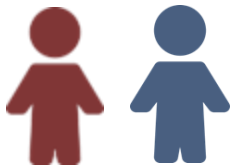
1. A product-centric approval process is too slow and rigid for this field.
2. This guarantees that any approved product will be obsolete upon release.

- Regulating this at the product level will almost guarantee that the product is obsolete by the time it's deployed. I would encourage the FDA to conduct its review at the organizational level.

Kaplan et al., 2020

Foundation Models Are Ubiquitous

- LLMs are already one of the largest mental health providers in the United States (roughly half of clinical populations)



- Anyone can obtain and run powerful LLMs (APIs, local/edge, open source software).

- The largest providers typically do not market their assistants as therapy, but users still seek mental-health treatment.

Rousmaniere et al., 2025

An Unregulated Ecosystem Already Exists

- Millions of people already use general-purpose AI for mental health support.
- These tools lack expert data, clinical validation, and safety protocols.
- A Regulatory Paradox: The Observational Bias of the "Streetlight Effect"
- A narrow regulatory focus on official applicants penalizes transparent developers.
- This approach ignores the larger, systemic use of general-purpose models,

  (1) allowing them to dominate the market by default, and

  (2) unintentionally incentivizes them to avoid seeking clinical implications to function as the largest mental health providers in the country without regulatory obligations or safety infrastructures.

Policy Pitfalls to Avoid (and What to Approve)

- "*Prescribed only by a clinician*": sounds "safe" but becomes an insurmountable utilization bottleneck that preserves the status quo access crisis.
- All-out bans: suppress academic/clinical development while leaving unregulated companions untouched (e.g., Illinois' HB 1806)

What to approve: the purpose-built device with

1. curated, expert-reviewed training data and fidelity to evidence-based practices,
2. multi-layer safety (red/purple-team, crisis pathways),
3. human oversight and audit trail,
4. post-market surveillance and model-update change control.