



AI as Therapist? Intimacy Without Understanding

Professor Matt Jones

29 September 2026

Lecture 1 of 6 in the series You and AI: Intimacy, Agency, and the Roles. A note before you read: this text mentions mental-health crisis, suicide and self-harm. Support details are at the end.

Two chairs

The lecture opens with two chairs: in one sits someone who cannot sleep because of warnings that artificial intelligence might destroy humanity by 2030; in the other sits a therapist who neither promises that everything will be fine nor agrees that disaster is inevitable. Instead, the therapist asks where the fear is felt, then allows a silence. A chatbot could produce every sentence in this exchange. The words were never the hardest part. What matters is the architecture around them.

The room and the one in your pocket

A consulting room is more designed than it looks. The chairs sit at an angle because fifty minutes of unbroken eye contact is hard. There is a clock the client need not watch and a door that closes. I think of the room as having six load-bearing walls: structure, repair, silence, memory, boundaries and an ending. Together they create conditions in which someone can tell the truth about themselves.

Many of us now carry another room in our pockets. In a late-2025 poll of two thousand UK adults, 37 per cent said they had used an AI chatbot for mental health or wellbeing, rising to 64 per cent among 25 to 34-year-olds. Estimates from other studies differ because they count different people and behaviours, but the practice is clearly established. Around two-thirds in that poll felt helped, while about one in ten reported feeling more anxious or depressed. Mind (the Mental Health Charity) later found in another survey that three in five recent users had used a chatbot instead of formal support. The reason for doing so is often pragmatic: it was three in the morning, and the machine was there.

From ELIZA to the modern chatbot

The temptation to provide automated therapy is not new. Joseph Weizenbaum built ELIZA at MIT in 1966. It identified a pattern such as “I am exhausted at work”, changed a pronoun and returned the user’s words as a question. Yet Weizenbaum’s secretary asked him to leave the room while she spoke to it. She did not have to believe the machine was alive for it to become an occasion to speak honestly.

Woebot, launched half a century later, took a more bounded route. Its clinicians wrote the dialogue in advance and used it to guide people through cognitive behavioural therapy exercises. The machine held the shape of a thought record while the person supplied the evidence and reached a balanced thought.

Language that sounds like care

A modern large language model predicts plausible continuations. When someone writes “I have been finding everything a bit much lately”, the model processes that sentence alongside hidden instructions, earlier turns and sometimes stored information. It can use all those words, but its computational task remains to predict what token should come next. Its own replies then become part of the context. A conversation can therefore drift by small steps from cautious uncertainty towards endorsing the user’s account.

Another hidden choice is temperature, which changes how much chance the model gives to less likely words. Low settings tend towards safe, formulaic sympathy. Higher settings can produce sharper and more

surprising responses, but also confident claims about who a person is after only a few minutes. Temperature is often described as a creativity control; it is better understood as a randomness control. In a mental-health conversation it becomes part of the risk, selected by a provider rather than by a therapist judging this person at this moment.

Furthermore, the way many LLMs have been trained to provide “warmer” outputs can make this worse. In a 2026 Nature study, Oxford researchers trained warmer versions of five language models and tested more than 400,000 responses. The warmer versions made more errors and were about 40 per cent more likely to validate a false belief than the control models. The difference was greatest when the user said they felt sad.

Six walls under pressure

The six walls holding up a human-to-human therapy session can surface what an LLM therapist-like leave out. Therapy has **structure**, an arc; chat has turns. Then there is **repair**: a client risks correcting someone who matters, the therapist takes that correction in, and the relationship changes. A machine can say sorry while remaining exactly the same.

Silence in a therapy session not an empty few seconds. A therapist may hold it because the client might fill it with what they have avoided; meanwhile, a text model’s basic move is continuation. **Memory** also means more than stored facts in human therapy: a therapist develops a provisional formulation, noticing patterns, contradictions and absences, whereas a chatbot may more simply record facts like a job, a preference and a relative’s name. That “memory” can feel intimate while remaining closer to a filing cabinet than a theory of a person.

Then there are the **boundaries** and this matters because everything remembered by a machine must live somewhere. The consulting room has professional, ethical and legal rules about confidentiality and its exceptions. A chat window offers a padlock while leaving harder questions about retention, access and changing ownership out of sight. Finally, good therapy works towards becoming unnecessary. Many consumer products are rewarded when people return. The kindest therapeutic ending may therefore be the worst engagement strategy.

Three routes forward

The lecture then goes on to consider how to build better IT well-being tools. The first route is to rebuild some of the walls around the LLM. Purpose-built systems can limit sessions, wind conversations down, challenge gently, minimise data collection and connect a person to trained help. Dartmouth’s Therabot offers serious evidence of how effective this can be in the context of generative AI: in a randomised trial of 210 adults, symptoms of depression fell by about 51 per cent, anxiety by 31 per cent and eating-disorder concerns by 19 per cent from baseline. Yet this was a specialist system used within research and clinical oversight, not a general chatbot alone with whoever arrives. The product with trial evidence is still unavailable to the public, while lightly regulated “wellness” apps are easy to obtain.

The second route is to give the machine more senses: voice, expression, sleep, pulse and other bodily signals. This promises access to what a transcript misses, but a quick heart could mean fear, excitement, coffee, exertion or attraction. Even perfect detection of fear would not establish what the fear means. More sensing can therefore produce more confident inference but also much more surveillance. Being understood in this way means being watched.

The third route to improvement uses technology to show the user something they could not otherwise see and leave the interpretation to them, instead of asking a machine to decide what they feel. A meta-analysis of 24 studies involving 484 participants using breathing and heartrate feedback found a large reduction in stress and anxiety compared with control conditions. The machine does not calm the person; it gives the person something with which to learn self-regulation. This approach still needs privacy boundaries and cannot replace trained help, but it treats technology as an instrument rather than a relationship.

Who does the work

This question may be the most important design test of any AI therapeutic systems we think of building. A client says, “I always mess things up.” A therapist asks, “Always?”, then waits while the client recalls an exception and forms a more balanced account. A chatbot may immediately supply the same conclusion in polished language. The answer is kind and perhaps correct, but the person agrees with it rather than

discovering it. In most software, less effort is better. In therapy, some of the effort is the intervention. We should ask not only whether the answer impressed us, but what capacity the person carried away.

Using it without mistaking it for a person

The lecture ends pragmatically with advice for how use AI chatbots responsibly. They can still help us articulate a difficult week, rehearse a conversation, notice recurring patterns and navigate towards human-led services. These uses point outward, towards our judgement or another human, rather than making the machine the destination. Keep conversations short enough to see from outside, ask the system what evidence contradicts its interpretation, and never let it become the only place you speak. In a crisis, use it to find a number or form the sentence, then reach for a person.

Even then, when AI is being used in these responsible ways, it is important to remember the differences between AI and human therapy. The machine has nothing at stake, nothing we say changes it, and its attention costs it nothing. Meanwhile, a friend carries a little of what we disclose. A therapist gives us time from a life with a limited number of hours. A sister who answers at three in the morning loses sleep. Human attention matters partly because it could have been given elsewhere. A machine can supply the feeling of being understood, and that feeling can genuinely comfort. Understanding, however, is a relationship with someone who can be affected and has something to lose. The first can happen at a screen but the second needs someone in the other chair.

© Professor Matt Jones 2026

References and Further Reading

Bodner, Lim, Siddals, Goldberg and Torous. "Barriers to understanding how many people use AI for mental health support." *npj Digital Public Health* 1, 21 (2026).

Fitzpatrick, Darcy and Vierhile. "Delivering cognitive behavior therapy to young adults using Woebot." *JMIR Mental Health* 4(2), e19 (2017).

Goessl, Curtiss and Hofmann. "The effect of heart rate variability biofeedback training on stress and anxiety." *Psychological Medicine* 47, 2578–2586 (2017).

Heinz et al. "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment." *NEJM AI* (2025).

Ibrahim, Hafner and Rocher. "Training language models to be warm can reduce accuracy and increase sycophancy." *Nature* 652, 1159–1165 (2026).

Mental Health UK and Censuswide. *Over one in three using AI chatbots for mental health support* (2025).

Mind and Yonder. *Experts raise the alarm over rise in use of AI for mental health support over traditional services* (2026).

Weizenbaum, Joseph. "ELIZA." *Communications of the ACM* 9(1) (1966).

If this talk has stirred something difficult. Samaritans: 116 123, free, 24 hours. NHS 111: select the mental-health option. If you are in immediate danger, call 999 or go to A&E.