[00:00:00]

John Moe: A note to our listeners, this episode contains mention of suicide.

Joseph Weizenbaum of MIT named his creation Eliza, after Eliza Doolittle from *My Fair Lady*. Human name, Eliza—even though Eliza wasn't a human at all. It was a natural language processing computer program. You could type things to Eliza, and it—she?—it would give language back to you through pattern matching and substitution. Eliza was one of the first chatterbots. We call them chatbots today. One of the most popular simulated conversations that Eliza ran was called Doctor, meant to sound like a psychotherapist, where the computer would take your questions and kind of turn them around on you. Now, people loved Eliza. Joseph Weizenbaum was pretty shocked actually by how much folks, including his own secretary, would insist that Eliza had real feelings and could really understand them. And Weizenbaum had to keep insisting that, no, it can't. It's just feeding information back to you.

But people really loved Eliza. They bonded with Eliza. They say that it helped them. Eliza debuted in 1966, 59 years ago. Now, a lot has happened—obviously—with computer technology since then. Computers and software are much more sophisticated, and so are bots. How sophisticated? Well, a program called Therabot was able to reduce symptoms of major depressive disorder by 51% in a clinical trial. It's still—like Eliza—definitely not a person. But dang, 51%.

It's *Depresh Mode*. I'm John Moe. I'm glad you're here.

Transition: Spirited acoustic guitar.

John Moe: Dr. Nicholas Jacobson is an Associate Professor of Biomedical Data, Science, Psychiatry, and Computer Science at Dartmouth. He directs the AI and Mental Health Innovation in Technology Guided Healthcare—or AIM HIGH—Lab at Dartmouth Center for Technology and Behavioral Health. And he's been building and developing Therabot for several years. He's the senior author of a new study of 106 people across the country on the effectiveness of this generative AI powered chatbot, the first ever clinical trial of this approach to therapy. And like I said, huge results. 51% reduction in depression symptoms, 31% for anxiety, 19% for eating disorders. And yet—and still—not a person.

Transition: Spirited acoustic guitar.

John Moe: Dr. Nick Jacobson, welcome to *Depresh Mode*.

Nicholas Jacobson: Thank you. Thanks for the invitation. Happy to be here.

John Moe: Let's start with defining this technology that we're talking about. What is Therabot?

Nicholas Jacobson: So, Therabot is a generative AI. So, many listeners may be familiar with ChatGPT. The idea of the underlying technology is somewhat similar in a big picture

perspective. Essentially, think about something like that, but instead of being designed to try to be it all, do it all; it's really designed specifically to deliver the highest quality treatment for mental health that you can imagine. I mean, that's essentially—that's the gist of what we're trying to do when we're designing Therabot.

So it's a—in terms of the interface, it's a smartphone application that folks install. And then it's a chat-based dialogue that folks can say anything they want in the same way that they could with any of these models. And from there, it's a conversation that flows. And the goal of Therabot is really that we've trained it to deliver evidence-based treatments, and try to essentially be both effective and safe.

John Moe: I want to get to the research that you've been doing and kind of find out a little bit how that all worked, but you used the word "treatment". It sounds like you don't think of this as just a sounding board or an exercise. This is designed to treat people with mental illnesses?

Nicholas Jacobson: That's correct, yeah. That is exactly what the system is intending to do.

John Moe: How can it do that when it's not a person? *(Chuckles.)* When it's just a computer program? How is it treating them?

Nicholas Jacobson: Yeah. So, I will say actually in trying to design a digital therapeutic, you don't need generative AI to design something that could be helpful for mental health symptoms. So, folks have been doing this within digital interventions for a long time.

[00:05:00]

You essentially could—a lot of the strategies that are evidence-based within mental health are behavioral and cognitive. And a lot of behavioral strategies are things that you can have standalone recommendations for an individual. That's been done for the past 30 years now at this point, where folks have done digital ways of trying to provide those behavioral strategies in kind of scalable ways. The thing that we're doing with Therabot specifically is trying to do this in a way that's much more like a human could provide. So, one of the downsides of prior approaches is then, digital interventions, they're kind of—the most personalized you'll get is modular, meaning you've got like some intervention module that is trying to treat your depression, and one for anxiety, and one for alcohol use.

But in therapy, you're experiencing things like—you will have—your social anxiety, for example, is totally related, potentially, to your alcohol use. And it's not two separate disorders that you have. So, a module on social anxiety and alcohol use disorder doesn't really work for somebody who's like drinking in social settings to try to avoid their anxiety, right? Like, these symptoms aren't experienced as unique buckets in real life. And so, generative AI provides a lot of the ability to have some of this level of ability to address comorbidities in real, personalized ways and contextual ways that's particularly meaningful in moving forward the field to really try to actually drive something that is <u>really</u> personalized and much more dynamic, much like a human does.

And a human does this all the time within therapy, right? They wouldn't think about these as two isolated issues that they would focus on one than the other. They'd say, “Of course, these are interrelated, and we need to address them together.”

John Moe: So, is it looking for language where an issue with depression and an issue with substance use overlap, and then it just generates a response based on what it knows when those things overlap? You know? Like, I’m just torn between whether it's thinking or whether it's regurgitating.

Nicholas Jacobson: So, yeah, I think at this point— I've been making the argument that models think for actually a long time based on the behaviors that we've seen with them and then working with that. I think it's very clear that models do their own version of thinking. I mean, it's certainly not the way that we think, but they do have internal processes that are going on that are— Folks will argue a lot of what they're doing is next token prediction, meaning like you're essentially trying to predict the next word when you're doing this kind of thing. But ultimately to be able to produce those responses, there's actually a lot of internal thinking that happens within the model. In terms of the examples that happen in the space surrounding a lot of this, we have a lot of evidence-based content surrounding treating these disorders individually, and also a lot of complex case representations that are part of our training data. And the training data is ultimately what makes Therabot work well.

So, it's the thing that has driven what— How the models behave is by example. So, they have to learn from something. And most of the foundation models—just as a counterpoint, so you can like hear what's kind of has typically like been done in other settings, is the internet. Everything that's available on the open internet. So, there's great content surrounding what evidence-based treatment looks like on the internet, but there's also really very terrible content surrounding what treat— If you say, “Act as a therapist,” you're getting all the fiction dialogues that are about what a therapist is. And you know, the same therapists that are having affairs with their clients. And you know, it's just like that could be part of that response generation in other settings.

Here, the data that we're training it on is saying, okay, here's what's actually been proven to work within randomized control trials and psychotherapy. And we're gonna take those same components, those same principles that have been shown to work, and we're gonna actually create vignettes that are—on the patient side, somebody—what they're experiencing and what the best system for—if we were actually gonna be live there, acting as Therabot in that moment, what would we want it to say? What do we want the system to say? And so, that's been our process is we have experts that are creating these dialogues based on evidence-based treatment. And then from there, before they even go into the training data, they have another expert that's actually reviewing the content before it makes it into the training corpus. So, we're really trying to make sure that these models are producing really high-quality, high-fidelity treatment.

John Moe: I was gonna ask where all the knowledge and language comes from. It's not combing the entire extant internet; it's based on a very limited source of information?

Nicholas Jacobson: Yeah. So, it's based on dialogues that we are inputting from the training team at its latest stance.

[00:10:00]

So, we are training—

John Moe: From humans, from humans *(inaudible; crosstalk)*.

Nicholas Jacobson: From humans. That humans have created. Yeah, a team of over 100 people that have spent over 100,000 human hours at this point, creating this. So, a lot of human effort from folks within the team.

John Moe: Is it therapy?

Nicholas Jacobson: I think it is. I think this is really generative psychotherapy, in terms of like what it looks like. If we were to take out a lot of the function in the form—like, the way that folks will describe the therapeutic process, we show that—in this trial—that folks exhibit a therapeutic bond with Therabot that's quite strong. For example, in ways that folks are essentially developing a working alliance, which is among the most well-studied construct in psychotherapy for how it works. And have shown that it, in this trial, actually, the therapeutic bond was about equivalent to what you would see in with humans. It's delivering evidence-based treatments through language in a form that is very similar to kind of how that works. We are very open about this being from a bot, no qualms about it. But I think it's pretty clear to me that this is actually therapy.

John Moe: Did you ever pause to consider that the name Therabot is kind of terrifying and sounds like it's from *The Matrix*?

Nicholas Jacobson: We have thought about the name a lot. I think one of the things that—the main thing that when we were trying to design the system that we wanted to go like completely all in on is that it's a bot. This isn't a human, ultimately. So, we named it Therabot, in part, trying to come up with and emphasize on what's gonna be the framing of what the purpose is that's gonna be— Also, making sure that folks don't anthropomorphize this system to the degree that they actually think that this is a human.

So, like in our design, for example, we have an animated robot. And we are really trying to make it very clear that this is ultimately not a human that they're interacting with.

John Moe: How was Therabot developed? How far back does this project go?

Nicholas Jacobson: We started in 2019, so this is like about three years before ChatGPT was released. A lot of folks don't know that generative AI existed back then, but it really did, and it was actually really promising as a technology at that stage. And I can tell you a little bit about like the failures we experienced in the early stages, thinking about what this could be. But that's—

John Moe: Yes, tell me about the failures.

Nicholas Jacobson: Yeah. So, before things started to go right, they went very, very wrong. And so, the early things that we did was thinking, within this space—so, these were with generative models that were existing at that time. We started training them on data that was collected on the internet. These are peer-to-peer forum data from folks that would be expressing, for example, things like information about their depression, and then receiving responses from peers in these areas during this process. So, we had readily available data on the internet in that way, thinking, "Okay, we just need lots of data to start with, and we will likely get some type of therapeutic response from this," in large part because there's actually good early literature surrounding that forums can really be promoting mental health outcomes.

So, the literature, for example, in cancer survivors who have access to these types of forums can find that their mental health outcomes are really much better than those that don't have it. And so, we started there. And our first model that we interacted with, you'd say, "I'm depressed. What should I do?"

This is the therapeutic response, but the very non-therapeutic therapeutic responses— The input, again, is, "I'm feeling depressed, what should I do?"

And the therapeutic response is, "I feel so depressed every day. I don't have the energy to get out of bed. I just want my life to be over."

This is like so not what we're going for, right? It's not only not providing a therapeutic response—if anything, it's escalating. And so, this was our first interaction with the robot is—I'm paraphrasing here, because I don't have the actual transcript in front of me, but it was like very, very similar to that. And it did not produce anything close to what we were going for.

So, the next shift we had was let's go where the experts are. Let's go with therapists. And so, we went to psychotherapy training videos. There's thousands of these. And so, these are videos that are psychotherapists who are learning how to do therapy. They would watch some of this content to see what they should do, how they should act. And what we found was the models at that point you'd say, again, "I'm depressed, what should I do?"

"Mm-hm," would be the response from the model.

John Moe: *(Chuckles.)* Not very helpful.

Nicholas Jacobson: "I'm depressed. No, I wanna know, what should I do? Go on."

[00:15:00]

But then usually three to five dialogue terms is "your problems stem from your relationship with your mother." And we hadn't even talked about the mother, to be clear. *(Chuckles.)* So, like this is like really not a success story, at that point. The thing that was super clear—and these are early learnings, but really important learnings—was that the models were emulating based on the data that they were seeing. The difference between the forum data and the psychotherapy data were vastly different from one another, in terms of how they behaved.

We just had the wrong data. And so, that's when we learned we needed to design the data ourselves, and that's when things started to go right.

So, about 1 in 10 of our responses that were produced from the psychotherapy training video—like, we did a lot of fidelity assessment during this time just to kind of see how good it was. About 1 in 10 was both personalized and acceptable. And about six months before the launch of ChatGPT, the evaluation at that stage, 9 in 10 of our models were really kind of personalized and this evidence-based content.

John Moe: When you talk about the therapy and the evidence-based treatment, you talk about the treatment. When someone is talking to the Therabot, at some point, does it give them a solution? Does it say, "Okay, I've heard you, and based on everything I know from what you've put in framed against all this other information that I have, here's what you should do"?

Nicholas Jacobson: Yes, it does that, but it does this in a way that's very different from how this might be done in other settings. So, one of the things that we learned early on in our development process was that if you are waiting to provide some type of intervention, it's completely unacceptable to users. So, like you're putting a lot into the system, right? It asks you all sorts of questions about who you are. But you're giving nothing back. Which is actually a lot of how psychologists might do assessment. They might have a three-hour or six-hour interview with you before they start to provide any treatment, trying to actually figure things out before they go into intervention.

And a lot of this is to try to get a problem list and differential diagnoses and things like this. We actually interact very differently in this way in that we're problem-focused. So, when folks are coming in and they have a problem, we can focus on trying to treat the problems that they're trying to address at that moment. Because of that—

John Moe: As opposed to the gestalt of the whole person. Just zeroing in on the issue?

Nicholas Jacobson: Well—yeah. Kind of like zeroing in on an issue early. And in a lot of ways, this provides a lot of traction for something that they can do a little bit more immediately, and they're not trying to wait for hours and hours with the system before they learn anything back. I think it's like really important within these settings, within some type of digital intervention, that you actually have something that really is actionable. So, yeah, that's a lot of how we've designed it.

Transition: Spirited acoustic guitar.

John Moe: We're talking with Dr. Nick Jacobson from Dartmouth, and we'll be back in just a moment.

(ADVERTISEMENT)

Transition: Gentle acoustic guitar.

John Moe: We are back talking with Dr. Nicholas Jacobson from Dartmouth about Therabot and the research that he's been doing with this.

[00:20:00]

And the results, the positive results that have been coming from that research. Let's talk about this research itself. What did the research set out to find? What was the objective?

Nicholas Jacobson: So, we wanted to conduct the first ever randomized control trial of a fully generative psychotherapy and really see, essentially, if it's both safe and effective. This has not been done in other settings. So, this is really something where we're really trying to use generative AI to actually provide psychotherapy.

We've been designing the system for years and years at this point, and then wanted to actually then test how well the system actually works in producing real outcomes in folks that are using it. So, the way we designed the trial was the primary targets of populations that we were recruiting are folks that are experiencing clinical depression, anxiety, or folks that are clinically at risk for eating disorders. And so, we enrolled them in the trial. We randomized them to either receive the ability to interact with Therabot or to be on a waitlist control. Which doesn't mean that they can't do anything. They could actually go and receive like in-person treatment or medication treatment if they could get access to it in that window. Or they could do nothing and kind of wait until they got access to it.

At the same time, then we looked at how well the folks that had access to Therabot versus those that didn't—their level of symptoms at four weeks and at eight weeks. So, we're looking at the change between when they came in and where they ended up at, you know, the first month and then the second month, essentially. So, the primary things that we found within the trial were that there was pretty large symptom reduction in depression and anxiety and eating disorders. And one of the things when we're thinking about how the entire field of psychiatry and psychotherapy are evaluated, when you're trying to see how well something works, they will quantify how well this works through an effect size, which is like the degree of change you experience in something versus another condition. And our effect sizes were really large. The same type of effect sizes and to the same degree of something that is really comparable to what we would see with our gold standard, best case in-person treatment delivered with the highest fidelity across a longer span of time—usually 16 weeks for example—with folks with psychotherapy within these outpatient settings.

So, this was a <u>really</u> strong initial result for what things look like in terms of the effectiveness of the work. The other things that we evaluated within the trial were the bond between Therabot and the user. And so, this is like the most well-studied construct in how psychotherapy works is the ability to get together and work on a common goal. So, essentially for the user to trust the therapist and come together. And actually, that bond is part of what drives and facilitates that.

John Moe: How can there be a bond with a bot? *(Chuckles.)* With a bunch of data? That's—I would argue, respectfully, that you're wrong. That it's impossible to bond with a robot.

Nicholas Jacobson: Oh, I think we've got evidence at this point that are pretty clear that that happens. And the folks are—in terms of how that happens, it's really— And this was surprising to me, the extent that it happens, but the things that people engaged with within the trial included behaviors that were related to the things that we do with other humans. So, things like checking in on someone. So, they would actually check in with Therabot and say, "Hey, Therabot, just checking in." A lot of people would nickname Therabot Thera. There was a lot of folks that just developed, clearly, like feelings within the relationships of that were kind of pro-social, like related to like these feelings of intimacy when we were reviewing how this was going during the course of the trial.

So, folks were developing that— I will say that folks have measured this outside of like Therabot and digital tools. And there is actually this prior literature on the digital working alliance, which is essentially the degree that people even develop a fondness for the tools that they're getting. Like, kind of almost—I think in *(inaudible)*, it's a little bit less— It doesn't feel like the same. So, like people, you know, it's kind of like you develop an affinity for your hammer or something like that, where you have a— And that's—I think of prior versions of the working alliance, when they think about that, it's like, "Oh, this tool is reliable," or something like this. I think, in this setting, it really is pretty different—

[00:25:00]

—in that people actually treat it as something that—a system that they can ultimately trust and that they have ability to put some stock in that they can work together effectively. So, I don't think you need to be a human to have the ability to have like that trust and that bond. And clearly, it seems to not be the case, because the norms that we got when we asked folks to rate this was like nearly identical to what you'd see with humans. It's like not only can they develop it, but it seems to be about as strong as what you'd see with a real human provider.

John Moe: So, they anthropomorphized it. They turned a non-living thing into a living thing in their minds.

Nicholas Jacobson: You could absolutely say that. I would say there's certainly no doubt that people anthropomorphized it, but we develop bonds with other things that are non-humans. I don't think that you need to be a human to develop a bond. Right? So, like I don't think that necessarily has to be an anthropomorphizing. I think that maybe the system acts in a way that is intelligent enough for us to develop a bond, I suppose.

John Moe: So, you're at Dartmouth. Were most of the subjects students, like young adults?

Nicholas Jacobson: No, we recruited across the United States, and we recruited online. The age range of the subjects did go from 18, but up to upper 60s within the trial, with the average age well above like kind of young adult territory.

John Moe: Okay. Was there a generational difference in how well people responded or how well they bonded to this bot?

Nicholas Jacobson: We don't have data on that yet. We haven't looked at the moderators, which are kind of like differences in how effective it was by age or things like that just yet. I will say like I have not noticed generational gaps when I was—during the course of the trial. I would say I didn't get a whole lot of sleep, because the first half of it I was reading near every message in near real-time. And so, this is something that I just wanted to do for my own feeling of like we needed to have oversight within the system. It's really new research. And making sure that we are doing this responsibly, with the option that we would intervene with something that would go wrong. But I think because of that, this is I guess just really something that was quite an intensive process.

John Moe: In the article I read coming out of Dartmouth about this, you're quoted as saying there's no replacement for in-person care, but there are nowhere near enough providers to go around. Which is true. We're in a huge shortage of providers. It sounds to me like you're talking about this as something that sort of is a replacement for in-person care. So, what's the distinction? Is it or is it not a replacement?

Nicholas Jacobson: So, I think when folks think about replacements, they think about replacing jobs. A lot of, for example, psychotherapists would worry about being replaced by something like this type of technology. And I think that folks will continue to benefit from human care, with or without very strong AI. So, like looking 10 years into the future or something like that, if this really continues to go as well as it could, humans will always be part of the care system, and a large part of it. I don't think the unemployment rate would change in any way. Of which, for example, because of the shortage, there is less than 1% unemployment rate for psychologists through recessions. Because of this, this is like—we're in an imbalance between how many people can have access to care and really how many people can receive it.

So, like I think some people will go to the idea of this feeling threatening to replacing humans. No, we wanna provide something that is a strong, evidence-based treatment for folks that generally don't have it. And so, the goal is really to provide something that's really a good treatment for folks that don't have anything. Do I think that it can be as effective, ultimately, as human care? I think we're probably already at a stage that is likely the case. But it doesn't mean that it's gonna benefit everyone. A lot of folks probably won't benefit from this and would benefit from a human provider. So, I don't think this is like a one-size-fits-all tool. It's not gonna be benefit from everybody, and humans are always gonna be part of the care system. But I do think that this could be really useful for a lot of people that don't have anything.

John Moe: Do you define Therabot and what it does as care?

Nicholas Jacobson: I think that I would frame it under that term. Yeah.

John Moe: Even though it's not coming from a person.

(Nicholas confirms.)

The numbers that came back, like 51% reduction in depression symptoms from people who had access to this—you know, it wasn't a massive, thousands of people study, but it was a study. Did those numbers surprise you?

[00:30:00]

Nicholas Jacobson: The size of the reductions were really pretty large, and these folks were also like—as a context point—not like folks that were just barely depressed. Like, moderate to severely depressed was about our average for the depression sample, like kind of decently in the thick of things. Because we've been working on these systems for so long, that part wasn't the surprising part. The degree to which folks developed the bond and the level of engagement throughout the trial, like on average interacting with it for over six hours— For a digital therapeutic, that's—in my mind—completely unheard of. I've never experienced this in prior work that we've done in this space.

And so, those were the things that were surprising to me in the course of the trial. But the effectiveness, actually, I was really expecting that to pan out.

John Moe: Now, I went online, and I googled Therabot, and there is an online version that you can use.

Nicholas Jacobson: That's not us.

John Moe: That's not yours? That's not you?

Nicholas Jacobson: No, that's not us. And it's—yeah, there are actually a few different folks that are using the same name. So, yeah, it's quite confusing, I think. We are not providing any of access to this at this point. Really, I think—frankly—it would not be judicious to release this technology at this point. There is a lot of folks that are entering this space and openly trying to provide access to treatment. And to be frank, we've been doing this for, at this point, close to six years. And I don't feel comfortable releasing this system, not because I think I'm like ridiculously conservative. The technology is brand new, and we don't know a whole lot about its safety and effectiveness. I think we need heavy oversight of how these tools actually behave. But folks are really marketing this technology in this way with things that are developed without this like really strong attention to detailed process and making sure that this is safe and effective.

A lot of this is essentially saying, “Hey, here's a large foundation model. We're gonna change the prompt, saying ‘act as a therapist,’” and off to the races. They essentially then create an interface, and that's it. That produces some behavior that is really, really risky. Those scenarios can go okay. Sometimes it provides okay care. Sometimes—a lot of times it's not delivering tools that are known to be effective. And sometimes, at a pretty high rates, these foundation models that are not trained in this way will be actively harmful. So, I think that the field really—in terms of the regulation, I think there needs to be a lot more in place to try to make sure that this is actually not harming folks. Because I expect that it is right now. There's actually pretty good evidence that it is, both from kind of prior work that we've done, but also some folks that have done this in kind of public settings that have had different

settings that have acted as a therapist that have contributed to actual real suicides that have happened.

John Moe: Wow. No, I tried this one out, and it said to me, “I genuinely care.”

And I thought, “No, you don't!”

(They chuckle.)

You’re incapable of emotions, because you are data!

What about your Therabot then? And maybe it's time to reconsider a name if there's all these other imitators out there. Do you see this as something that you are eventually going to bring out to the rest of the world?

Nicholas Jacobson: So, ultimately with continued data on its safety and effectiveness, we hope to strategically and openly begin to have access to it. We think there needs to be more trials before that happens. So, I want a larger trial in part because we had some data now on the safety and effectiveness, but the trial was—like, it was a decent size for initial work, but this is not like a convincing trial that I'd want on an appropriate oversighted release of this technology. So, we want a larger trial before we go and do this that's a bit more definitive, and there thereby start to actually scale it. So, that's a little of how we're replanning to handle it.

Transition: Spirited acoustic guitar.

John Moe: We have some questions from some of our listeners, and we're gonna get to those right after the break. We're talking with Dr. Nicholas Jacobson.

Promo:

> **Music:** Exciting, rhythmic synth.
>
> **John-Luke Roberts:** If you like too many podcasts, you'll love *Sound Heap with John-Luke Roberts*. It's got clips from all your favorite podcasts, such as *Diary of a Tiny CEO—*
>
> **Tiny CEO:** Leonard Sprague, tell me how you make your money!
>
> **Leonard Sprague:** I go to the beach, and I steal people's towels.
>
> **John-Luke Roberts:** *Rememberama*.

Speaker: Remember the trend of everyone whacking themselves in their head with hammers and mallets when they wanted to lose weight?

John-Luke Roberts: And *Elty Jom's Lobly Sonds*.

[00:35:00]

Elty Jom: I'm here today with Kiki Dee! Hello, Kiki Dee!

Kiki Dee: Hello, Eltonnn.

John-Luke Roberts: There's dozens of episodes to catch up on and brand-new episodes going out right now! So, if you want far, far, far too many podcasts, then look for *Sound Heap* on Maximum Fun. Boop-boop!

(Music ends.)

Promo:

Ella Hubber: Alright, we're over 70 episodes into our show, *Let's Learn Everything*. So, let's do a quick progress check. Have we learned about quantum physics?

Tom Lum: Yes, episode 59.

(Pencil scratching.)

Ella: We haven't learned about the history of gossip yet, have we?

Caroline Roper: Yes, we have! Same episode, actually.

Ella: Have we talked to Tom Scott about his love of roller coasters?

Caroline & Tom: *(In unison.)* Episode 64.

Ella: So, how close are we to learning everything?

Caroline: Bad news. We still haven't learned everything yet.

Ella: Awww!

Tom: WE'RE RUINED!

Music: Playful synth fades in.

Ella: No, no, no! It's good news as well. There is still a lot to learn!

(They cheer.)

I'm Dr. Ella Hubber.

Tom: I'm Regular Tom Lum.

Caroline: I'm Caroline Roper, and on *Let's Learn Everything*, we learn about science and a bit of everything else, too.

Ella: And although we haven't learned everything yet, I've got a pretty good feeling about this next episode.

Tom: Join us every other Thursday on Maximum Fun.

(Music ends.)

Transition: Gentle acoustic guitar.

John Moe: We are back talking with Dr. Nicholas Jacobson from Dartmouth about Therabot and the important distinction that, if you Google Therabot, that's not the thing that he's made; that's not the thing that we're talking about.

Nicholas Jacobson: No, that's not us.

John Moe: That's not you. That's a different Therabot. The article I read about this emphasizes that it all still needs clinical oversight. And one of our listeners read that and says, "Well, if it needs clinical oversight, how is this solving anything?"

Nicholas Jacobson: Good question. So, I think that it needs clinical oversight at this stage. This trial, for example, monitored every message that was sent to or from Therabot in near real-time. This is in part because this is the first time this has been done, right? So, like when we're thinking about this, we don't want to go and release something out into the open wild as a starting point. That doesn't—the starting point and the ending point don't need to be the same. So, for example, we have a lot of risk detection modules surrounding, you know, crisis events or things like that that might be happening.

And so, the scaled-down version of this is using those models to detect when we need some type of human review, as opposed to trying to review every single message in near real-time. It's like, no, we've got flags of, okay, a human could take a look at some of these messages rather than every message. And so, as we get better data on the safety and oversight, I think

the level of oversight and future trials will be scaled down. We have evidence that this really didn't produce some kind of huge negative outcome in this way, and we'll have—essentially, kind of titrate the level of oversight with the evidence as it increases. So, I think it becomes more scalable in its later forms, but we need to handle this really carefully at its origin.

John Moe: One of our listeners wrote in: "A lot of the AI powered chatbots out there tend to be rather sycophantic with their users. I'm curious how a therapy chatbot can challenge the beliefs or stories that a user might be telling. One of the most valuable aspects I've seen in therapy with human therapists is their ability to call me on my BS. I feel like this would be a critical area to address with an AI driven chatbot."

Nicholas Jacobson: I fully agree. A lot of current foundation models and things that are skins on foundation models have essentially just pure validation almost type of styles of responses. So, they continue to ask you to talk about yourself, and there is a form of therapy that entirely exists surrounding approaches like this. So, this unconditional positive regard is essentially a lot of what is done in relational style therapy that goes back for—I don't know—75 years or so at this point. This is—so, it is like something that is often done in treatment as a component of treatment at this point, but it's definitely not the most evidence-based techniques that we could deliver at this time.

So, yeah, we are actively trying to engage with cognitive distortions, for example. So, ways that folks kind of view the world that can contribute to their psychopathology, instead of saying, "Oh, we're gonna let those go and left untouched." How we do it isn't, necessarily—I would say—confrontational, like directly confrontational. A lot of how the best psychotherapists do it won't say, "Hey, you're wrong, and here's why." It's trying to get folks to actively challenge their own thoughts and really lead folks to do that. So, we do that in a similar style to how it's often been done within cognitive therapy, for example, which is Socratic questioning. So, getting folks to answer questions surrounding things as opposed to saying, "Here's the answer that I know to be true."

[00:40:00]

And a lot of that I think is not only because I think that's a way to do it, but I think it's one of the ways that you can do it and get folks to actually see it for themselves and like view it through their own lens and their own eyes. And those moments of insight I think don't happen necessarily as clearly as often without some kind of defensive response when it's coming from them. Like, they're ultimately the ones that connects the dots. You may hold up the two dots, but they're the ones that draw the line between it. And so, we try to do that.

John Moe: Do we know how— I mean, 'cause anyone who's sent an email or a text knows that tone is really difficult to capture in what you're saying. Like, if I go to a therapist, I can say, *(dryly)* "Oh, I'm doing <u>great</u>," and it's, you know, obvious that I'm being sarcastic. Do we know how your Therabot matches up against a human text-based therapy session?

Nicholas Jacobson: We don't have direct evidence like that with face-to-face care, for example. In the realm of sarcasm, I will say that's one of the things that has changed pretty dramatically over the course of the development of generative AI. Five years ago they were <u>incredibly</u> bad at understanding sarcasm. But that has gotten pretty dramatically better over

the course of the development here. In large part, they can more often contextually see it within the responses.

So, part of that has actually made them a little bit better at humor. I think humor is still a weakness area for them, for large language models generally. But part of what's made them better is that they've gotten better at things that are kind of reading between the lines, sarcasm and things like that.

John Moe: Has Therabot performance—this is a question from a listener. “Has Therabot performance been compared to older, more simple, and therefore more energy-efficient chatbots?” Like, I mean, the Eliza program goes back to the 1960s.

Nicholas Jacobson: 1960s! Yeah, absolutely. We haven't yet. You know, in terms of the iterations of really what we want to compare against, you know, there's a broad variety of things that could be done differently. We wanted to start somewhere, and one of the biggest things that we wanted to get a pure stance on how well this worked is compare it against a waitlist control—just to like not have some kind of active condition. So, we know the magnitude of effect.

Now, can we compare it to other things? Yeah. I mean, one of the—we actively do wanna have an active control. Actually, one of the next active controls we wanna do is directly against humans. The other side of this—like, so things that we expect to be less effective, Eliza being an example—I think is another good thing to do in the future. We've done the first thing, but not the last thing.

John Moe: Nick, when you use the phrase “against humans”, I get real nervous. I don't know what you're—*(laughs)*.

Nicholas Jacobson: Certainly, so what I mean is essentially have a randomized control trial where instead of saying, “Here's Therabot for some people and here’s a waitlist control,”—meaning “We'll see you in eight weeks, and then we'll give you access to the system”—we say, “Oh, here's a provider.: And the provider will actively see them. So, it's essentially giving folks—either randomizing them to be able to access care, or be able to access Therabot is one of the things that we wanna do as the next step.

John Moe: Do you think of Therabot as a person when you think of it?

Nicholas Jacobson: No, I view it as a model, ultimately, is like kind of how I think about it. Now, I will say—

John Moe: Because it sounds like some of the users think of it as a person.

Nicholas Jacobson: I don't know that they would necessarily describe it as a person. Like, I think that it has things that are like—*(flustered sounds)* there are aspects that have things that have been reserved in the past for things that are maybe in the space of a person for like an archetype of what a person is that in the past maybe have been segmented, because we haven't had the ability to like really interact with something that is as intelligent. But I think

everybody still knows that this is a bot, right? So, it's like I don't think there's any ambiguity that this isn't a human. But there are aspects that are human-like. So, there are certainly other aspects of interacting with something that has ultimately got some intelligence to it that have some kind of embodiments of human qualities. I would say that.

John Moe: What does Therabot 2.0 look like? What is the next version of this? Where is the technology going from here? Because I can't imagine you're done.

Nicholas Jacobson: No, we're not done. A lot of what we've done is to try to train it on every type of commonly experienced comorbidity. But we wanna have trials on how well it works for all other sorts of things.

[00:45:00]

Both because every time we do that type of trial and make sure it's fully adapted to another area that it's trying to actually target, it increases the ability of the system to work for every setting. So, in that way, like thinking about all of the different ways that folks access psychotherapy, I think is ultimately some of the ways that this type of technology could be helpful. Yeah. We want to continue to refine it, continue to do oversight and do work with it, but we've got other adaptations—for example—of different populations.

For example, one of the adaptations that's happening right now is related to folks that have cannabis use disorder. So, kind of reliance on cannabis, and real problems because of their use patterns within daily life, and could use some benefit from in intervention there. And another trial that I'm trying to adapt is to folks that are young adults, and I'm trying to really improve—kind of increase the customization to that group.

John Moe: One of our listeners says, "I'm fairly convinced that most big AI like this is an environmental nightmare with a veil of impartiality masking the huge biases injected in their training." So, this person is not on board. But what about that environmental aspect of it? How much energy does something like this use? 'Cause we hear some scary stories.

Nicholas Jacobson: Yeah. I think it's certainly—in terms of the energy to do all of this—as like an industry and a sector, the impact is huge. So, I do think that the work needs to be done in ways that really ultimately reduce the impact and burden. The field has gotten much better at being more efficient over time. So, kind of having greater potential at greater efficiency. The problem with that is, every time we have a jump in efficiency, we have greater expectations, and we want to do more. So, that's the tradeoff of kind of what that flip side looks like.

But ultimately, a lot of this, I think that it does have a real environmental impact. And so, I think that the training ultimately, hopefully, continues to be done in ways that are more sustainable over time, in terms of environmentally friendly. And so, there are industry sectors, for example, that are trying to do this in larger settings for the foundation models that are talking about, you know, reopening nuclear power plants and trying to get cleaner forms of energy.

Music: “Building Wings” by Rhett Miller, an up-tempo acoustic guitar song. The music continues quietly under the dialogue.

John Moe: Dr. Nick Jacobson, thank you so much for your time.

Nicholas Jacobson: Absolutely. It was a pleasure, and I really enjoyed the conversation. And thank you for the opportunity to talk to you.

John Moe: That's Dr. Nicholas Jacobson, Associate Professor of Biomedical Data Science, Psychiatry, and Computer Science at Dartmouth; Director of the AI and Mental Health Innovation and Technology Guided Healthcare—or AIM HIGH—Lab at Dartmouth Center for Technology and Behavioral Health.

Our show exists because of the donations of people like you—people listening to the show, people who think that this is providing a good service and helping other folks in the world. We appreciate your donations. We really need them in order to keep going. Please go to MaximumFun.org/join. You can donate; you can join at the $5 a month level on up. We really appreciate it. If you've already donated, thank you. You are making a difference in the world. Be sure to hit subscribe. Give us five stars, write rave reviews. That helps get the show out in the world, also where it can help folks.

Speaking of helping folks, the 988 Suicide and Crisis Lifeline can be reached in the US and Canada by calling or texting 988. It's free. It's available 24/7.

We're on BlueSky at @DepreshMode. Our Instagram is at @DepreshPod. Our newsletter is on Substack. Search that up. I'm on BlueSky and Instagram at @JohnMoe. Join our Preshies group on Facebook. A lot of good conversation happening over there. People talking about mental health, people talking about the show. I hang out there. I'll see you there. Our electric mail address is DepreshMode@MaximumFun.org.

Hi, credits listeners. That old Eliza therapist program still works, by the way. Someone found the code and put it up online. We have a link on our show page if you are so inclined.

Depresh Mode is made possible by your contributions. Our production team includes Ragu Manavalan, Kevin Ferguson, and me. We get booking help from Mara Davis. Rhett Miller wrote and performed our theme song, “Building Wings”. *Depresh Mode* is a production of Maximum Fun and Poputchik. I'm John Moe. Bye now.

Music: “Building Wings” by Rhett Miller.

I’m always falling off of cliffs, now

Building wings on the way down

I am figuring things out

Building wings, building wings, building wings

No one knows the reason

Maybe there's no reason

I just keep believing

No one knows the answer

Maybe there's no answer

I just keep on dancing

[00:50:00]

Jeri: This is Jeri from St. Paul, Minnesota. My message is: be kind in your mind, especially when you're feeling down.

(Music fades out.)

Transition: Cheerful ukulele chord.

Speaker 1: Maximum Fun.

Speaker 2: A worker-owned network.

Speaker 3: Of artist owned shows.

Speaker 4: Supported—

Speaker 5: —directly—

Speaker 6: —by you!