IS THE WORLD READY FOR AI-POWERED THERAPY?

Today's mental-health apps are the result of a seven-decade search to automate mental-health therapy. Now, large language models such as GPT-3 pose fresh ethical questions. **By Ian Graber-Stiehl**

ILLUSTRATION BY FABIO BUONOCORE

Since 2015, Koko, a mobile mental-health app, has tried to provide crowdsourced support for people in need. Text the app to say that you're feeling guilty about a work issue, and an empathetic response will come through in a few minutes – clumsy perhaps, but unmistakably human – to suggest some positive coping strategies.

The app might also invite you to respond to another person's plight while you wait. To help with this task, an assistant called Kokobot can suggest some basic starters, such as "I've been there".

But last October, some Koko app users were given the option to receive much-more-complete suggestions from Kokobot. These suggestions were preceded by a disclaimer, says Koko co-founder Rob Morris, who is based in Monterey, California: "I'm just a robot, but here's an idea of how I might respond." Users were able to edit or tailor the response in any way they felt was appropriate before they sent it.

What they didn't know at the time was that the replies were written by GPT-3, the powerful artificial-intelligence (AI) tool that can process and produce natural text, thanks to a massive written-word training set. When Morris eventually tweeted about the experiment, he was surprised by the criticism he received. "I had no idea I would create such a fervour of discussion," he says.

People have been trying to automate mental-health therapy for 70 years, and chatbots in one form or another have been a part of that quest for about 60. There is a need for the greater efficiency that these tools promise. Estimates suggest that for every 100,000 people worldwide, there are about 4 psychiatrists on average; that number is much lower in most low- and middle-income countries.

Recognizing this gap, smartphone-app developers have built thousands of programs offering some semblance of therapy that can fit in one's pocket. There were 10,000–20,000 mobile mental-health apps available in 2021, according to one estimate (see go.nature.com/3keu6cj). But for many of these apps, the evidence to support their use is quite thin, says Nicholas Jacobson, a biomedical data scientist at Dartmouth College's Center for Technology and Behavioral Health in Lebanon, New Hampshire. And the incorporation of large language models such as GPT-3, and the related chatbot ChatGPT, represents a new step that many find concerning.

Some are worried about increased threats to privacy and transparency, or about the flattening of therapeutic strategies to those that can be digitized easily. And there are concerns about safety and legal liability. Earlier this year, a Belgian man reportedly committed suicide after weeks of sharing his climate-related anxieties with an AI chatbot called Eliza, developed by Chai Research in Palo Alto, California. His wife contends that he would still be alive if he had not engaged with this technology. Chai Research did not respond to a request for comment.

Hannah Zeavin, a scholar of the history of human sciences at Indiana University in Bloomington, warns that mental-health care is in a fragile state. That makes it both an attractive and a vulnerable target for an industry that famously "likes to move fast and break things", she says. And although this technology has been building for decades, the swelling interest in emerging AI tools could supercharge its growth.

Scale in automation

The Eliza chatbot took its name from an early natural-language-processing program created by computer scientist Joseph Weizenbaum in 1966. It was designed to parody a type of psychotherapy called Rogerian therapy, which is rooted in the idea that people already have the tools to address their issues, if only they could access those tools properly. Weizenbaum's Eliza would take a typed message from a human and parrot back a version of it. He was not particularly enamoured with Rogerian therapy, but used it because it was easy to program and because he thought Eliza might prove his hypothesis – that human communication with a machine would be superficial.

> "THE VAST MAJORITY OF THE APP MARKETPLACE HAS NOT TESTED ITS PRODUCT."

To his surprise, however, users responded well to Eliza. Participants anthropomorphized the bot and were often eager to talk to 'her'. As Zeavin details in her 2021 book *The Distance Cure*, a number of other attempts to create automated chatbot therapists followed Eliza. These projects dovetailed with a campaign to make therapy more affordable and available than conventional psychoanalysis, with its reliance on a complex therapist–patient relationship. Throughout the latter half of the twentieth century, psychologists such as Albert Ellis, Kenneth Colby and Aaron Beck sought an approach that was more results-oriented – one that could be packaged into workbooks, recorded onto tapes, and displayed on the self-help aisles of bookshops.

These researchers ultimately converged on what would become known as cognitive behavioural therapy (CBT), which posits that psychological issues are, in part, due to counterproductive patterns of thinking that can be minimized by improving coping strategies.

Some psychologists, including Colby, subsequently attempted to automate CBT through therapy chatbots and stepwise digital programs. Weizenbaum's Eliza and the swarm of computerized therapists that followed offer a few lessons relevant to those developing automated therapy apps today: people easily open up to inanimate therapists; their experiences are largely contingent on their expectations of the platforms; and the language a bot uses to converse with a human is always, to some degree, a compromise between what might work best and what is possible to program.

The couch in your pocket

Thousands of phone apps now offer to be one's 'coach' or 'companion', or to 'boost mood'. They market themselves carefully, avoiding claims that might necessitate approval by health authorities such as the US Food and Drug Administration (FDA). Even those that do meet the FDA's definition of software as a medical device can often be approved without providing safety or efficacy data, provided that they can demonstrate substantial equivalence to products already on the market. This has enabled apps to allude to scientific claims without having to provide evidence.

Many apps, says Zeavin, are quick to co-opt the generally proven efficacy of CBT, stating that their methods are 'evidence-based'. Yet, one review[1] of 117 apps marketed to people with depression found that of the dozen that implement CBT principles, only 15% did so consistently. The low adherence could be explained by apps incorporating principles and exercises from multiple models of therapy. But another analysis of the claims made by mental-health apps found that among the platforms that cite specific scientific methods, one-third endorsed an unvalidated technique[2]. Another survey found that only 6.2% of mental-health apps publish efficacy data[3].

"The vast majority of the app marketplace has not tested its product," says Jacobson. That isn't to say that mental-health apps have no evidence as to their utility. Some perform better than others, and typically, it has been those apps with humans providing guidance and coaching that keep users engaged and progressing, says John Torous, director of Harvard Medical School's Division of Digital Psychiatry in Boston, Massachusetts. Several meta-analyses have shown that these 'guided' digital mental health programs perform comparably or better than conventional therapy[4].

Unguided apps have much less robust

evidence[5]. Some studies support their use, but, says Torous, without rigorous controls, many can be skewed by a digital placebo effect, in which people's affinity for their personal devices, and technology in general, inflates an app's perceived efficacy.

Automated therapist

Koko is far from the first platform to implement AI in a mental-health setting. Broadly, machine-learning-based AI has been implemented or investigated in the mental-health space in three roles.

The first has been the use of AI to analyse therapeutic interventions, to fine-tune them down the line. Two high-profile examples, ieso and Lyssn, train their natural-language-processing AI on therapy-session transcripts. Lyssn, a program developed by scientists at the University of Washington in Seattle, analyses dialogue against 55 metrics, from providers' expressions of empathy to the employment of CBT interventions. ieso, a provider of text-based therapy based in Cambridge, UK, has analysed more than half a million therapy sessions, tracking the outcomes to determine the most effective interventions. Both essentially give digital therapists notes on how they've done, but each service aims to provide a real-time tool eventually: part advising assistant, part grading supervisor.

The second role for AI has been in diagnosis. A number of platforms, such as the REACH VET program for US military veterans, scan a person's medical records for red flags that might indicate issues such as self-harm or suicidal ideation. This diagnostic work, says Torous, is probably the most immediately promising application of AI in mental health, although he notes that most of the nascent platforms require much more evaluation. Some have struggled. Earlier this year, MindStrong, a nearly decade-old app that initially aimed to leverage AI to identify early markers of depression, collapsed despite early investor excitement and a high-profile scientist co-founder, Tom Insel, the former director of the US National Institute of Mental Health.

The last role probably comes closest to what CBT pioneers such as Colby hoped to design, and what frightened people about the Koko experiment — the idea of a fully digital therapist that uses AI to direct treatment. Although Koko might have been the first platform to use an advanced generative AI that can create wholly original dialogue, apps such as Woebot, Wysa and Tess have all used machine learning in therapeutic chatbots for several years. But these platforms, says Torous, are probably powered by retrieval-based decision trees: essentially a flowchart that an AI navigates by logging markers of a conversation's pathway to help direct it through a set of established responses.

AI-powered therapy chatbots will need more-robust data. Koko's recent experiment, which was announced on Twitter with no published results, offered few metrics to contextualize its findings. It adds little evidence as to the efficacy of such approaches. What the experiment did accomplish, however, is to highlight the ethical questions. Notably, few of these questions are unique to AI.

> "I HAD NO IDEA I WOULD CREATE SUCH A FERVOUR OF DISCUSSION."

Rob Morris tested how people responded to GPT-3 in the mental-health app, Koko.

One concern is transparency. Both conventional, in-person therapy and automated versions have a vested interest in retaining patients. With the median retention rate for these apps dropping below 4% within 2 weeks[6], digital therapeutic platforms have a lot of room for improvement. The ethics of incentivizing patient retention are already complex — as popular mobile therapy platform TalkSpace discovered, when it came under fire for requiring therapists to insert scripts advertising its video chat features into discussions with clients. The ethics of programming a therapeutic AI chatbot to prioritize retention are murkier, particularly if bots can learn from experiences with other clients.

Privacy is a foremost consideration of all therapy. Yet, in early March, therapy app BetterHelp was fined US$7.8 million by the US Federal Trade Commission for allegedly sharing sensitive user information with advertisers. Likewise, in late March, mental-health start-up company Cerebral announced that it had leaked the data of 3.1 million people to third-party platforms such as Google, Meta and TikTok.

As machine learning becomes the basis of more mental-health platforms, designers will require ever larger sets of sensitive data to train their AIs. Nearly 70% of mental health and prayer apps analysed by the Mozilla Foundation — the organization behind the Firefox web browser — have a poor enough privacy policy to be labelled "Privacy Not Included" (see go.nature.com/3kqamow). So, wariness is warranted.

Overall, the digital therapeutic app marketplace is beholden to few clear standards. Although it would be unrealistic to hold every mental-health app marketed as a 'companion' or 'coach' to the standards that apply to conventional therapists, Insel has this month called for a new agency to oversee digital mental-health tools. The industry currently relies too heavily on a patchwork of standards proposed by psychiatric groups such as the American Psychiatric Association, and consumer guides by non-profit organizations such as One Mind PsyberGuide.

Even with apps that draw on evidence-based treatments, there is concern that as more platforms turn to AI, it could further cement CBT as the primary option for mental-health interventions. Zeavin, like others before her, argues that the quest to automate therapy could democratize it, granting access to more people. But this adds a wrinkle. Ideally, individuals should receive the diagnosis and treatment that is most accurate and effective. Automation comes with the same compromise Weizenbaum faced decades ago: balancing the best approach with the one that is easiest to program.

For all the potential benefits AI might hold in terms of access to mental-health tools, its application to therapy is still nascent, filled with ethical quagmires. GPT-3 is not yet smart enough to offer the answers, and as a Kokobot has warned: "I am just a robot."

Ian Graber-Stiehl is a science journalist in Chicago, Illinois.

1. Huguet, A. *et al. PLoS ONE* **11**, e0154248 (2016).
2. Larsen, M. E. *et al. npj Digit. Med.* **2**, 18 (2019).
3. Marshall, J. M., Dunstan, D. A. & Bartik, W. *JMIR Ment. Health* **7**, e16525 (2020).
4. Mohr, D. C. *et al. Psychiatr. Serv.* **72**, 677–683 (2021).
5. Weisel, K. K. *et al. npj Digit. Med.* **2**, 118 (2019).
6. Baumel, A., Muench, F., Edan, S. & Kane, J. M. *J. Med. Internet Res.* **21**, e14567 (2019).