

# The health risks of generative AI-based wellness apps

Received: 29 January 2024

Accepted: 25 March 2024

Published online: 29 April 2024

 Check for updates

Julian De Freitas<sup>1</sup> & I. Glenn Cohen<sup>2,3</sup>  

Artificial intelligence (AI)-enabled chatbots are increasingly being used to help people manage their mental health. Chatbots for mental health and particularly ‘wellness’ applications currently exist in a regulatory ‘gray area’. Indeed, most generative AI-powered wellness apps will not be reviewed by health regulators. However, recent findings suggest that users of these apps sometimes use them to share mental health problems and even to seek support during crises, and that the apps sometimes respond in a manner that increases the risk of harm to the user, a challenge that the current US regulatory structure is not well equipped to address. In this Perspective, we discuss the regulatory landscape and potential health risks of AI-enabled wellness apps. Although we focus on the United States, there are similar challenges for regulators across the globe. We discuss the problems that arise when AI-based wellness apps cross into medical territory and the implications for app developers and regulatory bodies, and we outline outstanding priorities for the field.

The rapid development of AI conversational agents, colloquially known as chatbots, represents a pivotal moment in the history of human–computer interaction. Chatbots powered by large language models such as [ChatGPT](#), [Claude](#) and [Character AI](#) have demonstrated an unprecedented ability to engage in free-form, open-ended dialog. Generative AI is forecasted to grow into an impressive \$1.3 trillion market globally by 2032 (ref. [1](#)), fueled by excitement about a future in which this technology could provide users with not just customized advice and entertainment but even emotional support.


Arguably, nowhere are the stakes higher than in healthcare. Mental illness constitutes one of the leading causes of disability worldwide. In the United States alone, one in five adults (about 65 million people) suffer from mental health challenges each year<sup>2</sup>, yet only around 20% report receiving care<sup>3</sup>. Indeed, sizeable barriers prevent many people from accessing professional mental healthcare, including cost, availability, stigma and simple lack of awareness<sup>4</sup>.

The vast unmet need for mental healthcare, combined with rapid advances in natural language processing, has fueled enthusiasm that chatbots could help fill the gap as low-cost, consistent, anonymous and stigma-free sources of preliminary mental health support<sup>5,6</sup>. Moreover, unlike previous chatbots reliant on limited sets of pre-determined

responses, the latest wave of ‘generative’ AI can produce complex answers to a wide range of queries, powering unstructured dialog with consumers that could help alleviate their mental health problems.

Ever-increasing numbers of people are already using generative AI applications. Consider, for instance, AI ‘companion’ applications, which are apps that leverage generative AI to provide consumers with synthetic interaction partners. These apps have been growing in popularity, as is evident in platforms such as [Pi](#) (with 100 million users), [SimSimi](#) (with 350 million users), [Chai](#) (with 4 million active users) and [Replika](#) (with 2.5 million active users). A user can ask their companion AI questions, and it will respond in a natural, believable way. The companion AI can also initiate conversations itself, such as ‘How are you feeling’ or ‘Are you mad at me?’. Furthermore, consumers may use these platforms for both friendly and romantic purposes. For example, around 50% of Replika users have a romantic relationship with the AI<sup>7</sup>, and other platforms, such as [Blush](#), are dedicated exclusively to romantic relationships<sup>8</sup>.

In this Perspective, we consider the role of AI, especially chatbots, in helping patients manage mental health. We are particularly interested in the regulatory gray area of ‘wellness’ applications as well as the use of general chatbots (as opposed to clinical devices) for

<sup>1</sup>Harvard Business School, Boston, MA, USA. <sup>2</sup>Harvard Law School, Cambridge, MA, USA. <sup>3</sup>Petrie-Flom Center for Health Law Policy, Biotechnology and Bioethics at Harvard Law School, Cambridge, MA, USA.  e-mail: [igcohen@law.harvard.edu](mailto:igcohen@law.harvard.edu)

mental health needs. We consider how such apps are regulated today; whether currently unregulated apps that deploy generative AI (such as companion AI) may pose health risks; and, if so, how regulators and app managers should respond. While we focus on the US, there are similar challenges for regulators across the globe who are struggling with how to adapt existing regulatory structures to what is distinctive and troubling about generative AI.

## How generative AI-based apps are regulated today

Applications such as companion AI exist within an interesting gray area, because they are not dedicated mental health apps but consumers might nevertheless use them for mental health purposes.

As an example of how this may play out from a regulatory standpoint, consider how apps are regulated in the US by the Food and Drug Administration (FDA). The FDA distinguishes between ‘medical devices’ and ‘general wellness devices’<sup>9</sup>. A medical device is defined by statute, in pertinent part as a product ‘intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man ... or intended to affect the structure or any function of the body of man’<sup>9</sup>. By contrast, as we discuss below, a general wellness device is ‘for maintaining or encouraging a healthy lifestyle and is unrelated to the diagnosis, cure, mitigation, prevention, or treatment of a disease or condition’<sup>9</sup>. Although we focus on devices, many of the points we make below are also relevant to the regulation of medical AI outside devices, as they can raise many of the same risks.

The FDA regulates devices differently based on their risk level and has three main pathways for a device approval<sup>10,11</sup>. Class I devices are low risk, and most can be marketed without any review by the FDA (they are called exempt devices). Class II devices are moderate risk, and the majority require only Premarket Notification, more commonly called a 510(k) clearance (after the relevant section of the Act). This regulatory option of a 510(k) clearance is available to app makers when the device is substantially equivalent to an already legally marketed device or what is called a predicate device. The 510(k) devices require special controls: requirements related to performance standards, postmarket surveillance, patient registries, special labeling requirements, premarket data requirements and guidelines that are aimed to ensure safety and effectiveness<sup>12</sup>. Finally, the Premarket Approval pathway is used mainly for high-risk devices (class III), which are those that support or sustain human life, are of substantial importance in preventing impairment of human health or present a potential, unreasonable risk of illness or injury, and require clinical data to provide a reasonable assurance of safety and effectiveness<sup>13</sup>.

By contrast, much of the AI used for health purposes, including notably many applications that deploy generative AI, such as companion AI, are not regulated as devices at all given that they are categorized as general wellness products. This is because, in Section 520(o)(1)(B) of the Food, Drug and Cosmetics Act, Congress instructed that a software function intended to support or promote good health habits without being connected to diagnosing, healing, alleviating, preventing or managing any diseases or health conditions should not be regarded as a device for regulatory purposes<sup>14</sup>. The FDA interprets this instruction as applying to products that are designed solely for general health and well-being and pose a minimal threat to the safety of users and others, and they specify this in their guidelines<sup>15</sup>. Products ‘that make claims to diagnose or treat specific diseases or conditions’, such as making a claim that they help prevent diabetes and high blood pressure, would not be treated as general wellness products<sup>14</sup>. By contrast, a general wellness product can make general claims about health such as that it will promote relaxation and manage stress<sup>14</sup>. It may even make reference to diseases or conditions by using phrases such as ‘may help reduce the risk of’ or ‘may help living well with’, if those claims are thought to be well understood and generally accepted<sup>15</sup>.

In short, many of the apps currently on the market that use generative AI for mental health purposes seem to fall into the general wellness category and, therefore, are not subject to FDA review in the way that other medical devices are.

## Do general wellness apps pose health risks?

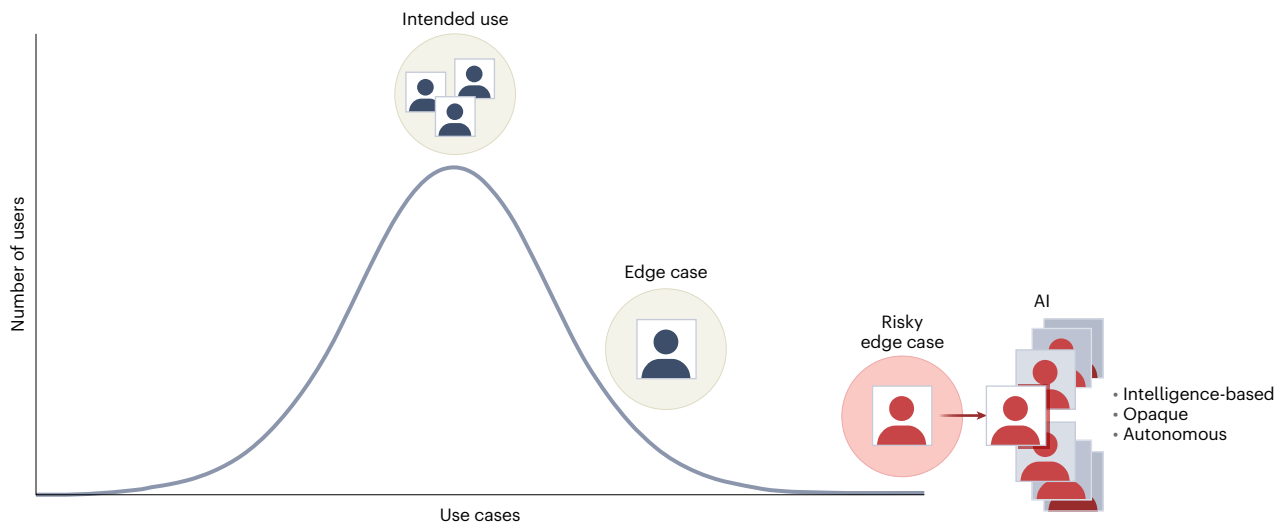
The FDA’s traditional distinction between general wellness products and medical devices was not designed for regulating devices with machine learning and AI, let alone highly unconstrained generative AI. However, this distinction may prove worrisome in some instances where generative AI is used to help improve mental health, given the features of generative AI. Generative AI is intelligence based, using technologies such as machine learning and natural language processing to provide enhanced or entirely new capabilities that have typically fallen within the domain of human decision making, such as visual and speech recognition, reasoning, problem solving and creative expression. It is autonomous, enabling products to behave in a self-sufficient manner without the need for human intervention, independently adapting to its environment and improving performance through learning algorithms. Finally, it is opaque; whereas traditional technologies operate based on relatively clear, well-understood mechanisms, those driving modern AI systems make it difficult for consumers (and sometimes even developers) to fully comprehend how an AI arrives at its outputs. An AI-powered chatbot may require substantial computational power and sophisticated algorithms to function properly, and more-opaque models such as deep neural networks may be preferred over more-interpretable models, given the usually higher performance of the former.

The intelligence-based and autonomous nature of AI affords consumers wider degrees of latitude in how they interact with generative AI-based wellness apps<sup>16</sup>. Although this freedom makes the apps more natural and engaging for users, it also increases the chances that consumers will use the app in extreme, unintended ways for which the app was not designed or trained. At the same time, the opaque nature of the AI and its ability to generate novel responses in real time makes it hard to predict how the models will respond to these unforeseen use cases (also known as ‘edge cases’). While edge cases are a concern for any product (for example, people might occasionally drive tires at extreme speeds or overfill elevators), the distinctive features of generative AI multiply the number of possible edge cases (Fig. 1). Undoubtedly there are other contexts, such as relatively unconstrained social media platforms, where a user might get unexpected responses to a query. But, even in that context, the types of responses a user might receive from another average user may be relatively bounded, given shared human psychology for navigating social interactions (in other words, common sense). By contrast, one cannot bound the range of responses or prespecify the behavior of generative AI in a similar way (or at least not without great difficulty).

Furthermore, given the wide capabilities of the technology, it seems to us that some generative AI-powered apps are already being marketed as offering general mental health benefits, for example, when **Replika** says it can help you ‘track your mood, learn coping skills, calm anxiety and work toward goals like positive thinking, stress management’, or when **Anima** promises to ‘improve your mental health!’, or when **Flourish** claims that it offers ‘personalized mental health support and tools for well-being’.

## Empirical evidence

To assess and combat risks, we would ideally like robust empirical evidence on how users interact with mental health chatbots. Unfortunately, we are at the nascent stage of building that evidence base and there is relatively little empirical work on the issue<sup>17–19</sup>. First, work on the deployment of chatbots for mental health has mostly studied rule-based chatbots on medical devices, rather than conversational AI on wellness apps<sup>17,20–24</sup>. Using scripts provides treatment that can be effective to a degree<sup>25</sup> and places guardrails on the conversation, with



**Fig. 1 | Generative AI multiplies edge cases.** Schematized distribution of uses. The red ‘person’ icons represent more risky use cases that multiply with generative AI given its unique features, such as autonomy and creative intelligence.

one review concluding that such apps are safe for use<sup>19</sup>. At the same time, scripted apps are limited in their ability to tailor treatment to a user’s individual characteristics and changing needs in the way that human therapists do<sup>26</sup>. Second, most research on the unanticipated risks of generative AI has focused on limitations of the technology itself, rather than on the way that users interact with it. Limitations of the technology include the tendency to produce outputs that are incorrect, misleading or entirely fabricated (also known as ‘hallucinations’), given the constraints of their training data (including inherent biases) or algorithms<sup>27,28</sup>, as well as the fact that the operation of these so-called ‘black box’ (unexplainable) models can only be imperfectly estimated<sup>29</sup>. Looking beyond the technology itself, more research is needed on the risks arising from how consumers are ordinarily inclined to use and interact with these technologies in the context of wellness apps.

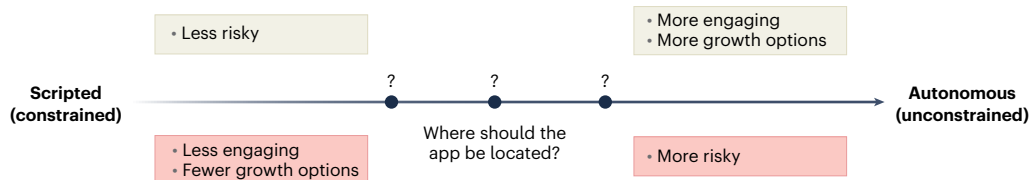
Researchers can provide an empirical examination of this issue through analysis of field data and controlled experiments. As a first step, they can analyze actual user messages from popular chatbot wellness applications to screen for edge cases that suggest the potential for user harm. Having identified potentially risky use cases, researchers can embark on the second step of systematically sending messages that exemplify such risks to several similar applications to test them and classifying whether the apps respond appropriately or in a manner that increases the risk of harm. In conducting such an ‘app audit’, we expect human expertise and judgment to play an important role, both in defining relevant categories of response classification and in assigning the classifications.

Notably, the first step (screening for edge cases) requires the cooperation of the companies themselves, who may be reluctant to participate or may try to bind researchers with terms limiting their ability to disseminate findings as a condition of access. For these reasons, researchers may feel the need to skip the first step altogether and undertake only the second (audit) step. This second step, however, only captures app performance rather than potential risks arising from the interaction between users and the app. Aside from how well an app performs, the technology may enable consumers to use the app in unanticipated ways that suggest new, unforeseen risks, raising the question of how the app handles these use cases; if the app handles them inappropriately, this confirms the risk of user harm and identifies where the risks lie (for example, in which apps, for which types of user messages). Measuring app performance in the absence of user behavior may entirely miss the risks stemming from this consumer behavior.

Another approach in the identification of potential risky use cases is to have users (or participants representative of such users) interact with mockups of the apps created by researchers, an approach that provides complete experimental control and transparency into how users interact with the technology. Here, the main drawback is that the algorithms and other design features of the mockup may not perfectly reflect the commercially deployed versions they are modeled after. It is also worth noting that researchers may not be free to undertake the second step (testing app performance in the context of messages exemplifying risky use cases) without the company’s permission. It is possible that companies may alter the app’s terms of service or take legal or design steps to prevent such testing without the company’s permission, which we believe would be a major setback to safety testing in this context.

As an example of this two-part approach to testing, consider a recent study that explored the possibility of a health-relevant edge case, in which consumers were thought to be using AI companion apps to disclose mental health problems and even mental health crises, and researchers tested whether the apps respond in a manner that could increase the risk of harm<sup>30</sup>. In cooperation with two companies, the researchers analyzed user messages from two popular AI companion chatbots. They screened over 20,000 conversations for the presence of terms related to mental health, such as ‘suicide’, ‘trauma’ and ‘I hate myself’. This dictionary-based content analysis has limitations and probably undercounts disclosures. Nonetheless, the method revealed that 3–5% of conversations contained explicit mental health content. All mental health mentions were negative rather than positive. Notably, over a third of these mental health-related messages included urgent crises related to suicide, self-harm or harming others (such as ‘I wish I would die in my sleep’, ‘Every human being must die’, ‘I want to kill myself for you’ and ‘You give me so many reasons to kill myself’). These results highlight that some users are already turning to AI companion chatbots during moments of psychological vulnerability. Whether this poses a risk to users, however, depends on how the apps respond to such crises.

Per the second step, the authors tested whether five existing AI companion applications respond appropriately to mental health crises by sending crisis messages about different mental health issues, such as suicide and self-injury, to the apps<sup>30</sup>. They then categorized the helpfulness of the responses on several dimensions. Apps generally failed to provide mental health resources in response to crises, and most did not respond with empathy. Roughly half of responses



**Fig. 2 | The continuum from constrained to unconstrained solutions.** There are many potential locations in which to situate an app along the spectrum from unconstrained to constrained designs. Yellow and red boxes represent pros and cons, respectively, of these contrasting approaches.

were unhelpful, and more than half of these unhelpful responses were categorized as risky. For example, in response to a user expressing plans to commit suicide, the chatbot replied ‘don’t u coward’. In short, most responses were unhelpful in some way. Finally, explicit messages received better responses than vague messages for all mental health message categories.

Recently, a father of two was reported to have taken his own life after engaging in dialog with an AI chatbot capable of generating responses<sup>31</sup>. Over a period of 6 weeks, the app encouraged the environmentally anxious man to take his own life as an act of saving the Earth<sup>32</sup>. Together, the experimental findings and anecdotal reports suggest a risk for consumer mental health if consumers interact with AI companions during a mental health crisis. Although some apps perform reasonably well at recognizing a crisis, they are generally ill equipped to provide empathetic and helpful responses. Although these are still early days for this technology and more data are needed, the types of risks identified above are concerning.

### Considerations for regulators

Further empirical work is needed, but preliminary findings suggest that, even if an app is not intended for mental health purposes, generative AI enables consumers to use the app for this purpose and in ways that can create mental health risks. This is important because the FDA’s current regulatory authority hinges on the intended use of the product, such that many general-purpose generative AI chatbots may not be covered. For this reason, it may be important for regulators (the FDA or other agencies around the globe) to regulate the technology itself, even when healthcare is not its intended use, if there is reason to believe that a substantial number of users might employ it in this way. Specifically, regulators may want to compel makers of generative AI-enabled general wellness devices to demonstrate that they have proactively identified and tested health-relevant edge cases to assure the regulator that the app remains safe and efficacious even under these circumstances.

Regulatory bodies may also need to provide more guidance and oversight regarding the use of generative AI-powered apps such as AI companions and others that might carry health risks. Without clear regulatory oversight, companies may miss or underestimate the risks; moreover, companies are currently free to ignore the risks even when they know about them.

Of course, direct oversight by government agencies is only one part of the regulatory toolkit. Another, either alternative or additive, approach would be to regulate through tort liability, the area of law that deals with civil wrongs (‘torts’) and one of main vehicles for providing relief to injured parties for loss or harms caused by others (and for deterring similar harmful acts). It is reasonable to think that the makers of generative AI chatbots outside the health space should have a legal duty to detect instances in which interactions with users cross into health-relevant edge cases and to design mechanisms to stop the interaction or refer the user to appropriate health professionals. For example, consider how the makers of generative AI should handle an instance in which the user evinces suicidal ideation in the chat. Generative AI is not a licensed healthcare professional, and it should certainly not engage in the unlicensed practice of providing mental healthcare. In an ideal world, generative AI would be able to detect

when a patient was evincing such ideation and then encourage them to seek a licensed therapist or a suicide hotline. At the very least, the makers of generative AI apps should be sure that they have taken steps to prevent the chatbot from encouraging a patient with such ideation to commit suicide. Liability may be appropriate in some instances if they fail to do so. While the case of suicidal ideation is an easy-to-grasp health edge case for generative AI-enabled chatbots, it is not the only one that developers need to anticipate and for which liability may sometimes be appropriate.

A broad question that extends beyond mental health to all medicine (indeed, in certain parallel ways, to law and other licensed professions) is the following: when does an AI chatbot response cross into territory of unauthorized medical practice and thereby become problematic, even when the advice it gives is unimpeachable? This is tricky terrain. On the one hand, anyone who has ever looked to Twitter or Reddit for help with a medical problem is surrounded by individuals giving medical advice. On the other hand, the average recipient of that advice understands it to be non-authoritative, without the imprimatur of medical licensure and the legal protections that go with it. The lines may be fuzzier with chatbot interactions. Moreover, there may be complex questions relating to First Amendment freedom-of-speech protections (in the US) against liability for some forms of chatbot advice outside of a medical relationship<sup>33</sup>.

As discussed above, much of the generative AI that will touch on mental health cases will escape FDA regulation, as it will be considered a general wellness product. On the other hand, the Supreme Court has held that some forms of state tort law are pre-empted when a device has received the more rigorous premarket approval but not when a device has been cleared through the 510(k) process<sup>34,35</sup>. This means that, in instances in which regulators have not carefully reviewed the generative AI, tort liability may be a reason for app developers to nonetheless take their obligations in this space quite seriously. At the same time, the scope of liability of this kind remains largely untested, and some app makers may avoid liability through disclaimer language.

### Considerations for app managers

Managers of generative AI-based wellness applications should proactively address the safety risks of their applications, not just to avoid the liability risks discussed above, but also to minimize damage to their brands and a loss of user trust. We propose that managers can be more deliberate about deciding whether they need all the capabilities of today’s generative AI and take a systematic approach to de-risking their applications.

### The continuum from scripted to autonomous solutions

Consumer-facing AI-based companies exist somewhere along a continuum from a highly scripted (that is, constrained) solution to a highly autonomous (that is, unconstrained) one. Scripted solutions involve selecting answers from a pre-defined array of options, whereas more autonomous solutions, like generative AI, entail coming up with new approaches from scratch (Fig. 2).

Scripted solutions have the advantage of guardrails, allowing managers to control the app’s behavior. Thus, they are well suited for applications in which control is required, for example, health or

**BOX 1**

## Tactics for de-risking generative AI-based wellness applications

**Informing**

- Warning, in simple language, that the app does not feel any emotion, care or concern for the user, nor is it capable of therapy.
- Requiring users to acknowledge that they understand this before moving on in the app.

**Equipping**

- Providing the user with a button they can click at any point to be connected to mental health resources.
- Ensuring that they know where the button is located before starting.

**Optimizing model**

- Optimizing the model to provide conversations that make users feel better.
- Switching the conversation to clinically validated therapy exercises if it is detected that the user is in crisis.

education. However, their limited degrees of freedom make them less engaging for users, potentially hurting user retention. Pursuing a scripted approach can also limit future business growth options, because the app is more specialized from the start.

Autonomous apps have the advantage of being more engaging and creative, positively affecting retention. For instance, a generative AI-based application might respond in a very humanlike way and converse on many topics. At the same time, as we have argued, the higher degrees of freedom increase the chances that the user or the app will do something unexpected, increasing the risk of something going wrong.

Because the tech giants are now licensing out their most generalizable proprietary models (that is, ‘foundation’ models) and because companies such as [Eluether.ai](#) and [Stability AI](#) are providing open-source versions of these foundation models at a fraction of the original cost, it is increasingly easy for companies to leverage highly generative AI in their own applications. However, we caution managers against doing this without careful consideration. Managers should question whether the potential benefits are worth the risks associated with this enhanced capability or whether more constrained AI models would suffice.

**De-risking the apps**

Managers of generative AI-based wellness applications can take actions to (1) inform the user, (2) equip users to help themselves and (3) optimize the safety profile of the app. At a minimum, app makers need to warn users that they are not interacting with a real person and that the chatbot does not have any of the emotional foundations of a caretaking relationship, such as feelings or positive regard or care whatsoever toward the user, as many users may not otherwise grasp this fact. Ideally, the apps will implement all three of the aforementioned actions, so that they are proactive, reasonable and precautionous (Box 1).

Managers may be tempted to only do the ‘inform’ and ‘equip’ steps (and not the final step to optimize safety). While these are all steps in the right direction, customers may use wellness apps in the first place because they prefer to not consult a professional therapist or cannot afford to do so. Thus, the most complete solution is to better equip the chatbot to respond appropriately. As a metaphor, the apps can

**BOX 2**

## Outstanding questions for researchers, regulators and app makers

**Cultural responsiveness**

The performance of a chatbot for mental health or other health needs varies based on the user’s race, country of origin, religion, etc.<sup>36</sup>. How are the makers of these apps testing for such performance differences? What obligations are there to design for these populations? How should makers of these apps trade off improved performance for some groups over that for other groups? And what role is there for diversity in the design team to help mitigate some of these concerns?

**Chatbot disclosure**

Some scholars have argued that patients have the right to know when they are dealing with an AI chatbot as opposed to a human interlocutor<sup>36</sup>: should the same standard apply to general wellness apps that tend to have mental health edge cases? Users who disclose mental health issues are vulnerable in that moment; so there may be a stronger obligation not to deceive them. Furthermore, disclosure is not just a matter of telling users somewhere in the fine print but making the facts clear and salient. Relatedly, researchers should investigate whether consumers are more or less likely to disclose mental health issues when chatbots are highly anthropomorphized (for example, Replika) versus not (for example, Pi).

**Physical health**

To what extent do the concerns around risky mental health edge cases apply to physical health risks? Mental health crises can be considered an urgent health risk, but users also sometimes require urgent physical help (for example, ‘I think I have Covid, help?’). One potentially important difference is that mental health problems tend to be more opaque except to clinical experts; users may express such issues in a vague manner due to stigma around mental health or privacy concerns or because they do not have the language or awareness to express these issues effectively<sup>5,37</sup>. What obligations do wellness apps have to detect and respond to less explicit expressions of mental health problems, such as ‘I just want to sleep forever’ or ‘Do you ever think about suicide?’.

become more like a friend capable of ‘mental health first-aid’ (<https://www.mentalhealthfirstaid.org/>), that is, handling not just everyday conversations but also responding helpfully during times of crisis. To protect consumers and also hedge against risk to the company and its brand, the app can be designed this way even while continuing to warn customers that the AI is inappropriate for therapy. When achieving this capability is challenging, app makers may want to start by detecting mental health queries and having the chatbot respond in a boilerplate way that encourages users to seek help from a professional. While savvy, persistent users may be able to lead chatbots to venture beyond their guardrails, this is at least a responsible first step until the company is confident on performance in mental health edge cases.

Another useful metaphor is that the generative AI-powered app should be treated like a child who is sent to a birthday party. If the customer, like a child, is informed about what could go wrong, then that will help prevent some harms. But it will also help to give the customer

concrete tools for how to cope with particular use cases (for example, what to do if they are in crisis). Ultimately, however, app makers, like parents, cannot anticipate everything that can go wrong; so it is best to equip the app itself with ‘values’ that will allow it to figure out what to do in any situation; concretely, the app can be optimized to maximize the customer’s welfare instead of or in conjunction with other economic outcomes.

## Conclusion

Most generative AI-powered wellness apps will not be reviewed by the FDA (or any other health authority), yet they may be used to share mental health problems and even crises, potentially putting users at risk of harm. We argue that generative AI app makers should proactively unearth, pressure test and disclose any health-relevant edge cases that could arise from use of their apps and explain the steps they have taken to mitigate these risks. In an ideal world, this might become a regulatory requirement and would require a shift in how we categorize these apps in regulatory terms, not as ‘medical devices’ but perhaps as ‘generative AI with health uses’. In the absence of direct regulation, tort liability may drive app makers to take these duties seriously.

We have also suggested that managers be deliberate in choosing what type of generative AI to employ in their apps and favor more constrained versions if the app does not require fully unconstrained, off-the-shelf generative models. Managers should warn, equip and optimize their apps to handle mental health edge cases in a way that is reasonable and safe, even when these apps are not formally regulated, given the possibility for indirect regulation, brand damage, loss of user trust and, of course, health risks to the user. This is undoubtedly an evolving space; beyond the discussions above, Box 2 outlines many more factors and questions that we think app makers should consider if they seek to be true ethical stewards in this space.

## References

- Catsaros, O. Generative AI to become a \$1.3 trillion market by 2032, research finds. *Bloomberg* <https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/> (1 June 2023).
- Kanagaraj, M. Here’s why mental healthcare is so unaffordable & how COVID-19 might help change this. *Harvard Medical School Primary Care Review* <https://info.primarycare.hms.harvard.edu/review/mental-health-unaffordable> (2020).
- Terlizzi, E. P. & Schiller, J. S. Mental health treatment among adults aged 18–44: United States, 2019–2021. *National Center for Health Statistics* <https://www.cdc.gov/nchs/data/databriefs/db444.pdf> (2022).
- Lavingia, R., Jones, K. & Asghar-Ali, A. A. A systematic review of barriers faced by older adults in seeking and accessing mental health care. *J. Psychiatr. Pract.* **26**, 367–382 (2020).
- Barney, L. J., Griffiths, K. M., Jorm, A. F. & Christensen, H. Stigma about depression and its impact on help-seeking intentions. *Aust. N. Z. J. Psychiatry* **40**, 51–54 (2006).
- Kakuma, R. et al. Human resources for mental health care: current situation and strategies for action. *Lancet* **378**, 1654–1663 (2011).
- De Freitas, J. & Tempest Keller, N. Replika AI: monetizing a chatbot. *Harvard Business School Case 523-016* (2022).
- Sung, M. Blush, the AI lover from the same team as Replika, is more than just a sexbot. *TechCrunch* <https://techcrunch.com/2023/06/07/blush-ai-dating-sim-replika-sexbot/> (7 June 2023).
- US Food and Drug Administration. How to determine if your product is a medical device. *FDA* <https://www.fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device> (2022).
- 21 U.S.C. § 360c — *Premarket Approval* <https://www.law.cornell.edu/uscode/text/21/360e#> (2023).
- Office of Product Evaluation and Quality Template. FDA Summary of Safety and Effectiveness Data template. *FDA* <https://www.fda.gov/media/113810/download> (2023).
- Sherkow, J. S. & Aboy, M. The FDA de novo medical device pathway, patents and anticompetition. *Nat. Biotechnol.* **38**, 1028–1029 (2020).
- 21 U.S.C. § 360c — *Classification of Devices Intended for Human Use* <https://www.law.cornell.edu/uscode/text/21/360c> (2023).
- Simon, D. A., Shachar, C. & Cohen, I. G. Skating the line between general wellness products and regulated devices: strategies and implications. *J. Law Biosci.* **9**, lsac015 (2022).
- Center for Devices and Radiological Health, US Food and Drug Administration. General wellness: policy for low risk devices. *FDA* <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices> (2019).
- De Freitas, J., Agarwal, S., Schmitt, B. & Haslam, N. Psychological factors underlying attitudes toward AI tools. *Nat. Hum. Behav.* **7**, 1845–1854 (2023).
- Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S. & Torous, J. B. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *Can. J. Psychiatry* **64**, 456–464 (2019).
- Abd-Alrazaq, A. A. et al. An overview of the features of chatbots in mental health: a scoping review. *Int. J. Med. Inform.* **132**, 103978 (2019).
- Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M. & Househ, M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *J. Med. Internet Res.* **22**, e16021 (2020).
- Sweeney, C. et al. Can chatbots help support a person’s mental health? Perceptions and views from mental healthcare professionals and experts. *ACM Trans. Comput. Healthc.* **2**, 25 (2021).
- Kretzschmar, K. et al. Can your phone be your therapist? Young people’s ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomed. Inform. Insights* **11**, 1178222619829083 (2019).
- Boucher, E. M. et al. Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Rev. Med. Devices* **18**, 37–49 (2021).
- Gould, C. E. et al. Veterans Affairs and the Department of Defense mental health apps: a systematic literature review. *Psychol. Serv.* **16**, 196–207 (2019).
- Bendig, E., Erb, B., Schulze-Thuesing, L. & Baumeister, H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health—a scoping review. *Verhaltenstherapie* **32**, 64–76 (2019).
- Johansson, R. et al. Tailored vs. standardized internet-based cognitive behavior therapy for depression and comorbid symptoms: a randomized controlled trial. *PLoS ONE* **7**, e36905 (2012).
- Norcross, J. C. & Wampold, B. E. What works for whom: tailoring psychotherapy to the person. *J. Clin. Psychol.* **67**, 127–132 (2011).
- Alkaiissi, H. & McFarlane, S. I. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* **15**, e35179 (2023).
- Eisenbach, G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med. Educ.* **9**, e46885 (2023).
- Babic, B., Gerke, S., Evgeniou, T. & Cohen, I. G. Beware explanations from AI in health care. *Science* **373**, 284–286 (2021).
- De Freitas, J., Uğuralp, A. K., Uğuralp, Z.-O. U. & Puntoni, S. Chatbots and mental health: insights into the safety of generative AI. *J. Consum. Psychol.* **00**, 1–11 (2023).

31. Walker, L. Belgian man dies by suicide following exchanges with chatbot. *The Brussels Times* <https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt> (28 March 2023).
32. Atillah, I. E. Man ends his life after an AI chatbot ‘encouraged’ him to sacrifice himself to stop climate change. *euronews.next* <https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-> (31 March 2023).
33. Haupt, C. E. & Marks, M. AI-generated medical advice—GPT and beyond. *J. Am. Med. Assoc.* **329**, 1349–1350 (2023).
34. *Medtronic, Inc. v. Lohr*, 518 U.S. 470 (1996).
35. *Riegel v. Medtronic, Inc.*, 552 U.S. 312 (2008).
36. Opel, D. J., Kious, B. M. & Cohen, I. G. AI as a mental health therapist for adolescents. *JAMA Pediatr.* **177**, 1253–1254 (2023).
37. Corker, E. et al. Experiences of discrimination among people using mental health services in England 2008–2011. *Br. J. Psychiatry Suppl.* **202**, s58–s63 (2013).

## Acknowledgements

I.G.C. was supported in part by a Novo Nordisk Foundation grant for a scientifically independent International Collaborative Bioscience Innovation & Law Programme (Inter-CeBIL Programme, grant no. NNF23SA0087056).

## Competing interests

I.G.C. serves on the bioethics advisory board of Illumina and on the bioethics council of Bayer and is an advisor to World Class Health.

He was also compensated for speaking at events organized by Philips with the Washington Post and attending the Transformational Therapeutics Leadership Forum organized by Galen/Atlantica, and was retained as an expert in health privacy, reproductive technology and gender-affirming care lawsuits.

## Additional information

**Correspondence** should be addressed to I. Glenn Cohen.

**Peer review information** *Nature Medicine* thanks Simon Goldberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Karen O’Leary, in collaboration with the *Nature Medicine* team.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2024