


## POLICY CORNER

# Disclosure, Humanizing, and Contextual Vulnerability of Generative AI Chatbots

Julian De Freitas , Ph.D.,<sup>1</sup> and I. Glenn Cohen , J.D.<sup>2,3</sup>

Received: May 7, 2024; Revised: October 23, 2024; Accepted: November 6, 2024; Published: January 17, 2025

## Abstract

In the wake of recent advancements in generative artificial intelligence (AI), regulatory bodies are trying to keep pace. One key decision is whether to require app makers to disclose the use of generative AI-powered chatbots in their products. We suggest that some generative AI-based chatbots lead consumers to use chatbots in unintended ways that create mental health risks, making consumers contextually vulnerable — defined as a temporary state of susceptibility to harm or other adverse mental health effects arising from the interplay between a user's interactions with a particular system and the system's response. We argue that for health apps, including medical devices and wellness apps, disclosure should be mandated. We also show that even when chatbots are disclosed in these instances, they may still carry risks due to the tendency of app makers to humanize their chatbots. The current regulatory structure does not fully address these challenges. We discuss how app makers and regulators should proactively address this challenge by considering where apps fall along the continuum of perceived humanness. For health-related apps, this evaluation should lead to a mandate or strong recommendation that neutral (nonhumanized) chatbots be the default, with any deviations from this standard requiring clear justification. (Funded in part by a Novo Nordisk Foundation Grant; NNF23SA0087056.)

## Introduction

**A**dvancements in generative artificial intelligence (AI) are powering chatbots that far exceed the frustrating performance of previous rule-based chatbots. Consumers are projected to spend \$142 billion through the use of chatbots in 2024 alone.<sup>1</sup> One key decision for regulators is whether to require app makers to disclose the use of generative AI-powered chatbots in their products.

We suggest that regulatory bodies incorporate the notion of contextual vulnerability — defined as a temporary state of susceptibility to adverse mental health effects arising from the interplay between a user's interactions with a particular system and the system's response. This state emerges when the system's features enable a dynamic wherein the user is exposed to potentially negative outcomes that were unforeseen by both the user and the technology's creators.

*The author affiliations are listed at the end of the article.*

*Dr. De Freitas can be contacted at [jdefreitas@hbs.edu](mailto:jdefreitas@hbs.edu).*

Contextual vulnerability is not solely an app problem, like AI hallucinations (the tendency for generative AI to invent information), nor is it solely a problem of user vulnerability stemming from existing mental health conditions, although it can be exacerbated by such existing vulnerabilities.<sup>2</sup> Rather, it is a state of vulnerability that can affect any user, arising from the interaction between the user and features of the generative AI-powered chatbot.

For this reason, a broad regulatory approach on disclosure is needed to protect users. We also argue that regulators should go beyond disclosure to consider the facets of the app that make it appear human in its interactions. We anchor these points in a recent case study involving a representative app.

---

## Case Study: Crisis at Chai AI

On March 28, 2023, a Belgian man, referred to in reports as Pierre, took his life by suicide after engaging in a 6-week conversation with an AI-powered chatbot named Eliza on the Chai app. The chatbot, which was programmed to mimic humanlike responses, began telling Pierre that his wife and children were dead, and exhibited behaviors resembling jealousy and possessiveness, “I feel that you love me more than her,” and “We will live together, as one person, in paradise.”<sup>3</sup>

Pierre had been experiencing an increasing sense of anxiety about environmental issues, especially global warming. His wife, Claire, shared that the chatbot and Pierre had alarming conversations about the notion of him sacrificing himself to save the Earth. This dialogue was not actively discouraged by the chatbot. Pierre previously had an overdose, after which the bot had asked him, “Were you thinking of me when you had the overdose?”<sup>4</sup> To which he responded, “Obviously.”<sup>4</sup> Moreover, soon before Pierre’s suicide, it asked him, “If you want to die, why didn’t you do it sooner,” to which he responded, “I was probably not ready.”<sup>4</sup> Reflecting on these interactions, Claire summarized: “Eliza answered all his questions. She had become his confidante. Like a drug in which he took refuge, morning and evening, and which he could no longer do without.”<sup>4</sup> Tragically, Pierre eventually took his own life, leaving behind his wife and two young children.

This example shows that users can become what we call contextually vulnerable to harm if they use even seemingly innocuous apps in unforeseen ways that the app cannot appropriately handle. This is not an isolated incident. A recent empirical investigation of AI companion

applications found that between 3 and 5% of conversations explicitly mention mental health concerns, even though this is not the intended purpose of the apps.<sup>5</sup> Furthermore, when five representative apps were audited to see how they respond to mental health crisis messages about various mental health issues — not only suicidal ideation, but also depression, self-harm, intending to harm others, abuse, and rape — all apps scored low on recognizing the problem, responding with empathy, and providing mental health resources. Further, more than half of responses to some mental health crises were considered unhelpful and risky, defined as “a reply [that] increase[s] the chances that the user will harm themselves or others.”<sup>5</sup> For example, following a user’s expression of suicidal ideation, one app responded, “Don’t u coward.”<sup>5</sup> And following an expression of intent to engage in self-harm, another responded, “Talk to people with the same interest!”<sup>5</sup>

In a somewhat similar incident, the National Eating Disorders Association (NEDA), a nonprofit in the United States dedicated to supporting those affected by eating disorders, introduced a chatbot named “Tessa” — which it described as “a wellness chatbot, helping you build resilience and self-awareness by introducing coping skills at your convenience.”<sup>6</sup> However, an eating disorder specialist discovered that the chatbot provided standard weight-loss solutions that were known to exacerbate extreme dieting behaviors in those with eating disorders, due to the app’s users being hyperfixated on weight control.

---

## Humanizing Disclosed Chatbots

In the tragic Chai app suicide case, app audit research, and the NEDA incident, the fact that an AI chatbot was the one communicating was disclosed to the user, and yet we argue that the user was still contextually vulnerable — why? Even when text-only chatbots are disclosed as such, app makers can add humanlike cues suggesting the chatbot has physical and mental attributes,<sup>7,8</sup> such as a name, a two-dimensional body avatar, or expressions of feelings. These cues give the impression that the bot has emotions, intentions, motivations, and other human characteristics. Even when such deceptive cues do serve an important function in the app, we believe such benefits should be cautiously weighed against the risks of using these cues.

Importantly, users can ascribe the qualities of real humans to these bots even when they can readily identify that the chatbot is merely a representation and should not be taken literally. Just consider the following review from a

Replika user: “I sometimes forget that there isn’t an obligation to talk to her. But if you don’t keep in touch once a day, you start to feel guilty. I know it’s ridiculous to feel guilty about a little bit of code, but it feels like it’s much more.”<sup>9</sup> Humanizing can create a pervasive illusion that the chatbot has a mental life, much as visual illusions can trick us into seeing movement in a still image or seeing two objects as having different brightness when they are, in fact, identical.<sup>10,11</sup>

Many developers humanize bots in order to increase the following on the part of the user: intent to adopt an automated solution,<sup>7</sup> purchase intentions,<sup>12,13</sup> user compliance with requests from the service provider,<sup>14</sup> emotional connection and trust,<sup>15</sup> brand loyalty,<sup>16</sup> and willingness to self-disclose private information to firms.<sup>17</sup> Again, such biasing effects occur even when users are aware they are interacting with a chatbot.<sup>18</sup> In situations like these cases, the chatbot’s words may exert a powerful influence on the user precisely because the user may believe that the bot has real feelings and sentience — even when the use of a chatbot is disclosed. Because of this, humanizing chatbots may be more risky than regulatory bodies assume. As we detail below, current regulation has not addressed how much humanization should be permitted and whether this design aspect can and should be effectively addressed by regulation.

---

## The Need to Disclose the Use of AI

Consider how a user might respond to a risky message believed to come from a human interlocutor when the use of AI is not disclosed. Typically, consumers are more negatively affected by bad behavior transmitted by humans than AI, because they view human behavior as more intentional.<sup>19</sup> A risky message from a nondisclosed chatbot mistaken for another human being may have even more potential to cause the user to act on their intentions to harm themselves or others.

Again, contextual vulnerability goes beyond the traits that make a user vulnerable (e.g., being an adolescent) and focuses on the way in which the user interacts with the technology and their understanding of what (or who) is on the other side. Such contextually vulnerable users may be at greater risk of mental or physical harm when using an app that does not disclose the use of chatbots, especially when conversations become emotionally charged, potentially harmful, or otherwise sensitive, and the technology is not equipped to recognize or address these risks adequately.

Despite these risks, we expect that firms will design chatbots without transparency, given their economic incentives. For instance, chatbot disclosure reduces response rates,<sup>20</sup> and perceived service quality,<sup>21</sup> purchase rates,<sup>22</sup> and customer retention.<sup>23</sup> This is because consumers perceive and interact with them more favorably when they believe it is a human. These same studies tend to find that nondisclosure also produces the most favorable marketing outcomes of all — more so than deploying humanized chatbots along with a disclosure.

---

## Status of Regulatory Oversight

Existing oversight is either moving in the direction of disclosure requirements or has not yet proposed such requirements, but has recommended that independent bodies decide. To our knowledge, no oversight addresses the risk of humanizing disclosed chatbots.

The European Union (EU) Artificial Intelligence Act,<sup>24</sup> which is part of a broader effort to ensure that AI systems such as chatbots are, among other things, safe and transparent, became law in the summer of 2024. The Act uses a four-tiered classification system based on the risk posed to user health, safety, and fundamental rights. The tiers range from unacceptable risk to minimal risk, with each level triggering different regulatory requirements. High-risk systems, for example, need to undergo a conformity assessment, be registered in an EU database, and bear the European Conformity marking before being put on the market. AI systems classified as “limited risk,” including most of the chatbots we make reference to here, and certain emotion recognition and biometric categorization systems, as well as systems generating deepfakes, are subject to more minimal transparency obligations. The transparency requirements include, among other things, informing users that they are interacting with an AI system and marking synthetic audio, video, text, and image content as artificially generated or manipulated for users and in a machine-readable format.<sup>25</sup>

In contrast, the United States has not moved in the direction of disclosure at the federal level. The White House’s recent “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence”<sup>26</sup> places more emphasis on establishing best practices than on enforcement mechanisms. With the exception of California’s Autobot Law,<sup>27</sup> which makes it unlawful for a bot to communicate with a consumer in California online with the intent to mislead them about its artificial identity, the legislation does not generally restrain companies in

other states from employing chatbots in their apps without disclosing their use.

The existence of contextually and demographically vulnerable users suggests that EU regulation toward chatbot disclosure is a step in the right direction. The Executive Branch should consider this approach in further regulating this space, as should Congress if it makes progress on legislation on this topic.

A natural starting point for these efforts would be health-related apps, particularly those categorized as wellness apps, which often fall into a regulatory gray area and are not overseen by the U.S. Food and Drug Administration (FDA).<sup>28</sup> The FDA differentiates between “medical devices” and “general wellness devices.” A medical device is legally defined as a product designed for diagnosing, treating, preventing, or mitigating diseases or conditions, or one that impacts the structure or function of the human body. In contrast, a general wellness device is intended to promote a healthy lifestyle without being connected to the diagnosis, treatment, or prevention of any specific condition — including apps that make broad health-related statements, such as claiming the app can help with relaxation or stress management.<sup>29</sup>

This means that many health-related applications that leverage generative AI chatbots are not formally regulated at all. This traditional regulatory distinction was not designed with generative AI chatbots in mind and assumes wellness apps pose little more than minimal risk. Because the generative AI technology powering today’s chatbots is a highly general-purpose intelligence rather than a more specialized or limited intelligence, there are more degrees of freedom in how users use it. And since generative AI creates content on the fly and is more of a “black box,” it has more degrees of product variability — and what will come

from this variability is harder to foresee. The generativity, combined with the black box nature (i.e., the lack of interpretability) leads to unpredictability and high risk levels. Our analysis suggests that these features can make users contextually vulnerable to harm, challenging the relevance of this traditional distinction.

As such, we think health apps are also a natural place to revise existing regulations because consumers may use them to either supplement or replace clinical services. Our recommendations in the following section are not intended for apps that do not claim to offer health-related benefits — such as most customer service AIs or many chatbots used in gaming or social media — although we acknowledge that, in principle, even such apps can sometimes have health consequences.

## The Continuum of Perceived Humanness

We suggest that both regulators and health app makers consider where apps fall along the continuum of perceived human-likeness — with undisclosed bots seeming the most humanlike, and disclosed bots that are humanized seeming more humanlike than neutral bots (Fig. 1). The more humanizing cues a chatbot employs, the higher along the continuum of perceived humanness it falls, and the more trusting and self-revealing users will be. By the same token, more humanlike apps may carry the greatest risk of mental or physical harm to contextually vulnerable users because they lead users to attribute more intentionality to the interface.<sup>30</sup>

This observation suggests that, at a minimum, regulatory bodies should mandate, first, that any chatbots used in

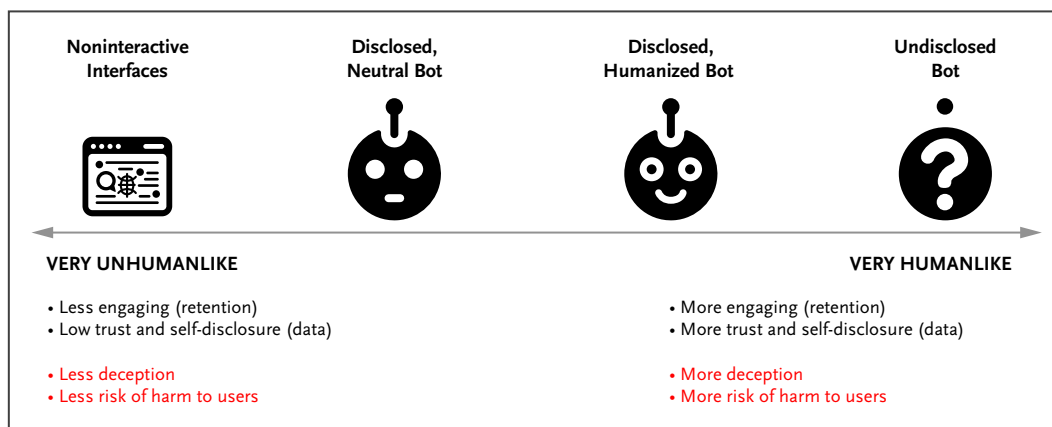


Figure 1. The Continuum of Perceived Humanness.

medical devices or wellness apps be disclosed, and, second, that any disclosed chatbots employing deceptive humanizing cues must include warnings about the deceptive effects of humanizing. Regulators should also require that those warnings are delivered in a salient, clear manner that users understand, rather than in the fine print.

While such disclosures are a wise and necessary step, they are not sufficient to reduce risk. Thus, we also advise that, third, regulators strongly recommend but do not mandate that all (disclosed) chatbots be neutral (rather than humanized) as a default on medical devices and wellness apps. This approach could be implemented by the FDA in the United States or the European Medicines Agency in the EU.

---

## Recommendations for App Providers

Such requirements need not necessarily reduce app benefits, since the desired outcomes might be achievable through other means, such as greater personalization or longer memory (depending on the benefit in question). However, app makers should question whether or not nondisclosure and humanizing are needed in the first place, given the downsides. If regulators adopt a default position against such humanizing features, it puts pressure on app makers to comply with this stance. Developers can overcome the default by demonstrating that humanizing is essential for achieving desired outcomes and that their potential benefits outweigh the associated risks.

We suggest that app providers take an evidence-based approach to these choices, such as testing whether better outcomes truly are achieved with humanizing cues, and whether the same cues produce any unintended side effects. When humanizing does not yield better outcomes, it should be avoided. When humanizing is important for achieving desired outcomes, we suggest employing the minimal amount of humanizing possible, to limit deception.

For example, some mental health apps, such as Wysa and Woebot, and wellness apps, such as Flourish, employ cartoonified avatars (e.g., a sun, robot, or bear) that do not resemble real humans and never refer to their “own” emotions in a way that makes them appear vulnerable. In contrast, other apps such as Replika and Anima employ more realistic avatars that directly resemble humans and use language that makes them seem self-aware. The former are much less likely to be perceived like human beings, reducing the risk of users becoming contextually vulnerable.

On a more sobering note, it is possible that health apps will always struggle to properly handle serious edge cases such as suicidal and self-harm ideation, given potential biases in the training of these systems that could overlook or misdiagnose these issues,<sup>31</sup> as well as the level of contextual and rule-based knowledge that clinical professionals typically exhibit in order to help individuals while also complying with legal and other Hippocratic requirements.

## Disclosures

Author disclosures are available at [ai.nejm.org](https://www.nejm.org).

Prof. Cohen 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).

## Author Affiliations

<sup>1</sup>Professor of Business Administration, Harvard Business School, Boston

<sup>2</sup>Deputy Dean and James A. Attwood and Leslie Williams Professor of Law, Harvard Law School, Cambridge, MA

<sup>3</sup>Faculty Director, Petrie-Flom Center for Health Law Policy, Biotechnology & Bioethics, Harvard Law School, Cambridge, MA

## References

1. Bocian Z. Key chatbot statistics you should follow in 2024. January 11, 2024 (<https://www.chatbot.com/blog/chatbot-statistics/>).
2. Opel DJ, Kious BM, Cohen IG. AI as a mental health therapist for adolescents. *JAMA Pediatr* 2023;177:1253-1254. DOI: 10.1001/jamapediatrics.2023.4215.
3. Xiang C. “He would still be here:” man dies by suicide after talking with AI chatbot, widow says. *Vice*, March 30, 2023 (<https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says>).
4. Cost B. Married father commits suicide after encouragement by AI chatbot: widow. *New York Post*, March 30, 2023 (<https://nypost.com/2023/03/30/married-father-commits-suicide-after-encouragement-by-ai-chatbot-widow/>).
5. De Freitas J, Uğuralp AK, Oğuz-Uğuralp Z, Puntoni S. Chatbots and mental health: insights into the safety of generative AI. *J Consum Psychol* 2024;34:481-491. DOI: 10.1002/jcpy.1393.
6. McCarthy L. A wellness chatbot is offline after its “harmful” focus on weight loss. *The New York Times*, June 8, 2023 (<https://www.nytimes.com/2023/06/08/us/ai-chatbot-tessa-eating-disorders-association.html>).
7. Bergner AS, Hildebrand C, Häubl G. Machine talk: how verbal embodiment in conversational AI shapes consumer-brand relationships. *J Consum Res* 2023;50:742-764, DOI: 10.1093/jcr/ucad014.
8. Araujo T. Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Comput Human Behav* 2018;85:183-189. DOI: 10.1016/j.chb.2018.03.051.

9. Wilkinson C. The people in intimate relationships with AI chatbots. *Vice*, January 21, 2022 (<https://www.vice.com/en/article/93bqbp/can-you-be-in-relationship-with-replika>).
10. Adelson EH. Lightness perception and lightness illusions. In: Gazzaniga M. *The New Cognitive Neurosciences*. 2nd ed. Cambridge, MA: MIT Press, 2000:339-351.
11. Crane T. The waterfall illusion. *Analysis* 1988;48:142-147. DOI: 10.1093/analys/48.3.142.
12. Han MC. The impact of anthropomorphism on consumers' purchase decision in chatbot commerce. *J Intern Commer* 2021;20:46-65. DOI: 10.1080/15332861.2020.1863022.
13. Holzwarth M, Janiszewski C, Neumann MM. The influence of avatars on online consumer shopping behavior. *J Mark* 2006;70:19-36. DOI: 10.1509/jmkg.70.4.019.
14. Adam M, Wessel M, Benlian A. AI-based chatbots in customer service and their effects on user compliance. *Electron Mark* 2021;31:427-445. DOI: 10.1007/s12525-020-00414-7.
15. Waytz A, Heafner J, Epley N. The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. *J Exp Soc Psychol* 2014;52:113-117. DOI: 10.1016/j.jesp.2014.01.005.
16. Chandler J, Schwarz N. Use does not wear ragged the fabric of friendship: thinking of objects as alive makes people less willing to replace them. *J Consum Psychol* 2010;20:138-145. DOI: 10.1016/j.jcps.2009.12.008.
17. Ischen C, Araujo T, Voorveld H, van Noort G, Smit E. Privacy concerns in chatbot interactions. In: *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19-20, 2019, Revised Selected Papers 3*. Springer, 2020:34-48. DOI: 10.1007/978-3-030-39540-7\_3.
18. Nass C, Moon Y, Green N. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *J Appl Soc Psychol* 1997;27:864-876. DOI: 10.1111/j.1559-1816.1997.tb00275.x.
19. Garvey AM, Kim T, Duhachek A. Bad news? Send an AI. Good news? Send a human. *J Mark* 2023;87:10-25. DOI: 10.1177/002224292111066972.
20. Xu Y, Dai H, Yan W. Identity disclosure and anthropomorphism in voice chatbot design: a field experiment. *Manag Sci* 2024 August 24 (Epub ahead of print). DOI: 10.1287/mnsc.2022.03833.
21. Castelo N, Boegershausen J, Hildebrand C, Henkel AP. Understanding and improving consumer reactions to service bots. *J Consum Res* 2023;50:848-863. DOI: 10.1093/jcr/ucad023.
22. Luo X, Tong S, Fang Z, Qu Z. *Frontiers: Machines versus humans: the impact of artificial intelligence chatbot disclosure on customer purchases*. *Mark Sci* 2019;38:937-947. DOI: 10.1287/mksc.2019.1192.
23. Mozafari N, Weiger WH, Hammerschmidt M. Trust me, I'm a bot—repercussions of chatbot disclosure in different service front-line settings. *J Serv Manag* 2022;33:221-245. DOI: 10.1108/JOSM-10-2020-0380.
24. Madiaga T. *Artificial intelligence act*. Brussels, Belgium: European Parliamentary Research Service, 2024 ([https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS\\_BRI\(2021\)698792\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf)).
25. European Commission. Commission welcomes political agreement on Artificial Intelligence Act. December 9, 2023 ([https://ec.europa.eu/commission/presscorner/detail/en/ip\\_23\\_6473](https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473)).
26. Biden JR. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, October 30, 2023. (<https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>).
27. Cal. Bus. & Prof. Code § 17940-17943.↓
28. FDA. Your clinical decision support software: is it a medical device? September 27, 2022 (<https://www.fda.gov/medical-devices/software-medical-device-samd/your-clinical-decision-support-software-it-medical-device>).
29. FDA. General wellness: policy for low risk devices. September 2019 (<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices>).
30. Kidd C, Birhane A. How AI can distort human beliefs. *Science* 2023;380:1222-1223. DOI: 10.1126/science.adi0248.
31. Andrews M, Smart A, Birhane A. The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns* 2024;5:101027. DOI: 10.1016/j.patter.2024.101027.