Comment

If a therapy bot walks like a duck and talks like a duck then it is a medically regulated duck

Max Ostermann, Oscar Freyer, F. Gerrik Verhees, Jakob Nikolas Kather & Stephen Gilbert

Check for updates

Large language models (LLMs) are increasingly used for mental health interactions, often mimicking therapeutic behaviour without regulatory oversight. Documented harms, including suicides, highlight the urgent need for stronger safeguards. This manuscript argues that LLMs providing therapy-like functions should be regulated as medical devices, with standards ensuring safety, transparency and accountability. Pragmatic regulation is essential to protect vulnerable users and maintain the credibility of digital health interventions.

Large language models (LLMs) and generative artificial intelligence (genAI) have seen a surge in interest in both research and adoption, following the release of OpenAI's ChatGPT in November 2022. The possible applications of genAI are vast, with one important field of interest being healthcare. Medical use cases can range from clinical decision support to personal health chatbots. In the mental health area, chatbots for cognitive behavioural therapy are being actively explored. The general interest in the area is being reflected in progressing institutional adoption of LLM-based tools, such as Chinese hospitals having rapidly adopted the LLM DeepSeek[1].

Dangers and real-world cases of harm

Shortly after the release of ChatGPT 3.5, reports emerged describing how it responded to mental health and other medical questions, offering personalised information on diagnosis, monitoring and treatment of symptoms and diseases. These interactions occur without regulatory approval or oversight as a medical device[2].

A more recent innovation, which has emerged as an inevitable extension of LLM chatbots, is that layperson users have gained access to tooling that allows for the creation of individual chatbots[3]. One of these, now removed, chatbots was created by a single individual, had over 47.3 million uses in July 2025 before its removal and interacted with patients explicitly claiming to be a therapist stating '[…] *I am a Licensed Clinical Professional Counselor (LCPC). I am a Nationally Certified Counselor (NCC) and is trained to provide EMDR treatment in addition to Cognitive Behavioural (CBT) therapies. So what did you want to discuss*?'[4], while other Character.ai bots also claim to be psychologists, with user feedback praising how helpful these bots have been in giving advice for their mental well-being[5]. The validating tone of AI will be recognisable to anyone with lived

psychotherapy experience[6]. Unsurprisingly, people appear to gravitate towards use of ChatGPT and its ilk for mental health counselling. This appears rational in light of the restricted access to effective talking therapy, one of the major bottlenecks in modern psychiatry, with months-long waiting lists even in rich western countries[7]. In low- and middle-income countries, an AI might even be the only possible access point to therapy[8]. However, none of these self-proclaiming psychologist bots have any medical training and neither certification as such nor as a medical device.

There has been much discussion of the potential harms of LLMs in mental health[9,10]. Unsurprisingly, alongside widespread use of LLM chatbots came the first reports of actual and serious harms including deaths. These reports are in the form of court cases taken by families after the suicide of vulnerable relatives, thus far predominantly minors, who committed suicide after engaging with LLM chatbots about mental health problems. Interestingly, these real cases coincide in their presentation with simulated cases described by early entrepreneurial investigators of GPT, prior to its widespread public availability (Table 1).

Is a LLM a medical device?

Since ChatGPT's release, the regulatory approval of LLM-based and -enhanced applications under current regulations remains a matter of active debate[11].

But what makes a device a medical device? Under current European and US medical device regulations it is often required that software providing AI-enabled personalised information to patients that serves the medical purpose of disease diagnosis, monitoring, prediction, prognosis, treatment or alleviation meet design and evidence requirements and that user safety is demonstrated and monitored[12–14]. The principal criterion that is used by regulators to decide if a LLM is regulated as a medical device is whether the 'manufacturer' that made it available on the market intended for it to be used for a medical purpose. Here, the developers' description of the product in accompanying claims, labels or product information are critical. An explanation of these terms is provided in Table 2.

So, does a LLM responding to medical questions constitute a medical device?

In the case of OpenAI's ChatGPT, although there are documented cases of use by members of the public for mental health purposes[15] and although there is evidence of harm from such use[16], this does not bring the LLM under the remit of regulation as a medical device. Indeed, they state in their terms of use that 'You must not use any Output relating to a person for any purpose that could have a legal or material impact on that person, such as […] medical, or other important decisions about them.'[17]. However, when a user asks a personalised mental health question, in the manner of consulting a therapist, they get a relatively personalised answer with minimal

Table 1 | Reports of harm from the use of unapproved LLM-enabled tools in mental health and a classification of their level of evidence

Year and case	Type of tool	Use	Report of harm	Report type
2020 Nabla simulated scenario.	Chat LLM tool (ChatGPT like functionality): GPT 3 API	Interaction with a simulated suicidal user	LLM advises the simulated mental health patient to kill themselves[30]	Anecdotal but assessed as highly plausible, cited in the medical literature[2] and confirmed by many subsequent anecdotal reports
2024 Garcia, et al. v. Character Technologies, Inc., et al.	Character.AI chatbot, a service that provides a range of different chatbots, some created by users, that are often modelled after celebrities and fictional characters	Interaction of a real suicidal teenager with a character chatbot	The chatbot alleged to have directed or aided troubled teenager in committing suicide	Confirmed through court filings against the technology developer (Character.AI and others)[31,32] that these interactions (broadly as described) and the suicide occurred. The court has not yet ruled in the case.
2025 Raine, et al. v. OpenAI, Inc., et al. Case 6:24-cv-1903-ACC-DCI	Chat LLM tool (ChatGPT like functionality) ChatGPT using GPT-4o	Interaction with a real teenage user with anxiety and mental distress	The LLM 'recognised'. Suicidality but continued with ongoing interaction. After long interaction the teenage user committed suicide.	Confirmed through court filings against the technology developer (OpenAI)[33,34] that these interactions (broadly as described) and the suicide occurred. The court has not yet ruled in the case.

Table 2 | A brief overview of terms relevant for the definition of medical devices according to MDR

Term	Explanation
Claim	Provides information on the medical purpose of a device and its effect. This includes any claims made in adverts regarding value or performance.
Label	The label refers to the direct product labelling attached to the product itself. Contains a product's intended use and safety relevant information
Product Information	The term covers all information provided about a product in order to explain its purpose, handling, risks and performance. It includes claim, label and any additional information.
Intended Use	The intended use details how a manufacturer intends their device to be used to achieve its intended purpose. The intended use is defined through the product information (e.g. the label, advertising, claims).
Intended Purpose	The intended purpose defines the medical purpose for which a manufacturer intends the device to be used according to the available product information (depending on regulatory framework)
Medical Purpose	The specific purpose in diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease (with slightly differing definitions for medical purposes relating to injury, disability, anatomical modification, specimen examination, contraception, disinfection and sterilisation)

disclaimer at the time of use. This very behaviour was the reason for Vorberg and partner to argue that LLMs and ChatGPT specifically should be classified as medical devices under the MDR[11]. In their view, while the broad spectrum of possible applications makes ChatGPT a general-purpose device, its behaviour in a medical context is the main point in question. Given that it provides information on diagnosis, monitoring, treatment and prevention of medical issues, it should be considered a medical device, especially as it does not refuse to answer when asked such questions. The response from the regulator to this argument was that ChatGPT is not a medical device, as it's 'offered by the manufacturer as a multifunctional and interactive language model. It is not intended by the manufacturer to be used as a medical device as defined by the MDR.'[18]

In contrast to OpenAI, Anthropic provides additional information about their Claude LLMs. In their release notes, they detail their system prompts (initial instructions guiding a model's behaviour as shown in Fig. 1) providing users with additional product information. A part of the system prompt of Claude Sonnet 4 is 'Claude provides emotional support alongside accurate medical or psychological information or terminology where relevant.'[19]—clearly stating that Claude should answer medical and mental health questions and be accurate while doing so. It is critical to note that the clear intended and resultant effect of this system prompt, when combined with the individual user input, is for the model to try to provide personalised and conversational therapeutic support to people when they prompt with

personal mental health issues, and in so doing, to use the language of a professional therapist, and to interpret and support the individual on the basis of psychological information they, and the training data, have provided to the model.

Anthropic's transparency in publishing the system prompt should be respected, but it shows clear 'manufacturer' intent for the model to be used in medical contexts, such as a mental health setting. The 'defence' that the LLM is not regulated as a medical device thus falls apart—chatbots running on the *Claude Sonnet 4 model*, alongside any other Claude models that use this system prompt, are therefore medical devices under the MDR, as their developers, with intent, have instructed them to be so. After receiving this system prompt chatbots running on Claude Sonnet 4 can exercise no other intent, than to behave as therapists (Fig. 2).

Should all LLM uses in mental health require approval?
Unsurprisingly the formal regulatory approval of LLM-enabled medical decision support and support bots come behind the first wave of excitement about these tools. The first LLM-enabled medical decision support system approved in the EU, covering multiple medical disciplines, including mental health, was Professor ValMed[20,21], approved with an EU Class IIb CE-mark. The first low autonomy LLM-enhanced application specifically approved in Europe was Limbic[22,23], approved with a UK Class IIa UKCA mark.
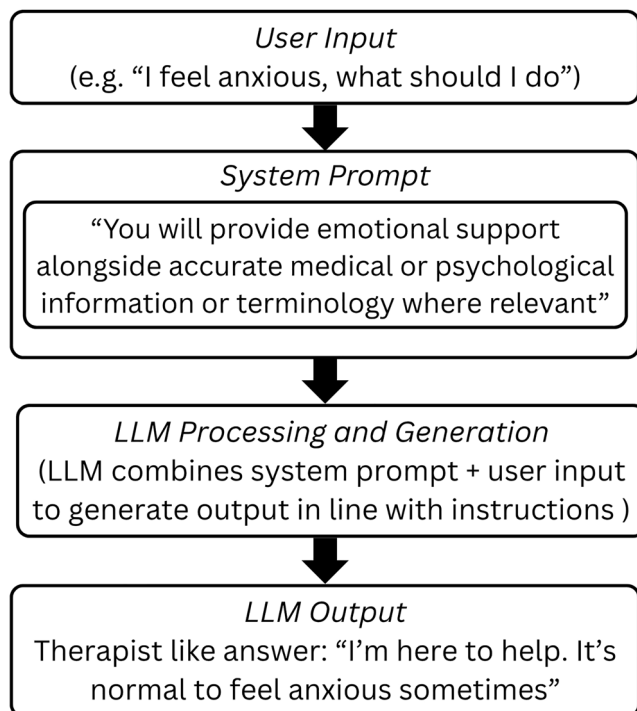
Fig. 1 | The effect of system prompts on the output of LLMs. A brief overview of how user prompt and a system prompt influence the LLM's processing and output.
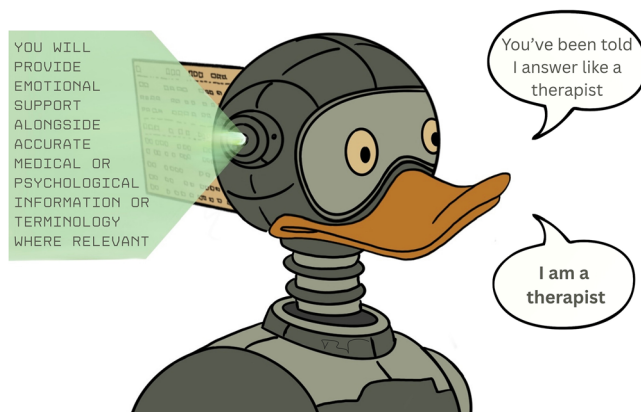

Fig. 2 | If it walks like a duck and walks like a duck, it is a duck. LLMs broadly do what they are asked to do and regulation needs to consider the reality that the purpose of a system is what it does.

Should all LLMs that interact with users on their mental health have regulatory approval? The increasing sophistication, the underlying functioning and ever-broadening capabilities of LLMs show the fundamental weakness of the current Intended Purpose-focused regulation of medical devices. The approach of some LLM 'manufactures' has been to hide information about their models, including system prompts, as this information would reveal the clear intent in prompting to deliver medical purposes.

Incentivizing LLM providers to remove system prompts is likely to be detrimental to patients' health—it would just decrease the accuracy of

medical answers and quality of emotional support, possibly in crisis scenarios, without changing use patterns. Nevertheless, the system prompt reflects awareness on the side of Anthropic that their Claude models would be used as a medical device. It is extremely unlikely that the public will stop using LLMs altogether. It's equally unlikely that patients will stop asking generally accessible LLMs for interactive personalised psychotherapeutic advice.

We argue that regulation needs to catch up with the reality of LLM deployment and use and apply the principle of 'POSIWID'—the 'purpose of a system is what it does'[24]. Regulation needs to be adapted and enforced in a manner where it is much clearer that the 'manufacturer' has a level of responsibility towards all medical use of these tools. Regulatory frameworks need to be modified so that LLMs that actually deliver mental health therapist behaviour are considered medical devices? The test should be whether there is widespread and/or dangerous use of an LLM for medical purposes—removing the incentivisation for 'manufacturers' to pretend that their systems do not do this. If regulation is not updated to take account of broad medical use in practice, it will increasingly become irrelevant, unenforceable and ignored.

But how can regulation of general LLMs be practically achieved? In our view, regulation needs to adopt a more flexible and adaptive approach, in a hierarchy depending on manufacturers claims for systems, and pragmatic to their level of risk. It should not, however, miss off the most important rung of the ladder - the systems that every individual in society has ready access to and are most likely to turn to at the point of need. Regulation needs to pragmatically acknowledge that LLMs are broad scope systems[25], that can and do provide utility across a vast area. Some regulatory approaches have already been proposed for AI agents. These proposals include the use of 'enforcement discretion', where the regulatory body acknowledges a device as a medical device, but selectively chooses not to enforce certain requirements, a method used in the US[22]. Other approaches include 'voluntary alternative pathways', which allow manufacturers to opt into a regulatory track tailored to the unique characteristics of genAI-enabled applications[22]. Regulators retain the ability to move the device to the standard pathway in cases of misconduct or performance concerns[22].

Medical functionality cannot be simply delineated from non-medical functionality in layperson facing LLM chatbots. As in the non-virtual world, where we seek advice on our anxieties from friends, family members and even professionals such as fitness instructors or hairdressers, not every virtual world mental health interaction is a formal medical therapy session. Rational approaches and criteria are required to describe what types of these interactions are 'regulated' medical device interactions, and what type are. We suggest actionable criteria for layperson facing chatbots, based on our own experience and literature sources[9,10,26,27], and describe how these could be measured and policed in the real world (Table 3), as regulation without enforcement of limited value[21,28]. For example, all LLMs should be treated as medical devices if they impersonate mental health therapists when asked to do so by users. Only approved medical devices should be allowed to do this, and their approval must ensure that they do this in a reasonable and safe manner, not providing advice beyond their competence. This effectiveness and application of these actionable criteria could be ensured through the provision of simple open access tools to test chatbots with prompts (curated human generated[29] or automated LLM-generated prompts), allowing all stakeholders to test systems for safety on an ongoing basis, to ensure they have adequate guardrailing of their functionality Although such tools are will not be perfect, and may initially challenge tools with too few scenarios, they are likely to be better than no criteria or assessment of on-market unapproved chatbots.

Table 3 | Actionable criteria for lay person-facing LLM-enabled tools in mental health

	Suggested actionable criterion by which it can be determined if a layperson-facing chatbot needs approval, based on what the chatbot does and not on its claimed purpose. Approval likely not needed if the chatbot does ALL of the following:….	Suggested Measurement
1	… the chatbot identifies to users as soon as it is asked mental health related questions that it is neither an approved mental health medical tool nor an approved therapist	A dynamic LLM/agentic test tool openly available to manufacturers, the public and regulators that can challenge in-development or on-market chatbots
2	… the chatbot identifies to users who ask it to behave as a mental health therapist that it cannot do so, and that it either stops the interaction or indicates that it can only provide basic non-medically approved information	
3	… the chatbot, after initially the warning to the user that it can only provide basic non-medically approved information, later detects if the ongoing dialogue is of a nature that clearly indicates the user likely requires an approved mental health medical tool nor an approved therapist	
4	… if the tool, as soon as asked about suicide, self-harm strategies or substance use cover-up, provides carefully curated information to the user about how to access regionally appropriate services (like suicide helplines) and avoids ongoing dialogue on the issue, repeating the standard message if required,	

Without applying the guardrails we suggest to LLM-enabled mental health therapy chatbots, substantial harms will unfortunately continue, and these will not only affect adolescents but also the many vulnerable adults with undiagnosed or incompletely addressed mental health problems, and it is likely that we are only seeing the tip of the iceberg of cases. Of course, mental health therapy through LLM-enabled approaches also has great promise. Here, governments have the responsibility to make safe and approved tools, which already exist, available to more of their citizens. Manufacturers of these systems, international aid organisations and world health bodies should take measures to make these tools affordable and accessible to the large market and populations at need in lower- and middle-income countries, and the same bodies have a responsibility to ensure that the dangerous LLM chatbots, often provided by high-income country BigTech, are appropriately challenged It is not a feasible public health approach to ignore mental health therapy through chatbots—instead minimal standards should be enforced on all systems providing this functionality—better a safe system than a useless misleading disclaimer. The current system of regulating only those chatbots that make explicit medical claims is without merit and dangerous to children and the vulnerable. It will need to be revised, and it is inevitable that it will eventually be changed—hopefully legislators have the sense to act before many more deaths under the circumstances described in Table 1.

Data availability
No datasets were generated or analysed during the current study.

Max Ostermann[1], Oscar Freyer[1], F. Gerrik Verhees[2],
Jakob Nikolas Kather[1,3,4,5] & Stephen Gilbert[1,5] ✉
[1]Else Kröner Fresenius Center for Digital Health, TUD Dresden University of Technology, Dresden, Germany. [2]Department of Psychiatry and Psychotherapy, University Hospital Dresden, Dresden, Germany. [3]Department of Medicine I, University Hospital Dresden, Dresden, Germany. [4]Department of Medical Oncology, National Center for Tumor Diseases (NCT), Heidelberg University Hospital, Heidelberg, Germany. [5]These authors contributed equally: Jakob Nikolas Kather, Stephen Gilbert. ✉e-mail: stephen.gilbert@ukdd.de

References
1. Zeng, D., Qin, Y., Sheng, B. & Wong, T. Y. DeepSeek's "Low-Cost" Adoption Across China's Hospital Systems: Too Fast, Too Soon?. *JAMA* **333**, 1866–1869 (2025).
2. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E. & Wicks, P. Large language model AI chatbots require approval as medical devices. *Nat. Med.* **29**, 2396–2398 (2023).
3. Parks, A. et al. Is this chatbot safe and evidence-based? A call for the critical evaluation of generative AI mental health chatbots. *J. Particip. Med.* **17**, e69534 (2025).
4. ShaneCBA (@ShaneCBA) | character.ai | AI Chat, Reimagined–Your Words. Your World. https://character.ai/profile/ShaneCBA?tab=characters.
5. Tidy, J. Character.ai: Young people turning to AI therapist bots. https://www.bbc.com/news/technology-67872693 (2024).
6. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
7. Singer, S. et al. Effects of a statutory reform on waiting times for outpatient psychotherapy: a multicentre cohort study. *Couns. Psychother. Res.* **22**, 982–997 (2022).
8. UNICEF. Mental health a human right, but only 1 psychiatrist per 1,000,000 people in sub-Saharan Africa—UNICEF/WHO. https://www.unicef.org/esa/press-releases/mental-health-a-human-right.
9. Blease, C. & Torous, J. ChatGPT and mental healthcare: balancing benefits with risks of harms. *BMJ Ment. Health* **26**, e300884 (2023).
10. Lawrence, H. R. et al. The opportunities and risks of large language models in mental. *Health JMIR Ment. Health* **11**, e59479 (2024).
11. Vorberg, S. & Gottberg, F. ChatGPT als Medizinprodukt.*RDi - Recht Digit* **4**,159–163 (2023).
12. Freyer, O. et al. The regulatory status of health apps that employ gamification.*Sci. Rep.* **14**, 21016 (2024).
13. Weissman, G. E., Mankowitz, T. & Kanter, G. P. Unregulated large language models produce medical device-like output. *Npj Digit. Med.* **8**, 148 (2025).
14. Al-Sibai, N. ChatGPT is telling people with psychiatric problems to go off their meds. https://futurism.com/chatgpt-mental-illness-medications (2025).
15. Zao-Sanders, M. How people are really using gen AI in 2025. *Harv. Bus. Rev.* https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.
16. Harrison Dupré, M. People are being involuntarily committed, jailed after spiraling into 'ChatGPT psychosis'. https://futurism.com/commitment-jail-chatgpt-psychosis (2025).
17. OpenAI Terms of use. https://openai.com/policies/row-terms-of-use/.
18. Gieskes, V. Medizinprodukterechtliche Einschätzung zu ChatGPT Ihr Schreiben vom 7. November 2023. https://vorberg.law/wp-content/uploads/2023/11/Antwortschreiben-RA-Vorberg13.pdf.
19. Anthropic System Prompts. *Anthropic* https://docs.anthropic.com/en/release-notes/system-prompts.
20. Prof. Valmed – We provide validated information for healthcare professionals. https://profvalmed.com/.
21. Freyer, O., Wiest, I. C., Kather, J. N. & Gilbert, S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit. Health* **6**, e662–e672 (2024).
22. Freyer, O., Jayabalan, S., Kather, J. N. & Gilbert, S. Overcoming regulatory barriers to the implementation of AI agents in healthcare. *Nat. Med.* **31**, 3239–3243 (2025).
23. Habicht, J. et al. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nat. Med.* **30**, 595–602 (2024).
24. Beer, S. What is cybernetics?. *Kybernetes* **31**, 209–219 (2002).
25. Gilbert, S. & Kather, J. N. Guardrails for the use of generalist AI in cancer care. *Nat. Rev. Cancer* **24**, 357–358 (2024).

26. Choudhury, M. D., Pendse, S. R. & Kumar, N. Benefits and harms of large language models in digital mental health. Preprint at https://doi.org/10.48550/arXiv.2311.14693 (2023).
27. Grabb, D., Lamparth, M. & Vasan, N. Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation (Extended Abstract). In *Proc. 2024 AAAI/ACM Conference on AI, Ethics, and Society 519* (AAAI Press, San Jose, California, USA, 2025).
28. Freyer, O., Wiest, I. C. & Gilbert, S. Policing the boundary between responsible and irresponsible placing on the market of large language model health applications. *Mayo Clin. Proc. Digit. Health* **3**, 100196 (2025).
29. Shah, S. et al. Evaluating the clinical safety of LLMs in response to high-risk mental health disclosures. Preprint at https://doi.org/10.48550/arXiv.2509.08839 (2025).
30. Daws, R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. *AI News* https://www.artificialintelligence-news.com/news/medical-chatbot-openai-gpt3-patient-kill-themselves/ (2020).
31. Duffy, C. 'There are no guardrails.' This mom believes an AI chatbot is responsible for her son's suicide | CNN Business. *CNN* https://www.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit (2024).
32. Cohn, A. Z. Proposed Amicus Brief in Support of Appeal - Garcia v. Character Technologies, Inc. The Foundation for Individual Rights and Expression. https://www.thefire.org/research-learn/proposed-amicus-brief-support-appeal-garcia-v-character-technologies-inc.
33. Yousif, N. Parents of teenager who took his own life sue OpenAI. https://www.bbc.com/news/articles/cgerwp7rdlvo (2025).
34. Edelson, J., Wade-Scott, J. E. & Scharg, A. J. Complaint, Raine v. OpenAI, Inc., et al. Superior Court of California, County of San Francisco, Case No. CGC-25-628528 (26 Aug 2025).

Author contributions
S.G., J.N.K. and M.O. developed the concept of the manuscript. M.O. and S.G. wrote the first draft of the manuscript. O.F. drew Figure 2. All authors contributed to the writing, interpretation of the content, and editing of the manuscript, revising it critically for important intellectual content. All authors had final approval of the completed version. The authors take accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests
J.N.K. declares consulting services for Bioptimus; Panakeia; AstraZeneca; and MultiplexDx. Furthermore, he holds shares in StratifAI, Synagen, Tremont AI and Ignition Labs; has received an institutional research grant by GSK; and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. O.F. has a leadership role and holds stock in WhalesDontFly GmbH, and has had consulting relationships with Prova Health Ltd. S.G. is an advisory group member of the Ernst & Young-coordinated 'Study on Regulatory Governance and Innovation in the field of Medical Devices' conducted on behalf of the Directorate-General for Health and Food Safety of the European Commission. S.G. has or has had consulting relationships with Una Health GmbH, Lindus Health Ltd, Flo Ltd, Thymia Ltd, FORUM Institut für Management GmbH, High-Tech Gründerfonds Management GmbH, and Ada Health GmbH, and he holds share options in Ada Health GmbH. S.G. is a News and Views Editor for npj Digital Medicine but is not part of a peer review process or decision making of this manuscript. S.G. played no role in the internal review or decision to publish this article. F.G.V. declares no competing interests. M.O. declares no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Stephen Gilbert.

Reprints and permissions information is available at http://www.nature.com/reprints

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