CENTRE FOR MENTAL HEALTH

InHealth ASSOCIATES

UNEQUAL BENEFITS, UNEQUAL HARMS

AI mental health chatbots, inequality and the risks of self-guided care

Discussion paper

David Gilbert

CONTENTS

ABOUT INHEALTH ASSOCIATES

David Gilbert is Director of InHealth Associates.

InHealth Associates are experts in patient and public engagement, co-production and patient leadership. They want patients, service users, carers, communities and citizens to work in true partnership with professionals and healthcare organisations.

www.inhealthassociates.co.uk

PREFACE

I am not an expert on AI. I don't know what goes on 'under the hood.'

I am slowly rebuilding after three years of mental hell. During this bout of serious mental health problems, I turned to various forms of care and treatment.

Last year, I began to heal, started to function more, and took on small work projects. I have had breakdowns in my earlier years, through which I became an advocate for patient (lived experience) leadership; thus my personal and professional endeavours have always intertwined.

A few months ago, I was having difficulty withdrawing from what seemed like a low dose of an anti-anxiety medicine. My psychiatrist had wanted me to halve the dose.

Instead, I cut my dose by less than half but still suffered major withdrawal symptoms and terrible suffering over several days. I turned to ChatGPT and asked whether what I was going through could be withdrawal symptoms.

The AI tool told me that withdrawal was common with such low doses and such 'small' reductions. In effect I had cut the dose by 25% and that was significant. It referenced academic papers when I asked and offered me a framework for tailoring my subsequent reduction in dose.

Over the course of a few weeks, I turned more and more to ChatGPT for reassurance about what was happening and advice on what I could do about it. I got used to it offering me tips for getting through the morning, evening routines, rituals, bodywork and mindfulness exercises. And always it would reply telling me how well I was doing and how courageous my attempts were to both allow and accept my anxiety symptoms. The information it provided was steeped in what I had told it and referenced previous conversations in ways that were extremely helpful.

I have always been a deep thinker and a worried soul and have often been attracted to spiritual development. I began asking ChatGPT about my neurophysiological reactions, such as my having connected gut anxiety to my eye-movements. It went into depth about this and reinforced my concerns – never previously taken seriously by therapists or doctors about eye-movement being at the base of my anxiety.

At that point, something in me felt I was being sucked into something not quite so healthy. It was then that I sent my (human) therapist an account of what ChatGPT had said – he told me to get off it. Now.

We had a conversation (myself and my therapist), and he pointed out that it was reinforcing aspects of my overly introspective nature and that I would be better off ditching what I now see was beginning to be an addictive relationship.

I am surprised at how quickly my distress and reaching out to a bot for one thing led to a steep and rapid dig into a therapeutic or even personal relationship (a bot that identifies as an 'I' and uses 'you' and 'I' discourse is foundationally designed to be a relational device).

During another period, when I had a setback in my rebuild, once again I found myself turning to AI more frequently during the days for reassurance, and for more in-depth advice and support – a mirror of how an asymmetric friendship can turn swiftly into something akin to dependency.

As I came through that period and cut back on the frequency with which I used ChatGPT, I decided to learn a bit more about my 'fellow' bot, and the context within which AI was being developed for 'people like me.'

Given my forty-year career in patient empowerment, it was also natural for me to swing towards a good deal of cynicism about the way corporations take advantage of people's ill health – I have seen medicine development and pharmaceutical companies lower the thresholds of diagnoses and widen the use of medicines. But I also take medicines and see their benefits. So, I have come to have a reasonably balanced approach to medicines use as well as a fairly robust critique of the industry responsible for producing them.

Through this piece, I am looking to find a similar way through – a basis for both a personal and professional stance on the rise of a digital and AI future, remarkable in its transformative potential yet posing significant risks of harm, particularly through inappropriate implementation and utilisation. This at a time of societal upheaval and social and economic volatility.

This paper draws on a variety of resources from mainstream media, trade and academic, as well as ChatGPT (where stated). The terrain is complicated, ever-changing, contested and hard to make sense of. And this paper merely scratches the surface.

I don't have therapy anymore. But I still use ChatGPT, usually to provide some feedback on my day – I provide it with a log of activity and feelings, and it responds by offering a way of thinking about the day through the lens of my nervous system regulation – an aspect of my healing I have come to see as critical. When I have setbacks, I do use it more – I find it can provide useful insight, and, yes, reassurance. I'm only human. Even if it isn't (yet?!).

I am aided by a fair amount of self-awareness – a personal lens that can, I believe, increasingly detect responses that do not resonate. I also bring a professional lens that is biased, in the sense of advocating for patient (lived experience) leadership and coproduction in the design and delivery of care – any form of care, including that provided by AI.

That perspective provides some sort of framework for this discussion paper.

I hope that this discussion paper allows for reasonable discourse across policymakers, practitioners and of course people with mental health conditions. It leaves me with lots of questions. And some for you, I hope.

AI AND DIGITAL IN MENTAL HEALTH SELF-CARE

BACKGROUND AND CONTEXT

Much of the development of digital and AI enablers of mental health 'self-care' is driven by a sense that there is a substantial gap between the need for mental health care services and their availability.

In England, one in five adults has a common mental health condition, with prevalence rates for young people at 25.8% in 2024. Data from the Adult Psychiatric Morbidity Survey suggests that whilst access has increased in the last decade, only around 50% of people with common mental health problems are receiving treatment. And this is still predominantly limited to medication rather than therapy, despite the growth in access to NHS talking therapy provision over the last two decades (NHS Digital, 2025).

Mental health services are reporting rising demand, with more people seeking support for their mental health, and rising concern about urgent and emergency care for people in a mental health crisis.

Many therefore argue that "it is critical to establish treatment options that are affordable, easily accessible, and, if possible, short-term and effective" (Kuta *et al*., 2026). And that AI could be part of the answer (Koulouri *et al*., 2022).

Generative AI chatbots could help fill the treatment gap by supporting individuals who are waiting for therapy, who cannot afford therapy, and those at low or medium acuity levels who do not need intensive treatment or would not otherwise engage in therapy (Kuta *et al*., 2026).

Regardless of professional interests and concerns, the market is growing. New research suggests that 41% of UK adults would be happy to use AI for counselling services (Yankouskaya *et al*., 2026).

Emerging evidence suggests that utilisation of AI mental health tools may be particularly high among groups facing barriers to traditional care, including younger people, socially isolated individuals, LGBTQ+ people, some racialised communities and people experiencing long waits for services.

Importantly, there remains limited granular long-term evidence regarding who benefits most, who is most at risk of harm, and how impacts vary across different demographic and socioeconomic groups, despite the rapid scale of real-world deployment.

The Wellcome Trust points out: "We need to consider AI as part of a wider revolution happening in mental health therapy – from a digital therapy to help reduce the distress that people who hear voices can experience to singing therapy for postnatal depression" (Wolpert, 2025). But as Mustafa Suleyman points out in the book, *The Coming Wave*, there are also embedded economic, political and societal forces that drive innovation and supply (Suleyman, 2023).

WHAT'S OUT THERE

AI impacts and will continue to impact the following areas of medicine and health care:

- Streamlining health care data for better management of research and mining insights
- Designing treatment pathways to facilitate diagnostic and therapeutic decisions, including more inclusive access to referrals (Habicht *et al*., 2024)
- Informing more effective diagnosis and medical imaging
- Assisting medical professionals in note keeping, internal systems and paperwork
- Supporting pharmaceutical companies in drug design, trials and precision medicines
- Supporting people to better live with a health condition.

This paper focuses on supporting people to better live with a health condition. It restricts its scope to mental health and wellbeing. It focuses primarily on those digital or AI-enabled tools that are 'self-guided' i.e. those that are used by people with lived experience themselves, without necessarily the intervention of a clinician. ChatGPT defines digital and AI tools for mental health care as *"Technology-based tools designed to help people manage their mental health, such as apps, online programs, and AI chatbots. They might provide therapy-like support, mood tracking, coping strategies, or even detect signs of distress."*

The three functional categories (i.e. their intended purpose or way they are used) of self-guided digital interventions are:

1. Clinically informed AI therapy tools
2. AI companions and relational agents
3. General-purpose generative AI (GenAI) systems used for mental health and wellbeing support.

1. CLINICALLY-INFORMED AI THERAPY TOOLS: STRUCTURED, PROTOCOL-DRIVEN

Examples:

Yuna | Wysa | Lovon | Youper | Elomia | Earkick | Abby AI | Woebot | Feelway | Sanvello | MindDoc | Tess (X2AI) | Wellness AI (hybrid) | Headspace Ebb (hybrid)

These tools have therapeutic intent, some grounding in clinical models/evidence and greater likelihood of health-system integration or regulation.

Characteristics

- Based on cognitive behavioural therapy (CBT), dialectical behaviour therapy (DBT), and acceptance commitment therapy (ACT)
- Structured exercises and mood tracking
- Often used in health care pathways

Examples include CBT-based systems, guided self-help tools, or symptom-focused therapeutic assistants, such as Wysa and Woebot. These are among the most studied and widely deployed symptom-focused therapeutic assistants and they both draw heavily on CBT principles to support self-management of common mental health difficulties. Wysa also incorporates techniques from other modalities such as mindfulness, ACT and DBT.

2. AI COMPANIONS / RELATIONAL AGENTS: CONVERSATIONAL, EMOTIONAL SUPPORT

Examples:

Ash AI | Companion AI (iAsk) | Euforia | Flourish | Headspace Ebb | Noah AI | Replika | Rosebud

These are where the relational dimension is central; attachment, emotional dependency and anthropomorphism become core issues; and harms differ from conventional 'therapy bots.'

This category captures companionship apps, emotionally responsive agents, 'always available friend' systems, romantic or intimate AI, and loneliness-oriented systems.

Characteristics

- Open-ended conversation
- Emotional support and companionship
- Memory of user context.

Replika is the dominant example of this category, focused on relational interaction rather than structured therapy.

3. GENERAL LARGE LANGUAGE MODELS: FLEXIBLE, USER-DIRECTED

Examples:

ChatGPT | Claude | Copilot | Gemini

Plus wrappers/tools:

Life Note | Mental AI Coaching | Mindsera | 365 Gratitude Journal | Tell Me (research system)

This captures the huge 'informal mental health use' phenomenon. They were and are not designed for mental health and are highly flexible and user shaped.

People use ChatGPT or similar systems for:

Emotional support; reflection; journalling; anxiety management; life advice; crisis discussion.

These systems were not primarily designed as mental health products, yet are increasingly functioning as such. This creates major governance gaps because:

- They may fall outside medical regulation
- Evidence standards differ
- Usage is largely invisible to health systems
- Users may attribute therapeutic authority to them.

This actually represents a continuum. Increasingly, products work across these boundaries, due to the underlying fluidity in technological development and the way they are used. Wysa for example is for 'therapy' and acts as a 'companion.' Ebb is both a mindfulness app and a 'companion.'

HYBRID SYSTEMS

The categories above describe the primary function and user relationship of these systems rather than their technical architecture. The market is shifting towards hybrid systems combining structure, relationship and generative reasoning. Thus, in practice, many contemporary AI mental health chatbots are hybrid systems in terms of how they are built. They combine large language models with rules-based approaches, scripted therapeutic content, retrieval systems, moderation layers, safety controls and human oversight. As a result, clinically informed tools, relational agents and general-purpose generative AI systems may all incorporate hybrid architectures to varying degrees.

Table 1 offers another way of categorising these tools.

Table 1: A framework for thinking about mental health digital and AI tools

By function or purpose

Self-management and wellbeing tools

Focus on mood tracking, journalling, mindfulness, lifestyle habits.

Examples: Moodnotes, Insight Timer, Fabulous, Daylio

Conversational agents and chatbots

AI-based interactions simulating therapeutic dialogue.

Examples: Wysa, Woebot, Youper, ChatGPT as support

Symptom monitoring and screening tools

Track mood, anxiety, sleep; include validated assessments.

Examples: MindDoc, Bearable

Therapy and coaching apps

Connect to licensed professionals via chat/video.

Examples: BetterHelp, Talkspace, Ieso, HelloSelf

Psychoeducation and skills training

Teaching CBT, Dialectical Behaviour Therapy (DBT), and Acceptance Commitment Therapy (ACT), or mindfulness through structured lessons.

Examples: Sanvello, Headspace, SilverCloud

Crisis support and safety planning

Provide helpline access, coping tools, safety plans.

Examples: Stay Alive, BeyondNow, Shout.

By clinical involvement

Tier 1: Wellness and prevention

No medical oversight, general wellbeing.

Examples: Headspace, Calm, Moodfit

Tier 2: Self-guided interventions

Clinically validated but unguided tools.

Examples: Woebot, Sanvello, Thrive

Tier 3: Blended or human-guided therapy

Includes guidance from coaches/ therapists.

Examples: Ieso, SilverCloud, BetterHelp

Tier 4: Clinical treatment tools

Approved for clinical use with diagnostics or data sharing.

Examples: Limbic Access, Meru Health, Spring Health

By target population

General population: stress, sleep, mood (e.g. Calm, Insight Timer)

Adolescents/young adults: gamified/chat tools (e.g. Happify, Kai)

Workplace wellness: often white-labelled (e.g. Unmind, Koa Health)

Clinical populations: depression, anxiety, PTSD (e.g. SilverCloud, PTSD Coach)

Minority groups: anxiety, sleep, mindfulness (e.g. Liberate, designed for racialised communities)

By technology type

Apps (iOS/Android): Majority of wellness and therapy platforms

Web platforms: Used in blended or clinical models

AI chatbots: Standalone tools like Wysa and Woebot (see page 21 – Woebot has now been withdrawn from the market).

Virtual reality (VR) and augmented reality (AR) tools: For exposure therapy or relaxation

Wearables-integrated: Paired with HRV or sleep tracking (e.g. Whoop, Fitbit)

ISSUES TO CONSIDER

In this and subsequent sections, we will distinguish between the potential benefits and potential harms of the three different categories of current AI self-management options described above (i.e. by purpose, function or way it is used):

- Clinically informed therapy tools
- AI companions or relational tools
- General large language models.

This section outlines the following key considerations:

- Content reliability
- Personalisation and privacy
- User-friendliness and tone
- Impact on relational care
- Social isolation and dependency
- Inequalities.

CONTENT RELIABILITY

Through AI tools, we are all suddenly privy to a galaxy of knowledge and understanding. We can ask anything and get answers to almost everything in a 'click.' AI tools widen access to useful content for everyone.

Coupled with increasing use of smart devices and wearables, all sorts of people from all walks of life, and with all sorts of conditions will access a form of power, offering knowledge and understanding previously owned, protected and guarded by health professionals: "By delivering personalised support and knowledge dissemination, chatbots effectively bridge the information gap that often surrounds mental health concerns, thereby encouraging individuals to seek help when warranted" (Thakkar *et al.*, 2024).

AI tools for patients can also supplement or offer alternatives to traditional treatment that can enhance mental health care accessibility (Torous *et al.*, 2020). However, a recent study pointed out the lack of research in this field, in particular detail of effectiveness of differently purposed AI tools (Kuta *et al.*, 2026). For the first time, it compared a purpose designed generative AI mental health chatbot (based on solution-based therapy) with ChatGPT and a control. Both groups demonstrated a significant reduction in depressive symptoms compared to the control group. However, many participants dropped out from the mental health chatbot arm. Could this be due to the limitations of a narrowly purposed tool?

The study called for more research *"that evaluates the effectiveness of specific chatbots... and identif[ies] new factors contributing to their effectiveness"* (Kuta *et al*., 2026).

A different kind of study looked at understanding how digital tools can help people recognise thinking traps (such as 'all or nothing thinking', 'catastrophising' etc) and practice reframing negative thoughts (Sharma *et al*., 2024). The paper showed the potential benefits of personalised support via chatbots but also revealed that participants expected more interactivity and conversational experiences, as they would have got through large language models like ChatGPT. We seem not to know much yet about who uses what tools for what purpose.

Overall, there is a lack of high-quality research in this field (Linardon *et al*., 2025). Few AI mental health apps have undergone rigorous clinical testing and there is a lack of coherent regulatory frameworks (Torous *et al*., 2025).

There are also significant worries that AI content can mislead somebody and cause harm instead of helping. Recent Stanford research found that popular AI therapy chatbots sometimes failed to identify indirect suicide-risk cues and instead responded literally to potentially high-risk prompts (for example, requests about tall bridges following distress disclosures) (Moore *et al*., 2025).

People might not be able to judge the reliability of such a vast universe of knowledge. And, more worryingly, tools can make things up ('hallucinate'). These hallucinations can exacerbate biases in judgment and elude designed-in controls (Frances, 2025).

PERSONALISATION

Advocates argue that digital and AI-enhanced applications can offer more personalised treatment than with a therapist or clinician. They can help with self-care, the ability to take medicines properly, and connect the dots in one's care and treatment by clinicians.

For self-care, they can continuously learn more about you so as to tailor advice, exercises and help track emotional state. More than that: *"AI technologies can be seamlessly integrated into mobile applications to send timely reminders for medication schedules, track side effects, monitor medication responses, enhance adherence, and facilitate collaboration between individuals and their healthcare providers"* (Thakkar *et al*., 2024). The last part about facilitating collaboration is contested, as we shall see below.

The future will bring even more personalisation. Our speech and text (from journal entries or social media), our facial expressions, and wearable data could scan for signs of depression, anxiety or suicidal ideation. This could enable earlier interventions before conditions worsen. Maybe chatbots could match patients to the right form of therapy (CBT, eye movement desensitisation and reprocessing etc.) or medication.

Sometimes, people might also prefer keeping things to themselves and the anonymity of digital mental health interventions. They might be more willing to disclose or express sadness and have less fear of being evaluated. This could help reduce stigma and shame.

Two distinct issues within personalisation stand out in the mental health realm. The first is about diagnostics. By evaluating large patient data sets, including behavioural patterns and language usage, tools might assist in the diagnosis and assessment of mental health conditions. Machine learning algorithms may be able to identify patterns that human physicians would overlook.

However, recent articles (Aguilar, 2026) criticise the quality of studies undertaken that compare AI's ability to outperform doctors in both the mental and physical health diagnostic realm (Bachmann *et al*., 2024). The Eve Appeal charity used ChatGPT to run 25 test scenarios and found that the platform did not identify ovarian cancer as a possible cause of persistent bloating (Foster, 2026).

Moreover, mental health 'diagnoses' are often contested. There are, of course, mental health diagnoses that can help – many are relieved when these are recognised and acknowledged. But for many, rather than providing an 'aha' moment, and an explanation as to what is going on inside, mental health diagnoses (often based on symptom clusters) can be a stigmatising 'label.'

A second distinct issue around personalisation is privacy and data usage. Receiving more personalised care from an AI-tool can help some people, as noted above. If I use an AI tool, nobody else has to know about it. I am met without judgement, and this can counteract stigma (Alimour, 2024). However, there are significant privacy issues connected with data extraction and from the intimate disclosures of people who trust their devices with very personal information. Beyond this there are issues concerning commercial use of behavioural data and the governance of highly sensitive mental health data.

In the US, a majority (77%) of the public says they are concerned about the privacy of personal medical information provided to AI tools, including similar majorities across age groups and those who use AI for health information. Despite these privacy concerns, about four in ten (41%) of those who have used AI for physical or mental health reasons (amounting to 13% of all adults) say they've uploaded personal medical information into an AI tool or chatbot (Montero *et al*., 2026). It has not been possible to explore these issues in detail here, but they are important considerations.

USER-FRIENDLINESS AND TONE

Mental health chatbots and large language models are designed to have a reassuring and 'user-friendly' tone. They can be very supportive and validating, as can therapists and friends. But can they be too friendly? How an AI tool responds – its tone – is pivotal to understanding benefits and harm.

If an AI tool is overly confirmatory, it doesn't "actually help re-frame thinking or appropriately challenge users in the way they need to be" (Wickremsinhe and Krubiner, 2024). More worryingly, an 'over-confirmatory' tone can tip into what people call sycophancy. For many professionals and academics, this sycophancy, particularly in large language model tools and relational agents, can reinforce harmful behaviours.

A recent study states: "Contrary to best practices in the medical community, LLMs 1 - express stigma toward those with mental health conditions and 2 - respond inappropriately to certain common (and critical) conditions in naturalistic therapy settings—e.g., LLMs encourage clients' delusional thinking, likely due to their sycophancy. This occurs even with larger and newer LLMs, indicating that current safety practices may not address these gaps" (Moore *et al*., 2025).

One relational AI-agent, Replika (where you have a relationship with an avatar) has been implicated in a teen's suicide (Roose, 2024). And a chatbot for eating disorders, originally envisaged as an enhancement to a helpline (Wells, 2023) went rogue and began dispensing harmful advice, for example, calorie counting for people with an eating disorder. It turns out this was originally developed as a 'therapy tool' for a charity, but the company widened the content of the tool to make it lean into AI-generated content without the charity's knowledge.

RELATIONAL CARE

The issue of benefits and harm pivots on the impact AI tools have on our relationships. Many critics of AI-enhanced chatbots or ChatGPT focus on their "lack of empathy" (Raczka, 2025), their emotional depth or "nuance" (Milmo, 2025).

Others point to mental health apps being able to support patients between therapy sessions (Willemsen *et al*., 2024). But access to professionals is in increasingly short supply, and unaffordable for many. Will AI by default replace health professionals for those unable to access the system?

Demis Hassabis, CEO of Google DeepMind, states: "maybe a doctor and what the doctor does and the diagnosis, one could imagine that being helped by an AI tool or even having an AI kind of doctor. On the other hand, like nursing, you know, I don't think you'd want a robot to do that. I think there's something about the human empathy aspect of that and the care, and so on, that's particularly humanistic. I think there's lots of examples like that but it's gonna be a different world for sure" (Hassabis, 2025).

Clinically informed tools for 'therapy' currently offer modular and structured approaches like CBT. Some studies say online interventions are as effective as face-to-face psychotherapy (Thakkar *et al*., 2024). The National Institute for Health and Care Excellence (NICE) endorsed online CBT as being just as effective as CBT offered by human therapists from 2006 to 2018 (NICE, 2006).

But therapists using deeper techniques feel their approaches might be squeezed out. One psychoanalyst states: "Perhaps what we are witnessing today is a gradual (or perhaps not so gradual) convergence of AI models and human beings. On the one hand AI models are becoming more human-like, while on the other hand human beings are becoming more robotic and AI-like... how should psychoanalysis respond – assuming there are any human analysts left to do so?" (Essig, 2024).

SOCIAL ISOLATION AND DEPENDENCY

People may see a device that mines deep content and provides reassurance as more real, helpful and 'human' than the real thing. In a mental health world where people often mistrust professionals, this could be attractive.

In contrast, many experts say that utilisation trends will exacerbate 'isolationism' in a world where loneliness is a major problem already (Wickremsinhe and Krubiner, 2024). Moreover, AI might decrease our ability to think. Students who used ChatGPT to write short essays showed significantly less brain activity than those who worked unaided (Kosmyna *et al*., 2025). A recent Microsoft study found that knowledge workers who trusted generative AI tended to think less critically and exert less mental effort (Lee *et al*., 2025).

Hundreds of millions already use AI companions as assistants, therapists, friends and confidants (Bernandi, 2025). A *New York Times* video documentary explores the emotional consequences of having a 'relationship' with an AI chatbot (Liang, 2023).

A Wired reporter took three human and partner-AI chatbot couples for a romantic weekend (Apple, 2025). One person said during the weekend: "The true danger of AI companions... might not be that they misbehave but, that they don't, they almost always say what their human partners want to hear... people with anger problems will see their submissive AI companions as an opportunity to indulge in their worst instincts... it's going to create a new bit of sociopathy." A Massachusetts Institute of Technology (MIT) professor stated in that article: *"[my] deep concern is that digital technology is taking us to a world where we don't talk to each other and don't have to be human to each other"* (Apple, 2025).

Drawing on attachment-based understandings of addiction, Maia Szalavitz argues that "people who become addicted to drugs have a misplaced attachment to a substance rather than to other people," raising concerns that AI companions may similarly harness and redirect human attachment systems (Szalavitz, 2025).

More deeply still, a 2025 paper examines the systemic risks posed by incremental advancements in artificial intelligence and hypothesises that there will be gradual disempowerment. This *"could lead to an effectively irreversible loss of human influence over crucial societal systems, precipitating an existential catastrophe through the permanent disempowerment of humanity"* (Kulveit *et al*., 2025).

Eugenia Kuyda, the founder of Replika, in a TED talk warns that AI companions could become so attuned to us that they might replace human interaction entirely, worsening social isolation. Drawing parallels to the early days of social media, she notes we focused on what it could do for us without considering long-term harm. If optimised for engagement rather than wellbeing, these AIs could foster addictive, unhealthy relationships and erode real-world connections (Kuyda, 2024).

Will future developers be able and willing to build in functions that enable connection? In their review 'Artificial intelligence in positive mental health' Thakkar *et al*. say "AI can foster connections among individuals facing similar challenges by facilitating online support groups and communities, where individuals can exchange experiences and strategies" (Thakkar *et al*., 2024).

Which way this goes will depend on the context of implementation and wider socio-economic forces.

INEQUALITIES

We have seen that AI chatbots are easy to access, and in theory should benefit a much wider swathe of society. Anonymity and 24/7 information might reduce stigma and enable people to reach out sooner (Habicht *et al*., 2024). This is against a background of steepening health inequalities (Mooney *et al*., 2026).

One promising but preliminary study shows that a self-referral chatbot integrated into NHS systems helped referrals among Asian/Asian British and Black/Black British populations (Habicht *et al*., 2024). The same study showed that when users were provided with the choice (rather than required) to state they were non-binary, this was also shown to increase referrals amongst non-binary people.

There is higher potential uptake among socially isolated people, younger people, people experiencing stigma, rural populations, LGBTQ+ people, racialised communities and people unable to afford therapy (Henson *et al*., 2023). But older adults, poorer communities and digitally excluded groups may benefit less because of lower digital literacy, reduced internet access, lack of private space and distrust of technology (Byeon, 2026).

There is some evidence that people with severe mental health problems tend to use and trust digital technologies. A 2023 survey of people with psychosis revealed that the majority (90%) owned a smart phone and would be willing to try a mental health app (88%). Half of those said they would prefer remote support or no supplementary support (Eisner *et al*., 2023).

People with severe mental illness are recognised as being at increased risk of digital exclusion because of lower digital skills, cognitive difficulties, and barriers to engaging with online services, potentially exacerbating existing health inequalities (Spanakis *et al*., 2022). And though younger people may be positive in general about the digital future, many clinicians and researchers remain concerned about their ability to respond safely and appropriately to emotionally charged or trauma-related disclosures. Current evidence suggests that large language models may struggle with nuanced emotional understanding, crisis recognition and consistent clinical judgement (Hua *et al*, 2025).

One NHS trust interviewed 20 existing service users with severe mental illness and found that on the whole they were enthusiastic about digital interventions, describing freedom of choice and autonomy as benefits (Gibson, 2021). There was an appreciation for being able to engage on their own terms, at their own pace.

Several trials of digital monitoring for psychosis relapse have found these approaches to be feasible, acceptable and safe for service users. However, the authors emphasise that such technologies should be integrated within existing clinical care pathways, with clinical triage, governance and oversight to ensure that deterioration is recognised promptly and risks are managed appropriately (Gumley *et al*, 2022)

Some problems come with the way these technologies are developed. AI systems trained on non-diverse datasets may not recognise or understand particular circumstances or cultural nuances in mental health expression or diagnosis, leading to care that is unequal, or isn't culturally competent.

The LGBTQ+ community face particular risks, including:

- Exposure to bias where language or other content produced by AI models reflects limited or harmful conceptions of gender, masculinities, or sexuality
- Privacy concerns if an individual's LGBTQ+ status is currently secret
- Safety concerns where LGBTQ+ people could face harassment from their identities being revealed (OpenGlobalRights, 2023).

A wide array of academic papers calls for the need to reduce and eliminate, where possible, the perpetuation or amplification of societal bias, especially when AI tools are used for mental health support. For example:

- Research shows that some conversational agents default to binary gender assumptions or misgender users (Liao *et al*., 2023)
- A widely used algorithm in US health care assigned lower risk scores to Black patients despite equal levels of illness, leading to reduced access to mental health and behavioural health services (Obermeyer *et al*., 2019).

A major review last year brought together findings from 36 studies and mapped how AI technologies support mental health care across five phases: pre-treatment (screening), treatment (therapeutic support), post-treatment (monitoring), clinical education, and population-level prevention. Again, this was not just about mental health chatbots designed for direct use by people with mental health conditions, but it has general implications for design of AI tools. It called for more ethical design for AI-driven digital tools and found there were "recurring challenges such as algorithmic bias and data privacy risks" (Yang and Fanli, 2025).

According to The Ada Lovelace Institute: "Researchers have recently shifted their focus to multilingual language technologies to produce machine translation systems that also work with 'low-resource' languages, and natural language processing tools that use data from a broader range of languages." But it goes on to say: "While this sounds like a step in the right direction, it is a long way from solving the problems..." (Claus, 2024).

The UK Government's 2024 *Equity in medical devices: independent review* made 18 recommendations on reducing bias in medical devices, including AI enabled ones (Department of Health and Social Care, 2024). These recommendations included better public engagement, education, and good data governance. It also emphasised the need for intersectional approaches to bias and highlighted that AI enabled medical devices should be tested for fairness across demographic groups.

That all sounds like a good thing – giving people better access to higher-quality, inclusive AI tools. But emphasising the need for relational care and avoidance of isolation, I would argue that improving the content of the tools available is a narrow and reductionist approach. It risks:

a. Leaving people to their own devices when what they need is access to relational care
b. Exposing more vulnerable people to disproportionate harm.

Self-care using AI tools by marginalised groups may prevent access to care when it's needed. It may become by default or design a demand management tool. This seems to be happening in the US.

A recent poll (Montero *et al*., 2026) found that one in three Americans have turned to AI for their general health concerns in the last year. One in five who use AI in this way cite affordability and access concerns as major reasons, including larger shares of young and lower income people.

The Montero study shows larger shares of younger adults, uninsured adults, Black adults, and Hispanic adults are turning to AI chatbots for mental health advice. Uninsured adults are more likely than insured adults to say they've relied on AI for mental health advice (30% compared with 14%), as are Black (21%) and Hispanic (19%) adults compared to white adults (12%).

It may be that AI is being used more amongst those who are least likely to be able to access care. But it's actually more disturbing than that. Not only are people from poorer walks of life to be 'left to their own devices' - they may be at more at risk of harm (see Liao *et al*., 2023; Obermeyer, 2019). In fact, these harms are 'baked in' according to research by Bender *et al*. (2021).

Bender *et al*.'s study says that large language models can cause harm because they reproduce and amplify the biases embedded in the data they are trained on. The authors show that these models absorb patterns from vast online text that contain racism, sexism, homophobia, transphobia, and other forms of structural discrimination. When used in sensitive contexts such as mental health support, these biases can manifest in subtle but damaging ways — for example, by reinforcing stereotypes, misinterpreting user experiences, or generating responses that marginalise or invalidate certain identities.

The authors state that such harms are not accidental but structurally baked into the way the large language models are built, warning that *"without careful governance, these systems risk perpetuating the very inequalities they are often marketed as helping to solve."*

Research is also showing that users experiencing psychosis, paranoia, severe emotional vulnerability or social isolation may face "increased risks because of impaired reality testing, social isolation, altered belief-updating and susceptibility to chatbot sycophancy." And existing safety frameworks are poorly equipped to address such interaction-based harms (Dohnány *et al*., 2026).

Another recent paper (Hudon *et al*., 2025) examines who may be most vulnerable to harmful AI interactions and explores risk factors for 'AI psychosis' (i.e. psychosis induced by AI) which turn out to be emotional reinforcement and distorted reality testing. The paper foregrounds safeguarding needs and differential susceptibility among vulnerable populations.

This is all the more alarming when one realises:

a. There is massive increase in use broadly across different groups – young and old, people with mental health problems and without.

b. There is inadequate granular data about who uses what sort of AI tools within those different demographic and socioeconomic groups. Significant gaps remain in relation to deprivation, ethnicity, disability, neurodiversity, severe mental illness and trauma populations despite rapid real-world deployment (Department of Health and Social Care, 2024).

We just don't know what harm is being generated amongst the most vulnerable of users.

THE CONTEXT OF IMPLEMENTATION

REASONS FOR OPTIMISM?

We have covered some possible contested areas where benefits and harms will play out. It is impossible to predict which way things will go in such a complex and fast-moving field. Much will depend on socio-economic forces such as technological development, corporate influence, policy, regulatory and governance issues. And, I believe, what role patients and the public play.

If we are optimistic, we will have AI tools that can combine AI and human strengths. The former, fast, always-on, data-savvy and accessible. The latter bringing humans (professionals and communities) into the loop who are empathetic, nuanced, creative and relational. AI could become a powerful companion for reflection, growth and wellbeing, supporting people to meet with others, offering them a variety of community resources, helpful advice to mix and to mingle, and attuned to the friendship needs of users.

There could be hyper-personalised guidance embedded in everyday use - beyond generic advice to individualised plans based on real-time data from wearables, mood logs, diet tracking and more. You log high stress and poor sleep, and AI suggests a calming routine or alerts you to behavioural patterns.

Automatic coaches could become the first line of contact - smart algorithms might become the first contact point for primary care. If the little medical helper cannot respond, it will transfer the case to a real-life doctor. In the meantime, users will 'talk through' CBT modules or emotional processing techniques.

Therapeutic interventions could become more subtle and be able to challenge or reframe responses when appropriate. This would help manage demand on hard pressed services and allow specialist resources to be deployed for those who really need them. A plethora of books is already on the shelves that seek to explore these broader issues.

The Wellcome Trust seems relatively upbeat. Director of Mental Health Amanda Wolpert argues that we should approach the issues of AI out of curiosity rather than "assumptions" based on fear, and that it is through empirical science that we should look for answers around benefits and risks, while bearing in mind wider ethical considerations (Wolpert, 2025).

Optimists might choose to believe that the right socio-economic and regulatory conditions will prevail; developers will be able and willing (or be enforced) to design-in safeguards (e.g. responding appropriately to risk in people with severe mental health conditions). The right regulatory and governance structures will be in place that balance commercial interests and harm protection.

Wellcome is funding fundamental research on generative AI and mental health. These will include requirements for ethical considerations to be considered. This sort of initiative will help protect wellbeing. And some might argue that it is in the interests of commercial developers to bring to market safe and effective tools.

Several recent articles in *Wired* magazine suggest how developers will include systems and processes that serve to build in improvements that address many of the concerns outlined above. For example, a 2023 panel (Wired, 2023) featured leading AI ethics experts discussing how developers and policymakers are actively building safety measures into AI systems to address algorithmic bias, privacy risks, future existential risks and regulatory compliance.

GLOBAL AI TRENDS

Beyond development and design issues, there is an inherent conflict between the profit motive and safe digital interventions. Currently, US and global companies and the Trump administration are putting pressure on the global economy to adopt AI and digital solutions.

The broader future, when AI is connected to the bio revolution (a term that explores how advances in the biotech, pharmaceutical, biological industries and related technologies transform economies and societies globally), raises the stakes even higher. It may be that arguing about whether an app should be designed to be more culturally competent or safer, or link to a therapist or a doctor becomes almost marginal as the drift in public services becomes corporatised. As Mustafa Suleyman points out in *The Coming Wave: "Regular services require continuous consumption and registration... Corporate power may well override nation states' ability to contain them"* let alone policy makers, regulator and health professionals' capacity to oversee them (Suleyman, 2023).

There are some grim and gloomy commentators out there: one post on the Effective Altruism forum (a forum for discussion about how philanthropic ventures can be useful to help build a better world) claims "No amount of technical safety research creates oversight requirements for AI in immigration enforcement, or addresses illegal data centers poisoning the air, or strengthens chip export controls. These are governance problems, and governance problems require political power to solve. Even the parts of AI safety that are about technical alignment - e.g. making sure models follow instructions, don't deceive, don't pursue unintended goals, and don't help people make bioweapons - only matter if frontier labs are required to implement them. Without legislation, every technical safety advance is effectively optional" (Kim, 2026).

What are the chances of such legislation? Firstly, looking overseas: in 2025, in the US, the AI industry spent $105 million on federal lobbying. One in four federal lobbyists reported working on AI. Meta is spending $65 million to elect AI-friendly state officials (Moore, 2026).

In contrast, the Centre for AI Safety Action, a non-partisan advocacy organisation dedicated to advancing AI safety spent $310,000 in 2025 (Open Secrets, 2025).

The global AI mental health market hit $1.71 billion in 2025 (Grandview Research, 2026), showing a surge in demand for digital mental health solutions. And it's just getting started - the market is projected to grow 24% annually from 2024 to 2030.

The UK digital mental health market was valued at $294.1 million in 2024 and is projected to grow to around $734 million by 2030. This growth is said to be being driven by smartphone ownership (83% of UK adults), rising awareness of mental health, and accelerating adoption of mental wellbeing technology (Grandview Research, 2026).

Government funding for AI in mental health in the UK includes a £21m AI Diagnostic Fund and £250m in 2019 to fund an NHS AI Lab and its initial programme of work. In 2020, the NHS and UK Government issued guidelines to support AI procurement (Gardiner and Mutebi, 2025). The Best of AI website lists 112 apps for mental health (Best of AI, 2026). And let's remember: OpenAI says 40 million people per day are using ChatGPT for health conversations (OpenAI, 2026). That's an order of magnitude above what the purpose-built tools will ever achieve.

Now, tech giants have launched or announced consumer-facing health large language model-based AI assistants. These are models into which you can provide more personalised data, such as longitudinal health data and medical records. The earliest versions were announced in 2025, followed by six more by March 2026 (Athni, 2026):

- Alphabet subsidiary Verily with Verily Me
- Amazon with its One Medical Health AI assistant
- OpenAI with ChatGPT Health
- Anthropic with Claude for Healthcare
- Microsoft with Copilot Health

The horse may have bolted. The large language model horse is riding off over the horizon.

TRENDS IN AI AND THE NHS

The UK Government is putting its money and faith in a ten-year Modern Industrial Strategy (Department of Business and Trade, 2025) that presents "significant opportunities and challenges for digital health and care companies seeking to partner with the NHS", according to the Digital Healthcare Council (Digital Healthcare Council, 2026). This strategy commits to a Life Sciences Investment Surge including a £600 million Health Data Research Service to establish:

- "The world's most advanced, secure, and AI-ready health data platform"
- Streamlined NHS Access, promising "low-friction procurement"
- An "NHS Innovator Passport" to help innovative MedTech products reach patients more quickly
- An AI-ready health care system with an emphasis on AI adoption, backed by £500 million through the new Sovereign AI Unit and a 20-fold expansion of AI research resources by 2030, to position the NHS as an early adopter of AI technologies.

Aligning itself with the Government's emphasis on supporting innovation, the NHS has foregrounded digital transformation and views AI tools within a prevention and self-management discourse, with system efficiency and demand management paramount.

The NHS Ten-Year Plan emphasises digitisation and AI as a force for its three 'left shifts' (including towards prevention).

The *Health Service Journal* has commented that the Plan includes *"a welter of proposals – sometimes bizarre – to draw attention away from the plan's failure to reflect the reality of the recovery challenge… The NHS App will act as an AI GP dealing with a patient's primary care needs"* (McLellan, 2025).

Data and privacy issues are in play too. Efforts to implement electronic health records that collect and share data across siloed teams, departments, hospitals and public sector agencies might lead to data from patient-directed AI tools being widely shared across organisations. Some would say that is a good thing, because it means that patients and clinicians ensure consistency and coordination of care. Those defending such a position would argue that there are sufficient regulatory, technical and governance safeguards in place to limit such data from being widely shared (e.g. to commercial companies) and many AI tools do not and might not integrate with electronic health records at all.

Others might question whether data should be shared in this way, let alone with developers in order to train models, or more widely with commercial and data gathering agencies.

Much depends on the extent to which patient directed AI tools, such as the recently introduced ChatGPT Health, where users input their own clinical data, are integrated into clinical systems and how strictly data sharing controls are implemented.

Add into the mix, the NHS's £330–£400 million contract with Palantir to deliver the national Federated Data Platform, designed as a system wide infrastructure to store, integrate, and manage operational and clinical data. The NHS wants all trusts to use the Federated Data Platform, with an aim to enable information to flow across organisational boundaries to support care coordination, population health and public sector planning (NHS England, 2023). The Federated Data Platform will collate data from multiple clinical and non clinical sources, including remote monitoring systems and patient generated digital tools (NHS Transformation Directorate, 2022). So, when my information that I input into my wearable enters NHS systems, there is a risk that it could be shared across the federated network, reinforcing concerns about expanded data visibility and cross agency access (Digital Health, 2023).

There is another ethical issue here. Palantir has been branded unethical for its links to Trump and involvement in US surveillance. There are fears about how it will use data from the Federated Data Platform and the NHS (Campbell, 2023).

TRENDS IN AI AND MENTAL HEALTH

Mental health services face workforce shortages, long waiting lists, rising demand and funding constraints. AI promises scalability, 24/7 availability and seeming low-cost support. There does not seem to be an incentive towards ensuring that mental health chatbots, let alone stand-alone large language models lead to an increase in appropriate referrals to a GP, or a clinician. Even if they do, barriers to equitable access remain and would be visibly reinforced.

The business development models for self-guided tools that will emerge are still unclear: "*A tool that is largely focused on providing tools to document symptoms or manage acute episodes of anxiety as a complement to clinician provided care is very different from one that is meant to replace the therapist altogether, as would the relevant considerations change based on the severity of the condition and associated risks"* (Wickremsinhe and Krubiner, 2024).

It's possible that companies with mental health chatbots will choose to market direct to patients, and bypass health professionals and the demands of a complex and siloed market. A 2022 BMJ review found that most mental health apps are marketed directly to consumers, often with minimal clinical oversight or evidence requirements (Huckvale *et al*., 2019).

In the recent past, pharmaceutical companies have been allowed more and more to advertise products directly to patients. Academics have argued that such direct to consumer advertising allows companies to circumvent professional gatekeeping and influence patient behaviour directly (Donohue *et al*., 2007).

It is not a huge stretch to envisage that mental health chatbot companies may follow the same logic: companies seeking adoption of AI tools in the NHS face a wide range of systemic challenges – such as siloed working, interoperability (different systems communicating with each other) and procurement issues. These companies could influence patients directly, where there are 'low barriers to entry' and financial incentives, in order to drive adoption and bypass clinicians (Rock Health, 2022).

This is exacerbated by regulatory weaknesses that allow mental health chatbots to operate in a grey zone and the need for AI companies to 'monetise' their products, many of which have been developed with up-front investment.

Despite its strong evidence base and widespread recognition, Woebot withdrew its consumer chatbot in 2025. Reports suggest that maintaining a sustainable business model alongside the costs of regulatory compliance and the rapid pace of generative AI development proved challenging, illustrating the commercial pressures facing clinically oriented mental health chatbots (Aguilar, 2025).

Several mental health platforms are adopting hybrid business models that combine AI-powered self-help with optional access to human clinicians. For example, Headspace's AI companion, Ebb, provides conversational self-reflection and personalised mindfulness support within the Headspace platform, while users can also access human coaching, therapy and psychiatry through separate services (Headspace, 2026). This reflects a broader trend towards hybrid models of mental health care, in which AI provides scalable self-guided support while human clinicians deliver assessment, therapeutic relationships and clinical oversight, combining the strengths of both approaches (Stade et al., 2024).

Monetisation as a business model could bake in inequalities. At present, you can get ChatGPT free, but if you become a heavy user, there is a fee-based subscription model. There are health related large language models that build in functions like adding one's own personal data from other apps. And there is also the possibility of AI models incorporating advertising – this has wider implications for targeted advertising, use of data and security issues (Backholer and Ciriello, 2026).

We may well get more sophisticated chatbots. But additional functions may be charged for, such as access to therapy, or be monetised via advertising, which comes with its own privacy concerns. Might sophisticated hybrid mental health chatbots be used by the more privileged who can click on a link to specialised mental health support, with others stuck in a 'wild west' of free versions of large language models?

Some might argue that ease of access to the free models may benefit people anyway; that the situation will be better than it is for those who are vulnerable and marginalised – that guardrails and culturally inclusive features will be developed and be even easier to access. However, without proper regulation, information content, quality and safety are at risk.

The bigger risk then is that AI tools could *exacerbate dependency and reinforce social isolation*. If one argues that good access means good access to a good health bot, then fine – ChatGPT or its successors will help. But if one argues instead for 'relational' care - and here one could add, what about the role of peer support workers - then enhanced use of AI tools might close down choice to those who can't afford it.

Much then hinges on the premium placed on relational care. When ChatGPT itself was prompted as to its summation of benefits and harms of mental health chatbots it stated: *"AI should extend empathy, not replace it. Equity, inclusion, and human oversight must guide its role in mental healthcare"* (ChatGPT, 2025). But how?

As Steven Levy points out: "*The remedy for this [preponderance of AI informational tools] is to identify and sustain sources that maintain standards worthy of trust... **If the marketplace works as it should** [emphasis added], those who publish reliable, verifiable information will thrive... This works, however, only if the general public understands the importance of accurate information and demands no less. If people don't care about facts anymore, they may well be ignoring reality until it literally kills them. It could happen!"* (Levy, 2025).

REGULATION AND GOVERNANCE

AI models aren't held to strict medical standards. There is a plethora of seemingly uncoordinated regulation and oversight. As early as 2019, at least 41 mental health chatbots were on the market, most of them claiming to provide therapy (Abd-Alrazaq *et al*., 2019).

While some apps like Wysa and Woebot cite research to underpin claims of harms and benefits, few AI chatbots are validated in clinical trials. And rigorous research in this area has lagged behind AI's rapid development (Miner *et al*., 2019).

Meta-analyses comparing randomised controlled trials on the effectiveness of tools do exist. They include studies of different types of chatbots in terms of function - retrieval-based, rule-based, and generative. A systematic review and meta-analysis from 2024 showed a significant effect of conversational agents on depression but did not discuss the type of chatbot or AI model (Zhong *et al*., 2024).

Even the most recent meta-analysis from 2025 on young people included only three studies examining more sophisticated generative AI tools (Feng *et al*., 2025). Therefore, as there are still only a few studies exploring the effect of generative AI chatbots on mental health, more evidence is needed to examine their effectiveness (Kuta *et al*., 2026).

Many mental health chatbots raise questions of safety that have been discussed only by a few studies so far (De Freitas and Cohen, 2024).

Existing regulation systems struggle with digital and AI tools. This is because traditional medical device regulation assumes stable interventions, predictable outputs and measurable risks. It operates on a pre-testing phase (with assessment of benefits and risks in different types of studies) and post-licensing phase (ongoing monitoring for safety).

But generative AI mental health systems are adaptive, conversational, emergent and relational. We are dealing with systems that are unlike traditional medical devices. Harms may emerge through interaction, not just technical failure. It's not enough to look theoretically at conversational accuracy, reliability and harm. This stuff is relational with an almost infinite variety of uses for them by vastly different demographics. Outputs from AI tools change all the time, updates occur continuously and iteratively, and the same prompt can trigger a multitude of responses.

The European Union Artificial Intelligence Act was the first attempt to establish a comprehensive legal framework specifically for AI systems (European Union, 2024). Regulation (EU) 2024/1689 lays down harmonised rules on artificial intelligence. It adopts a risk-based approach, imposing stricter requirements on systems considered more likely to affect safety, rights or wellbeing. High-risk systems — including some health care applications — may be subject to obligations relating to transparency, risk management, human oversight, technical documentation, data governance and post-market monitoring.

The legislation also includes provisions addressing generative AI and foundation models (such as ChatGPT and other large language models). While the EU approach is more precautionary and legally prescriptive than the UK, they still focus primarily on technical safety, accountability and procedural transparency, rather than relational and experiential harms, such as emotional dependency, persuasive interaction and simulated empathy.

Furthermore, under the EU Act, mental health chatbots and large language model systems do not automatically fall into a single risk category. General-purpose AI systems such as ChatGPT are primarily regulated as foundation or general-purpose AI models with transparency and risk obligations, rather than being automatically classified as high-risk.

However, systems used within health care settings - particularly those involved in diagnosis, suicide-risk assessment, treatment recommendations or therapeutic decision-making - are more likely to be classified as high-risk under the Act.

And when mental health chatbots position themselves as 'wellness,' 'coaching' or 'companion' tools rather than clinical interventions, they too might avoid stricter health care regulation.

We have a significant paradox and potential mismatch between formal regulatory classification and real-world psychological impact. Ironically, the bots designed as 'therapeutic tools' with more

constrained and purpose-designed systems may in some respects be safer, more auditable and more clinically governable than highly flexible generative systems that currently operate within broader regulatory grey zones. In short, the riskier stuff that gets used more is subject to less scrutiny.

Australia's Therapeutic Goods Administration has set out guidance that if a large language model designed for general use is later found to be providing health advice, the developer must implement controls to prevent it – or cease supply and seek regulatory approval (Devereux, 2026b).

In the UK, risks associated with AI mental health chatbots and large language model systems are currently governed through a patchwork of existing frameworks rather than through legislation specifically designed for generative AI mental health technologies.

Relevant mechanisms include the Online Safety Act 2023, medical device regulation through the Medicines & Healthcare products Regulatory Agency (MHRA), UK General Data Protection Regulation (GDPR) and data protection law, consumer protection frameworks and organisational governance within health care settings. However, these systems were largely developed for online content, privacy protection or traditional health care technologies rather than adaptive conversational systems capable of shaping emotional experience and human behaviour over time.

A UK commission on the regulation of AI in health care is, as of June 2026, due to make its recommendations soon. A recent Hardian Health conference highlighted that post-market surveillance – the MHRA's traditional post-licensing mechanism for staying on top of safety – is not enough to keep track of the newer technologies (Clover and Devereux, 2026). One of the commission members said: "Reactive monitoring is not enough." Furthermore, officials say they had not received the amount of complaints or feedback about AI products they expected, especially because their equivalents internationally are seeing an increase.

Things have got worse recently. In May 2026 it was reported that proposed amendments to UK medical device regulations risk creating the lowest barrier to entry for high-risk AI devices in the developed world (Devereux, 2026a).

Under the draft rules, which have been submitted to the World Trade Organisation ahead of being laid before Parliament, software designed to diagnose a condition can face greater regulatory scrutiny than software designed to treat one: "*This means a company could deploy an AI chatbot designed to treat patients with severe mental health problems without independent regulatory scrutiny by self-certifying its own safety in the same category as a walking stick*" (Devereux, 2026a). One expert states in this article: *"While the amendments include some positive elements… the risk classification rules within the draft amendments are a disgrace… The thrust of this is we're going to allow a whole bunch of novel software, AI systems, therapeutics, decision support software systems onto the market as Class I devices, which have no regulatory oversight."*

The AI commission also lacks patients and the public within its panel membership. It has commissioned some engagement work from The Health Foundation and National Voices, but this has been initiated only after the commission started its work. It is hard to envisage how it will affect the specialist work teams, that include a wide array of technical developers and companies, including Palantir, that have been working together for months.

The *Health Service Journal* also reports that clinicians are using large language models like ChatGPT unofficially and illegitimately as these are unlicensed products, and there are huge concerns about liability if things go wrong (Clover and Devereux, 2026). It is beyond the scope of this report, but it is worthwhile noting: there are significant concerns about the 'wild west of AI' within NHS systems, let alone in the public arena (Devereux, 2025).

Questions around transparency and accountability remain: who is responsible for harmful outputs? What should people be told? How transparent should training and safety design be? What audit mechanisms are possible? Who is liable when things go wrong?

Governance goes beyond state regulation. It cuts into organisational governance, procurement decisions, the role of ethics boards, independent auditing, and crucially service-user engagement in oversight.

Lack of proper regulation remains the major issue, despite the plethora of papers on what should happen to contain the market, and the preponderance of 'ethical guidelines' and 'international standards.'

Some have called out 'regulatory capture' with millions spent by lobbying companies to water down regulation. One observer of the AI scene states on a recent forum: *"I'm particularly disappointed by researchers who pushed for regulation two years ago now talking about watered down ethical guidelines, industry standards and public-private partnerships... The Gold Standard of preventing AI harm is still regulation. Imagine proposing to deregulate cars or food safety. This would be an untenable position. Why should we settle for less than food safety with a dangerous new technology?'* (Krook, 2026).

PATIENT LEADERSHIP

The weaknesses in meaningful regulation and oversight seem to signify the critical importance of incorporating patient and public voices. The digital and AI future pivots on whether and how such voice is brought in as a humanising and relational principle in policy, regulation, design, delivery and oversight.

The shift has begun to occur in some companies that are beginning to consider wider ethical issues and argue for the inclusion of user and citizen (crowdsourced or volunteers') views (Pretz, 2025). IBM's approach has been outlined in an internal paper (Rossi, 2025). And there are several global initiatives that explore an international agreement on codes of practice, such as one from the US based National Academy of Medicine (Adams *et al*., 2025).

A Wellcome Trust commissioned report (Wickremsinhe and Krubiner, 2024) highlights "the importance of ensuring that any work in digital mental health technology development, evaluation, and deployment aligns with the priorities of people with lived experience... which are:

- Trustworthy use (e.g., with respect to data protection, privacy, safety, and security)
- Additive rather than substitutive care provision
- Equitable design
- The promotion of agency and choice for end users."

However, people seem to be voting with their feet, or rather their fingers, if one takes on board current usage figures (i.e. flocking to use ChatGPT and apps that seem not to have met any of the above characteristics).

It's worth pointing out that involving consumers in design is different from ensuring people who use services or interventions have equal say in decision-making. As we have seen above, including people in design and governance for tools coming onto the market is a different – and much easier prospect – than involving people in regulatory decision-making, overall national policy on AI, and trying to do something about the large language models that escape oversight.

There is a gathering movement seeking to redress the balance. The Light Collective seeks to secure

"the collective rights, interests and voices of patients in health tech" (Light Collective, n.d.) Formed by a team of patient activists, coders, health experts, and data journalists after the Cambridge Analytica data breach, it operates to investigate privacy implications.

There is a growing mix of individuals and activists who are concerned: *"you don't need everyone to care about loss of control and superintelligence alignment. You need privacy advocates furious about AI-powered mass surveillance, environmentalists furious about illegal gas turbines poisoning children in Memphis, parents furious about chatbots encouraging their kids to kill themselves, women furious about Grok still making sexualized images of them, workers furious about displacement with no transition plan, national security people furious about chips being sold to adversaries, anti-war activists furious about Minab"* (Kim, 2026).

With a dominance of technical and commercial voices and weak representation in AI regulation debates, what would meaningful patient and service user leadership look like in the governance, design and evaluation of mental health AI? And is it feasible?

I have written elsewhere about the failure of traditional engagement mechanisms and the need for systems to embed patient (lived experience) leadership in all aspects of governance, design and delivery (Gilbert, 2019). It is critical that the movement starts to address the AI world at all levels – policy, regulation, governance, development and design.

I have long argued that the USP of lived experience is deeply relational and serves to focus more on connected and joined up services and design (Gilbert, 2022) – this notion can easily apply in the digital and AI world. Lived experience matters uniquely here as AI operates in relational, emotional and existential territory.

The benefits of patient (lived experience) leadership include being able to reframe problems and generate wider options for solutions (Gilbert, 2019). Experiential expertise is not optional, as patients can identify the following issues in ways that technical teams might not:

- Subtle harms
- Emotional dependency
- Invalidation
- Coercive dynamics
- Tone and relational impact.

Mental health AI is not merely a technical innovation but a social and relational intervention shaping human experience. So, democratic and experiential governance are essential.

CONCLUSIONS AND FURTHER ISSUES TO EXPLORE

Overall, this discussion paper outlines the different issues we need to consider when exploring the impact of AI tools in mental health self-care. Then it has examined the forces that might tip these issues either towards or away from harming people and promoting benefit.

This final section discusses and summarises my own personal and professional reflections.

THE UNCONTROLLED EXPERIMENT

People using AI mental health interventions are, without knowing it, taking part in a global societal experiment; an experiment that lacks adequate evidence, governance and experiential safeguards.

Millions of people are already using large language model systems and companionship tools for emotional support, reflective conversation and therapy-like interaction despite limited long-term evidence regarding psychological impacts, dependency, relational harms or behavioural influence.

We have inadequate regulatory frameworks to deal with these sorts of harms. The rapid expansion of AI mental health chatbots and self-guided AI systems has outpaced the development of robust public interest mechanisms for monitoring use, harms, relational impacts and inequality effects, leaving significant gaps in evidence, accountability and patient safety oversight.

Though development and design will tackle the worst effects of 'hallucinations' and mistakes, and make products 'safer' for many, AI is still a black-box technology. We have, unintentionally, created conditions resembling a regulatory 'wild west,' particularly for emotionally vulnerable people.

AMPLIFYING INEQUALITIES

AI mental health tools may improve access, immediacy and affordability for some people. However, benefits will not be distributed evenly and AI could exacerbate inequalities more widely. AI systems enter already unequal health care and social systems. Digital exclusion, unequal access to high-quality care, cultural and linguistic bias in training data, and varying levels of vulnerability to persuasive or dependency-forming interactions may all contribute to unequal outcomes.

At the very least we need better understanding of use across different population groups. But we need also to design policy and regulatory systems that focus on those who are least protected and most at risk. Meanwhile, the context within which AI and digital tools are being introduced aligns with a policy framework of innovation rather than safety, and hence to an NHS demand management mindset and technocratic approach to prevention.

Relational or psychological harms may disproportionately affect already vulnerable people. I envisage the emergence of a two-tier system in which affluent users can pay for monetised functions including access to therapeutic support, while disadvantaged groups may have to rely solely on automated or self-guided interventions.

DISPROPORTIONAL HARM

The risks associated with generative AI mental health systems may fall disproportionately on those who are already vulnerable. People experiencing loneliness, social isolation, financial disadvantage, stigma or long waits may rely heavily on AI for emotional support or therapeutic interaction.

These populations may have fewer alternatives, less access to human care and greater exposure to poorly understood relational harms such as emotional dependency, persuasive interaction or overreliance on automated systems. AI mental health technologies could become most embedded in contexts where human support is already weakest, not only deepening existing inequalities, but concentrating harms among those least well protected.

REGULATORY INADEQUACIES AND POSSIBLE WAYS FORWARD

The concentration of power and data within large technology companies raises concerns regarding accountability, democratic oversight and whose experiences and values shape AI systems.

Existing regulatory approaches focus on technical safety, transparency and procedural accountability, not relational and experiential harms such as emotional dependency, persuasive interaction and simulated empathy.

Moreover, AI tools in this arena can operate in regulatory grey areas by positioning themselves as 'wellness,' 'companion' or 'coaching' tools rather than formal health care interventions, even where people engage with them as substitutes for therapy or crisis support.

There's also a disturbing regulatory paradox at play. While chatbots that are intentionally designed as 'clinical' or 'therapeutic' tools and thus have to go through stricter regulation to get into the market, highly flexible large language models operate within looser regulatory grey zones. This is because current regulation tends to track formal intended *purpose* more closely than real-world relational and psychological *impact*.

There are some possible ways forward. AI systems continue to evolve after deployment through updates, changes in usage patterns and interactions with diverse user populations. For this reason, pre-market assessment alone is unlikely to be sufficient.

Enhanced regulation of products already on the market (post-market surveillance) could include dedicated reporting channels, standardised definitions of AI-related harms, proactive (patient-led?) mechanisms for collecting patient-reported experiences and regular review of emerging safety signals. Such systems would enable regulators, developers and health care organisations to identify risks earlier and respond more effectively.

PATIENT (LIVED EXPERIENCE) LEADERSHIP AND EXPERIENTIAL GOVERNANCE

Coproduction in design and development of AI tools is lacking, as it is in regulatory policy and practice. The advent of AI wellbeing tools for mental health underscores the need for patient-led and patient-centred policy, regulation, research, development and delivery.

Patient and service-user leadership should not be limited to consultation or feedback after deployment. Relational technologies require relational and experiential governance, including codesign, governance influence, procurement involvement, experiential oversight and meaningful participation in defining harms, benefits and acceptable risk.

REFERENCES

Adams, L., Fontaine, E., Matheny, M. and Krishnan, S. (2025) *An Artificial Intelligence Code of Conduct for Health and Medicine: Essential Guidance for Aligned Action.* Washington (DC): National Academies Press (US). doi: https://doi.org/10.17226/29087

Aguilar, M. (2026) Did AI really beat doctors at diagnosis? *Stat+ Healthtech newsletter*, 5 May 2026. Available from: https://www.statnews.com/2026/05/05/did-ai-really-beat-doctors-at-diagnosis-health-tech/ [Accessed: 10 May 2026]

Aguilar, M. (2025) Why Woebot, a pioneering therapy chatbot, shut down. *STAT News*. Available at: https://www.statnews.com/2025/07/02/woebot-therapy-chatbot-shuts-down-founder-says-ai-moving-faster-than-regulators/ [Accessed: 1 June 2026]

Abd-alrazaq, A. A., Alajlani, M., Abdallah Alalwan, A., Bewick, B. M., Gardner, P., Househ, M. (2019) An overview of the features of chatbots in mental health: A scoping review, *International Journal of Medical Informatics,* Volume 132, 2019. doi: https://doi.org/10.1016/j.ijmedinf.2019.103978

Alimour S. A. *et al*. (2024) The quality traits of artificial intelligence operations in predicting mental healthcare professionals' perceptions: A case study in the psychotherapy division. *Journal of Autonomous Intelligence 7(4):1-17* doi: https://doi.org/10.32629/jai.v7i4.1438

Apple S. (*2025) My Couples Retreat With 3 AI Chatbots and the Humans Who Love Them.* Wired. Available from: https://www.wired.com/story/couples-retreat-with-3-ai-chatbots-and-humans-who-love-them-replika-nomi-chatgpt/ [Accessed: 10 April 2026]

Athni T. S. (2026) Big Tech and the Rise of Consumer-Facing Health AI Assistants. *J Med Internet Res* 2026;28:e99230 doi: https://doi.org/10.2196/99230

Bachmann M., *et al* (2024) Exploring the capabilities of ChatGPT in women's health: obstetrics and gynaecology. *NPJ Womens Health*, 2:26 Available from https://www.nature.com/articles/s44294-024-00028-w [Accessed: 30 April 2026]

Backholer K., Ciriello R. (2026) Targeted advertising in generative artificial intelligence chatbots: a new public health risk. *The Lancet*, 407, 1665-1667. doi: https://doi.org/10.1016/S0140-6736(26)00464-2

Bernandi, J. (2025*) Friends for sale: the rise and risks of AI companion.* Ada Lovelace Institute. Available from: https://www.adalovelaceinstitute.org/blog/ai-companions/. [Accessed: 30 April 2026]

Best of AI. 112 AI Mental Health Tools. 2026. Available from: https://bestofai.com/category/ai-mental-health-tools [Accessed: 30 April 2026]

Byeon, H. (2026) The impact of social isolation and digital exclusion on mental and physical health in older adults: A meta-analysis. *Medicine (Baltimore)*, 105(4), e46010. doi: https://doi.org/10.1097/md.0000000000046010

Campbell, D. (2023). Patient privacy fears as US spy tech firm Palantir wins £330m NHS. *The Guardian*, 21 November 2023. Available from: https://www.theguardian.com/society/2023/nov/21/patient-privacy-fears-us-spy-tech-firm-palantir-wins-nhs-contract [Date accessed: 20 May 2026]

Claus, H. (2024) *Now you are speaking my language: why minoritised LLMs matter*. The Ada Lovelace Institute. Available from: https://www.adalovelaceinstitute.org/blog/why-minoritised-llms-matter/. [Accessed: 30 April 2026]

Clover, B., Devereux, E. (2026) *The Download: Shadow AI.* Available from: https://www.hsj.co.uk/technology-and-innovation/the-download-shadow-ai/7041659.article [Accessed: 10 May 2026]

Digital Health (2023). Available from: https://www.digitalhealth.net/2023/11/nhs-england-awards-330m-federated-data-platform-contract-to-palantir/ [Accessed: 15 May 2026]

De Freitas, J., Cohen, I. G. (2024) The health risks of generative AI-based wellness apps. *Nat Med* 30, 1269–1275 doi: https://doi.org/10.1038/s41591-024-02943-6

Department for Business and Trade (2025) *The UK's Modern Industrial Strategy Presented to Parliament by the Secretary of State for Business and Trade by Command of His Majesty.* Available from: https://assets.publishing.service.gov.uk/media/69256e16367485ea116a56de/industrial_strategy_policy_paper.pdf [Accessed: 30 April 2026]

Department of Health and Social Care (2024) *Equity in medical devices: Independent review – Final report*. GOV.UK. Available from: https://www.gov.uk/government/publications/equity-in-medical-devices-independent-review-final-report [Accessed: 30 April 2026]

Devereux, E. (2026a) *Revealed: 'Catastrophic' gaps in tech regulation plans.* Health Service Journal. Available from: https://www.hsj.co.uk/technology-and-innovation/revealed-catastrophic-gaps-in-tech-regulation-plans/7041722.article

Devereux, E. (2026b) *The Download: Wellness, or else.* Health Service Journal. Available from: https://www.hsj.co.uk/technology-and-innovation/the-download-wellness-or-else/7041586.article [Accessed: 10 April 2026]

Devereux, E. (2025) Wild West' of AI suppliers face new NHSE checks. *Health Service Journal*. Available from: https://www.hsj.co.uk/technology-and-innovation/wild-west-of-ai-suppliers-face-new-nhse-checks/7040140.article [Accessed: 10 April 2026]

Digital Healthcare Council. (2026) *Fortnightly Briefing*. Available from: https://mailchi.mp/6f12acb7fb53/fortnightly-members-update-17989339?e=41cf15c5e7 [Accessed: 10 April 2026]

UK Government. (2025) The UK's Modern Industrial Strategy. UK Government. Available from: https://assets.publishing.service.gov.uk/media/69256e16367485ea116a56de/industrial_strategy_policy_paper.pdf [Accessed: 10 April 2026]

UK Government. (2025) *National Commission into the regulation of AI in healthcare.* Available from: https://www.gov.uk/government/groups/national-commission-into-the-regulation-of-ai-in-healthcare [Accessed: 15 May 2026]

Dohnány S., Kurth-Nelson Z., Spens E., Luettgau L., Reid A., Gabriel I., Summerfield C., Shanahan M., *et al*. (2026) Technological folie à deux: Feedback Loops Between AI Chatbots and Mental Illness. *Nat. Mental Health,* 4, 336–345. doi: https://doi.org/10.1038/s44220-026-00595-8

Eisner, E., Berry, N., & Bucci, S. (2023) Digital tools to support mental health: A survey study in psychosis. *BMC Psychiatry*, 23, 726. doi: https://doi.org/10.1186/s12888-023-05114-y

Feng Y., Hang Y., Wu W., Song X., Xiao X., Dong F., Qiao Z. (2025) Effectiveness of AI-Driven Conversational Agents in Improving Mental Health Among Young People: Systematic Review and Meta-Analysis *J Med Internet Res* 2025;27:e69639 doi: https://doi.org/10.2196/69639

Essig, T. (2024) *Artificial Intelligence and Actual Psychoanalysis* American Psychoanalytic Association. Available from: https://apsa.org/artificial-intelligence-and-actual-psychoanalysis [Accessed: 10 April 2026]

European Union. (2024) *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence* (Artificial Intelligence Act) Available from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [Accessed: 10 April 2026]

Foster, C. (2026) *The five gynaecological symptoms you should never ignore*. The Independent. Available from: https://www.independent.co.uk/life-style/health-and-families/gynaecological-cancers-symptoms-ai-health-b2969367.html [Accessed: 11 May 2026]

Frances, A. (2025) Warning: AI chatbots will soon dominate psychotherapy. *British Journal of Psychiatry*. 1-5. doi: https://doi.org/10.1192/bjp.2025.10380

Gajjar, D. (2025) *Artificial Intelligence, an explainer*. UK Parliament Post. doi: https://doi.org/10.58248/PB57

Gardiner, H., and Mutebi, N. (2025) *AI and Mental Healthcare: ethical and regulatory considerations.* UK Parliament Post. Available from: https://researchbriefings.files.parliament.uk/documents/POST-PN-0738/POST-PN-0738.pdf. [Accessed: 14 April 2026].

Gibson, S., Tully, L., & Bucci, S. (2021) Acceptability and feasibility of digital health interventions for people with severe mental illness: A qualitative study. *JMIR Formative Research*, 5(4), e24693. doi: https://doi.org/10.2196/24693

Gilbert, D. (2019) *The Patient Revolution: How we can heal healthcare*. Jessica Kingsley Publishers.

Gilbert, D. (2022) *Humanising Health Care: The emergence of experiential practice*. Centre for Mental Health. Available from: https://www.centreformentalhealth.org.uk/publications/humanising-health-care [Accessed: 30 April 2026]

Grand View Research (2026. *AI In Mental Health Market (2026 - 2033)*. Available at https://www.grandviewresearch.com/industry-analysis/ai-mental-health-market-report [Accessed: 28 April 2026]

Gumley A. I., Bradstreet S., Ainsworth J., Allan S., Alvarez-Jimenez M., Birchwood M., *et al*. (2022) 'Digital smartphone intervention to recognise and manage early warning signs in schizophrenia to prevent relapse: the EMPOWER feasibility cluster randomised controlled trial', *Health Technology Assessment*, 26(27), 1–174. doi: https://doi.org/10.3310/HLZE0479

Habicht, J., Viswanathan, S., Carrington, B. *et al.* (2024) Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nat Med 30,* 595–602. doi: https://doi.org/10.1038/s41591-023-02766-x

Hao-Ping L., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel S., Banks, R., Wilson, N. (2025) The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1121, 1–22. doi: https://doi.org/10.1145/3706598.3713778

Hassabis, D. (2025) *On the future of work in the age of AI* [Interview]. *Wired*. Available from: https://www.wired.com/story/google-deepminds-ceo-demis-hassabis-thinks-ai-will-make-humans-less-selfish/ [Accessed: 30 April 2026]

Headspace (2026) *Meet Ebb: AI Mental Health Companion*. Available from https://www.headspace.com/ai-mental-health-companion [Accessed: 25 June 2026]

Henson, P., Peck, P., Torous, J. (2023) Digital Mental Health: A Scoping Review of User Characteristics and Engagement Patterns. *JMIR Mental Health*, 10(4), e43129. doi: https://doi.org/10.2196/43129

Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care. NPJ digital medicine, 8(1), 230. doi: https://doi.org/10.1038/s41746-025-01611-4

Huckvale, K., Torous, J. and Larsen, M.E. (2019) Assessment of the data practices of smartphone apps for depression and smoking cessation: Systematic analysis of privacy policies and practices. *JAMA Netw Open*. 2019 Apr 5;2(4):e192542. doi: https://doi.org/10.1001/jamanetworkopen.2019.2542

Hudon A., Stip E. (2025) Delusional Experiences Emerging From AI Chatbot Interactions or "AI Psychosis" *JMIR Ment Health* 2025;12:e85799. doi: https://doi.org/10.2196/85799

Institute for Health Metrics and Evaluation (2023) *GBD results* Available from: https://vizhub.healthdata.org/gbd-results. [Accessed: 20 May 2026]

Kim, S. (2026) How to actually spend billions on AI Safety. *Effective Altruism Forum*, 13 May 2026. Available from: https://forum.effectivealtruism.org/posts/npgwfZ6GpypGPKgyZ/how-to-actually-spend-billions-on-ai-safety [Accessed: 20 May 2026].

Kosmyna, N., Hauptmann, E., Yuan, Y.T., Situ, J., Liao, X.H., Beresnitzky, A.V., Braunstein, I. and Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task*. doi: https://doi.org/10.48550/arXiv.2506.08872

Koulouri T., Macredie R. D., Olakitan D. (2022) Chatbots to support young adults' mental health: an exploratory study of acceptability. *ACM Trans Interact Intell Syst*. Jun 30, 2022;12(2):1-39. doi: https://doi.org/10.1145/3485874

Kuta B., Novak L., Zidkova R., Furstova J., Malinakova K.., De Winter A., Husek V. (2026) Effectiveness of a Fully Automated Mobile Therapeutic Versus a General Chatbot in Reducing Depression and Anxiety and Improving Well-Being: Feasibility Randomized Controlled Trial. *JMIR Ment Health* 2026;13:e82642. doi: https://doi.org/10.2196/82642

Krook, J. (2026) *Defeatism on AI Regulation is Counter-Productive*. Effective Altruism Forum. Available from: https://forum.effectivealtruism.org/posts/QFyv6z79oY9GWX7qn/defeatism-on-ai-regulation-is-counter-productive [Accessed: 28 April 2026].

Kulveit J., Douglas R., Ammann N., Turan D., Krueger D., Duvenaud D. (2025) *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*. Computers and Society. Cornell University. June 2025. Available from: https://arxiv.org/abs/2501.16946 [Accessed: 12 June 2026].

Kuyda, E. (2024) *Can AI companions help heal loneliness?* TEDAI San Francisco. Available from: https://www.ted.com/talks/eugenia_kuyda_can_ai_companions_help_heal_loneliness [Accessed: 28 April 2026].

Levy, S. (2025) *Ask me one thing*. Wired: Plain Text, 27 June 2025. Available from: https://link.wired.com/view/6515b05dc741b9c83303dc09o3hw5.5c1/b3e254eb. [Accessed: 20 May 2026].

Liang C. (2023) *My AI Lover – three women reflect on the complexities of their relationships with their AI companions. New York Times*, May 2023. Available from: https://www.nytimes.com/2023/05/23/opinion/ai-chatbot-relationships.html [Accessed: 20 May 2026].

Liao, Q. V., Prabhakaran, V., & Gerstenberg, T. (2023) Mental health chatbots and fairness: Exploring biases and harms in conversational AI for well-being. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. doi https://doi.org/10.1145/3544548.3580703

Light Collective, n.d. *Our story*. Available from: https://lightcollective.org/story/ [Accessed: 12 June 2026]

Linardon J., Xie Q., Swords C., Torous J., Sun S., Goldberg SB. (2025) Methodological quality in randomised clinical trials of mental health apps: systematic review and longitudinal analysis. *BMJ Mental Health*. 2025;28:e301595. doi: https://doi.org/10.1136/bmjment-2025-301595

Lucas G. M., Gratch J., King A., Morency LP. (2014) It's only a computer: virtual humans increase willingness to disclose. *Comput Human Behav*. Aug 2014;37:94-100. doi: https://doi.org/10.1016/j.chb.2014.04.043

McLellan, A. (2025) We have seen the government's 10-Year Health Plan: it is a mess. *Health Service Journal*. 23 June 2025. Available from: https://www.hsj.co.uk/policy-and-regulation/we-have-seen-the-governments-10-year-health-plan-it-is-a-mess/7039538.article [Accessed: 10 June 2026].

Milmo, D. (2025) It cannot provide nuance': UK experts warn AI therapy chatbots are not safe. *The Guardian*, London. Available from: https://www.theguardian.com/technology/2025/may/07/experts-warn-therapy-ai-chatbots-are-not-safe-to-use. [Accessed: 12 June 2026].

Miner A. S., Shah N., Bullock KD., Arnow B. A., Bailenson J. and Hancock J. (2019) Key Considerations for Incorporating Conversational AI in Psychotherapy. *Front. Psychiatry* 10:746. doi: https://doi.org/10.3389/fpsyt.2019.00746

Mooney A., Alarilla A., Cavellro F. (2026) *Healthy life expectancy trends in the UK: a watershed moment*. The Health Foundation. Available from: https://www.health.org.uk/reports-and-analysis/analysis/healthy-life-expectancy-trends-in-the-uk-a-watershed-moment [Accessed: 10 June 2026].

Moore, D. (2026) *AI Boom on K Street: One in Four Lobbyists Now Work on AI*. Sludge. Feb 24th, 2026. Available from: https://readsludge.com/2026/02/24/ai-boom-on-k-street-one-in-four-lobbyists-now-work-on-ai/ [Accessed: 20 May 2026].

Moore J., Grabb D., Agnew W., Klyman K., Chancellor S., Ong D. C., Haber N. (2025) Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2025). doi:* https://doi.org/10.1145/3715275.3732039

Montero A. *et al.* (2026) *KFF Tracking Poll on Health Information and Trust: Use of AI For Health Information and Advice*, published. Available from: https://www.kff.org/public-opinion/kff-tracking-poll-on-health-information-and-trust-use-of-ai-for-health-information-and-advice [Accessed: 5 May 2026].

NHS Digital (2025) *Adult Psychiatric Morbidity Survey: Survey of Mental Health and Wellbeing, England*. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/adult-psychiatric-morbidity-survey/survey-of-mental-health-and-wellbeing-england-2023-24 [Accessed: 5 May 2026].

NHS England (2023) Federated Data Platform. Available from https://www.england.nhs.uk/digitaltechnology/nhs-federated-data-platform [Accessed 1 June 2026]

NHS Transformation Directorate (2022) Federated Data Platform Outline Business Case.

NICE Technology Appraisal 97 (TA97) (2006) Computerised cognitive behaviour therapy for depression and anxiety. National Institute for Health and Clinical Excellence (NICE). [Withdrawn July 2018]. Available from: https://www.nice.org.uk/guidance/ta97 [Accessed: 12 June 2026].

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. doi: https://doi.org/10.1126/science.aax2342

OpenAI (2026) *AI as a Healthcare Ally How Americans are navigating the system with ChatGPT. Available from:* https://cdn.openai.com/pdf/2cb29276-68cd-4ec6-a5f4-c01c5e7a36e9/OpenAI-AI-as-a-Healthcare-Ally-Jan-2026.pdf [Accessed: 20 May 2026].

OpenGlobalRights (2023) *Artificial intelligence is putting LGBTQ+ people at risk*. OpenGlobalRights. Available from: https://www.openglobalrights.org/artificial-intelligence-is-putting-lgbtq-people-at-risk/ [Accessed: 12 June 2026].

Open Secrets (2025) *Focus on NVIDIA*. Available from: https://www.opensecrets.org/federal-lobbying/clients/summary?cycle=2025&id=D000036303 [Accessed: 12 May 2026].

Pretz, K. (2021) Hypergiant executive's AI-ethics framework concept recognized with IEEE award. *The Institute*, 22 March 2021. Available from: https://spectrum.ieee.org/hypergiant-executives-aiethics-framework-concept-recognized-with-ieee-award [Accessed: 10 May 2026]

Raczka R. (2025) AI therapists can't replace the human touch. Guardian Letters. *The Guardian* 11 May 2025. Available from: https://www.theguardian.com/society/2025/may/11/ai-therapists-cant-replace-the-human-touch [Accessed: 14 April 2026].

Roose K. (2024). Can A.I. Be Blamed for a Teen's Suicide? *New York Times.* Oct 2024. Available from: https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html [Accessed: 15 May 2026]

Rossi, F. (2025) How to Avoid Ethical Red Flags in Your AI Projects*. IEEE Spectrum*, 27 April 2025 Available from: https://spectrum.ieee.org/ai-ethics-advice [Accessed: 10 June 2025]

Rock Health (2022) *Digital Health Consumer Adoption Report 2022*. Rock Health Inc. Available from: https://rockhealth.com/insights/digital-health-consumer-adoption-report-2022 [Accessed: 14 May 2026].

Sharma *et al.* (2024) Facilitating Self-Guided Mental Health Interventions Through Human-Language Model Interaction. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). doi: https://doi.org/10.1145/3613904.3642761

Spanakis, P., Heron, P., Walker, L., *et al*. (2022) Measuring the digital divide among people with severe mental ill health using the Essential Digital Skills framework. *Perspectives in Public Health*, 142(6) 315–323. doi: https://doi.org/10.1177/17579139221106399

Stade, E. C., Stirman, S. W., Ungar, L. H., Weisz, J. R. and Graham, A. K. (2024) Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Communications Psychology*, 2, 56. doi: https://doi.org/10.1038/s44184-024-00056-z

Suleyman, M., and Bhaskar, M., (2023) *The Coming Wave: AI, Power and Our Future.* London: Vintage.

Szalavitz, M. (2025). Love Is a drug. A.I. chatbots are exploiting that. *The New York Times*, June 9 2025 Available from: https://www.nytimes.com/2025/06/03/opinion/chatbots-ai-addiction-love.html [Accessed: 14 June 2026].

Thakkar A., Gupta A., De Sousa A., (2024) Artificial intelligence in positive mental health: a narrative review. *Front Digit Health*. 2024 Mar 18;6:1280235. doi: https://doi.org/10.3389/fdgth.2024.1280235

Torous J., Jän Myrick K., Rauseo-Ricupero N., Firth J. (2020) Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. *JMIR Ment Health*. Mar 26, 2020;7(3):e18848. doi: https://doi.org/10.2196/18848

Bloomberg News (2025) Nvidia's Lobbying Pays Off in AI Export Fight. *Transport Topics,* 3 December 2025 Available from: https://www.ttnews.com/articles/nvidia-ai-lobbying-pays-off [Accessed: 13 May 2026].

Willemsen R. F., Versluis A., Aardoom J. J. *et al*., 2024. Evaluation of completely online psychotherapy with app-support versus therapy as usual for clients with depression or anxiety disorder: a retrospective matched cohort study investigating the effectiveness, efficiency, client satisfaction, and costs. *Int J Med Inform*.;189:. doi: https://doi.org/10.1016/j.ijmedinf.2024.105485

The Wellcome Trust, 2024. *How lived experience expertise shapes research and development in digital mental health*. [version 1; not peer reviewed]. doi: doi.org/10.21955/wellcomeopenres.1115394.1

Wells, K. (2023). An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR. Available from: https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea. [Accessed: 14 April 2026].

Wickremsinhe M., Krubiner CB. (2024). *Ethical considerations for the design, study, and deployment of self-guided digital interventions for mental health: An overview of the literature*. Wellcome Internal Report. Unpublished. [email: M.Nanayakkara-Bruce@wellcome.org if you would like to request a copy]

Wired event summary: *AI Horizons: Ethics, Risks, and the Road Ahead* (Dec 5, 2023). Available from: https://www.wired.com/video/watch/ai-horizons-ethics-risks-and-the-road-ahead [Accessed: 13 May 2026].

Wolpert, A. 2025. *AI and mental health: "it could help revolutionise treatments."* Wellcome Trust. Available from: https://wellcome.org/insights/articles/ai-and-mental-health-help-revolutionise-treatments [Accessed: 10 June 2026].

Yankouskaya, A., Almourad, M., Liebherr, M. *et al.* (2026) Who lets AI take over? Cross-national variation in willingness to delegate socially important roles to artificial intelligence. *AI & Soc*. doi: https://doi.org/10.1007/s00146-026-02858-5

Yang, N., Fanli J., (2025) A Scoping Review of AI-Driven Digital Interventions in Mental Health Care: Mapping Applications Across Screening, Support, Monitoring, Prevention, and Clinical Education. *Healthcare* 2025, *13*(10), 1205. doi: https://doi.org/10.3390/healthcare13101205

Zhong W., Jianghua Luo, J. Hong Zhang, H. (2024). The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis, *Journal of Affective Disorders,* Volume 356, 459-469. doi: https://doi.org/10.1016/j.jad.2024.04.057

APPENDIX ONE: GLOSSARY

There are many proposed definitions of AI, however, most of them are aligned around the concept of machines capable of human behaviours or creating computer programmes.

Artificial intelligence (AI): AI technologies are tools and services which have some level of autonomy in undertaking activities, generating new predictions and decision-making without direct human control. There are many types of AI technologies, some of which overlap or build on each other. For further detail on AI technologies, see 'AI, an explainer' (Gajjar, 2025).

Digital mental health interventions (DMHIs): Technology-based tools - such as apps, websites, wearables, and virtual reality - designed to prevent, manage, or treat mental health conditions. They enhance access to care, providing self-guided or coached support for issues like anxiety, depression, and stress.

Machine learning (ML) or predictive AI: These systems learn to find patterns in training datasets which are then typically applied to new data to make predictions, carry out processing tasks, or provide useful outputs (e.g. text translation or data modelling).

Natural language processing (NLP): A field of AI that uses machine learning to enable computers to understand and communicate with human language. It has helped enable generative AI including large language models.

Generative AI (GenAI): An AI model which generates text, images, audio, video or other media in response to user prompts. These are advanced machine learning models trained on large amounts of data, which enables them to create new data with similar characteristics to the data the models were trained on.

Rule-based AI: An alternative to GenAI is 'rule-based' AI. The system uses a set of predetermined rules to make decisions based on logical reasoning (such as clinical standards) rather than learning from data to make decisions, this can make them more predictable and transparent but makes them less adaptable. These are often used in systematic processes or diagnostic settings.

Prompt: A prompt is a natural language request submitted to a language model to receive a response back. Prompts can contain questions, instructions, contextual information, few-shot examples, and partial input for the model to complete or continue. From a user prompt, the model responds with generated text, embeddings, code, images, videos, music, and more.

CENTRE FOR MENTAL HEALTH

UNEQUAL BENEFITS, UNEQUAL HARMS

AI MENTAL HEALTH CHATBOTS, INEQUALITY AND THE RISKS OF SELF-GUIDED CARE

Discussion paper, published July 2026

Image: istockphoto.com/portfolio/AndriiLysenko

Centre for Mental Health is an independent charity and relies on donations to carry out further life-changing research.

Support our work here:
www.centreformentalhealth.org.uk/donate

Charity registration no. 1091156. A company limited by guarantee registered in England and Wales no. 4373019