UK Parliament
POST

POSTnote 738

By Hannah Gardiner,
Natasha Mutebi
31 January 2025

AI and mental healthcare: ethical and regulatory considerations

Overview

- Stakeholders across academia, industry and healthcare have been exploring and experimenting with artificial intelligence (AI) in mental healthcare practice and provision in the UK.
- There is potential for AI to support practical delivery and administration tasks, and to radically change how mental health is researched, diagnosed and treated. For more on these opportunities and the related delivery considerations see PN737.
- Notably there are AI tools built for NHS use, less regulated wellbeing apps, and general-purpose apps being used for unintended mental health purposes. Each have different regulatory implications.
- Ethical concerns have been identified in the use of AI tools in mental health. These include data privacy, bias and discrimination, equity of access, and a need to address transparency, accountability, and liability issues.
- Researchers have stated that large-scale studies over longer time periods will be needed to determine the potential benefits and risks of AI for mental healthcare. There has been a substantial regulatory response to this issue, including internationally.
- Several collaborative projects and partnerships between regulators are currently underway in the UK.

Background

Artificial intelligence (AI) has had significant public and policy interest, particularly since the development of Generative AI (GenAI), which brings unique opportunities and challenges for supporting the delivery of mental healthcare (see PN737, HS55 and key AI terms in Table.1).

AI technologies across healthcare settings have received government funds over the past few years (PN637), with aims of streamlining administration and enabling earlier interventions. Funding includes a £21m AI Diagnostic Fund launched by the Department of Health and Social Care (DHSC) and NHS in June 2023,[1] and £250m in 2019 to fund an NHS AI Lab and its initial programme of work.[2,3] In 2020, the NHS and UK Government issued guidelines to support AI procurement.[4,5]

The aims of streamlining and improving efficiency are relevant to mental healthcare delivery, which has faced increasing demand and workforce capacity and retention issues.[6–8] See more about challenges in mental healthcare delivery in PN737 and CP06988.

Use of AI tools across society has led to various ethical, social and legal concerns, including around security, privacy, transparency, bias, liability, labour rights, intellectual property and disinformation (PN708). These are relevant to AI's use in mental healthcare.

For detailed definitions of AI technical terms, please consult the POST AI Glossary.

Table 1 Key AI terms

Term	Description
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. Some of which can continue to adapt after being 'trained' on datasets and 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' PB057
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).
Generative AI (GenAI)	An AI model which generates text, images, audio, video or other media in response to user prompts[a]. These are advanced ML 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.

[a] "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."[9]

Rule-based AI	An alternative to GenAI is 'rule-based'[b] AI. The system uses a set of predetermined rules to make decisions based on logical reasoning. These are often used in systematic processes or diagnostic settings.[10]

Relevant policy and regulation

Emerging technology including AI is a UK-wide policy area whereas healthcare is a devolved policy area. This POSTnote refers to mental healthcare policy in England, although the research described is likely to have broad applicability across the UK.

Recent policy activity related to AI and mental health is summarised in Table.2.

Table 2 Examples of guidance, policy, and regulation of AI in mental healthcare in the UK and EU

Area	Activity
Mental health	• Mental health has been a key healthcare mission since 2021, now called the 'Mental Health Goals Programme'. [11] In 2022 the UK government launched Healthcare Goals, with Mental Health one of the five key priority areas.[12] In 2023 the Mental Health Mission was launched as part of this which includes "funding to support mental health research, services, and digital technology" and two demonstrator sites.[13] • The UK government published a Suicide Strategy in 2023 that included an ambition to further develop "knowledge of the potential benefits and risks of [AI] in relation to suicide prevention".[14] • The Mental Health Bill 2024-25 passed its second reading and reached committee stage in the House of Lords in January 2025 (see: PN671, PN685, PN695, PN722, CBP9132)[c].[15,16] The bill would make changes to compulsory detention and treatment of people with a mental disorder.
Artificial Intelligence (AI)	• In 2023, the UK government hosted the first global AI Safety Summit, resulting in the Bletchley Declaration.[17] The declaration lays out a framework of principles and commitments on AI safety. • In 2024, the government set up a Regulatory Innovation Office to support innovation,[18] and established an AI Safety Institute.[19] • As of November 2024, the Artificial Intelligence Bill 2023-24) passed its second reading in the House of Lords and a third reading was in progress.[20] (see RB16)

[b] Rule-based AI systems use a set of predetermined rules to make decisions (e.g. clinical standards), rather than learning from data to make decisions, this can make them more predictable and transparent but makes them less adaptable.

[c] See four distinct POSTnotes analysing the Mental Health Act reforms from multiple perspectives and a House of Commons Research Briefing on the reforms (Reforming the Mental Health Act, 2024).

- A notable development is the EU's Artificial Intelligence Act 2024, which designated different levels of oversight depending on the risk classification of any particular AI system (a risk-based approach).[21,22]

Standards, legislation and regulations[d], apply to Digital Mental Health Interventions (DMHIs) such as AI tools, particularly if they are supplied to the NHS (see Table.3).

Notably, any tools classified as medical devices by the MHRA are subject to additional regulatory oversight.

The regulatory response to rapid developments in this technology, includes the establishment of a multi-agency advisory AI and Digital Regulations Service to support innovators (Table.6) in health and social care,[24,25] and projects investigating how regulation and guidance could be updated – see section 'Regulatory Challenges and Responses'. [26]

Table 3 Examples of standards, regulation and guidance

NHS Standards	– DCB0129: requirements for clinical risk management for manufacturers.[27] There is also a connected standard (DCB160) for the health organisations who deploy and implement health IT systems, compliance with both is mandatory under the Health and Social Care Act (2012).[27] – Digital Technology Assessment Criteria (DTAC) which have been under review since 2024.[28] – Digitally enabled therapies assessment criteria (DET) applies to Digitally Enabled Therapy tools intended for use in NHS Talking Therapy services.[29]
National Institute for Health and Care Excellence (NICE)	– NICE is the NHS in England's health technology assessment agency. NICE evaluates digital health technologies, including AI, through its Medical Technologies Evaluation Programme. This programme focuses on technologies that could offer benefits to patients and the health and social care system and topics are prioritised to ensure alignment with national priorities for health and care. Manufacturers can support NICE to identify suitable technologies by registering their product in the NHS Innovation Service and health and social care staff, patients, and the public can suggest topics for guidance development.[30] – NICE has also published an Evidence Standards Framework (ESF) for Digital Health Technologies which was updated to be suited for AI technologies in 2022.[31] The ESF provides information that can be

[d] Other relevant laws include: Equality Act (2010), Intellectual Property Act (2014); privacy and common law; (PN708). As well as the Contracts Act (1990), and Consumer Protection Act (1987) (including product liability). The Online Safety Act (2023) and The National Health Service (Clinical Negligence Scheme for General Practice Regulations (2019) might also apply depending on the service or products offered.[23]

		used to support health professionals and other decision-makers to assess evidence supporting new technologies. This includes AI technologies aimed at delivering system efficiencies (for example streamlining communication) and AI technologies with direct benefits for patients.[32]
Medicines and Healthcare products Regulatory Agency (MHRA) – specifically for AI tools classified as medical devices	–	All medical devices on the market in the UK need to be registered with the MHRA[e]. Class I (lowest risk) medical devices can be self-assessed; Class IIa, Class IIb, and Class III require 'Approved Body'[f] certification.[35,36]
	–	Once products satisfy a conformity assessment they can place a UKCA product marking.[37]
	–	The MHRA also require companies carry out post-market surveillance and report 'adverse events'[g].[39]
	–	For products sold in Europe a 'CE' mark is needed, this is not from the Approved Bodies but rather a European Notified Body.[40]
International Organisation for Standardisation (ISO)	–	ISO 14971:2019 evaluating and managing product risks.
	–	ISO 14155:2020 standards for clinical investigation of medical devices for human subjects (good clinical practice).
	–	In 2024, the ISO began drafting a new standard for clinical evaluation of medical devices.[41]
	–	In 2021, the European Committee for Standardization published guidelines on health and wellness apps (CEN ISO/TS 82304-2).[42]
Care Quality Commission (CQC)	–	Any service where a listed healthcare professional is completing or overseeing one of 14 regulated activities must register with and be assessed by the CQC, including supplying evidence annually.[43] The CQC have specific evidence categories for mental health services,[44] excluding psychologists.[45]

[e] See a public access database of registered manufacturers and medical device types here.

[f] Approved Bodies were formerly called 'Notified Bodies', they are designated independent certification bodies.[33] The Approved Bodies layer of the regulation process is unique to medical device regulation. In 2024 they formed an industry body to support consistency in interpretation called Team AB.[34]

[g] Adverse events include any unexpected occurrence in a research trial participant, they are related to adverse reactions which are "Any untoward and unintended responses to the trial intervention, at any dose administered, including all AEs judged by either the reporting investigator or the sponsor as having a reasonable causal relationship to the trial intervention."[38]

Information Commissioners Office (ICO)	–	Organisations who process personal data need to register with the ICO, and ensure they meet the legal requirements of the Data Protection Act (2018).[46] There are also specific provisions for research activities.[47]
British Standards Institute (BSI)	-	Published a validation framework for the use of artificial intelligence (AI) within healthcare in 2023.[48]

Key ethical concerns

The ethical challenges outlined below are similar to those discussed about AI in general (PN708), and across healthcare more broadly (PN637). [49–52] This section explores how these apply in the specific case of mental healthcare, for example concerns on safeguarding and risk assessments.

Accessibility, inclusion and exclusion

Research has emphasised the need to ensure equitable distribution of and access to mental healthcare.[53–56]

Cross-sector sources suggest service digitalisation can offer some accessibility benefits, such as for those who struggle to be in groups or to travel.[57–61] For example, research suggested that for some individuals with mental health conditions an in-person meeting or call could be more stressful than interacting on a computer.[60]

Compared to other DMHI, AI solutions can offer additional benefits[62] such as the anonymity of chatbots reducing shame and enabling people to reach out sooner.[60,63,64]

Different NHS trusts have varied digital engagement, which could lead to disparities in access to digitally enabled healthcare.[65] Research on DMHI's in the US has identified the potential for reductions in geographic disparities in healthcare access.[66,67]

However, cross-sector stakeholders highlight the potential for digital exclusion[h] when introducing DMHI.[53,57–60,69–76]

A 2024 survey identified people with severe mental illness as a high risk group for digital exclusion due to skills-based barriers, regardless of access.[77] Service users and therapists also describe low literacy and visual or cognitive impairment as challenges,[60,61] although current solutions include DMHI's designed with big fonts, icons and clear multimedia content can facilitate engagement.[61]

[h] More on digital exclusion[68] and PN725.

Stakeholders suggest the introduction of person-centred digital skills training,[77] and hybrid approaches with staff supporting service users.[78]

Academic and third sector stakeholders emphasise that to aim for equity in access, other forms of mental healthcare and service engagement need to be maintained and resourced, for example, face-to-face counselling.[57,59,74,79–83]

Conversely, in a 2023 survey of people with psychosis, the majority (90%) owned a smart phone and would be willing to try a mental health app (88%).[84] Half of those surveyed also said they would prefer remote support or no supplementary support, although complex interactions between mental health and technology were among barriers identified for some people.[84]

Several trials of digital monitoring for psychosis relapse risk conclude that digital approaches are feasible and acceptable to service users,[85,86] although some adverse events were reported.[85] Implementation of a DMHI by one NHS Trust found enthusiasm to use it amongst people with severe mental illness, with participants describing freedom of choice and autonomy as benefits.[87]

Research has suggested that demographic and contextual differences can influence how likely people are to engage with technology.[88–91] For example, technological interventions need to be tailored to support older adults to access or engage with them.[61,88]

Underserved and underrepresented populations

Disparities in mental healthcare treatment outcomes for minority ethnic groups are highlighted by many reports (PN695).[8,92–94]

Research describes stigma around mental health as complex and culturally influenced,[95] and that cultural identity can influence individual responses to AI.[96] Therefore, AI tools can be designed or used in ways that do not work for particular populations or in specific contexts.[97–100]

Stakeholder proposals include diversity, equity and inclusion being integrated across product development.[101] Others highlight more inclusive AI tools should incorporate cultural diversity, nonverbal communication, and offer multiple languages.[74,102,103] For example, GenAI has mainly been trained on the English language.[104]

Notably, AI tools used in UK healthcare settings need to comply with the Equality Act 2010. This means that the outcomes of AI tools cannot be discriminatory against a particular protected characteristic, such as age or ethnicity.

The Children's Commissioner for England stated in 2024 that young people are particularly underserved by mental health services[105] (PN685) (RB7196). Experts suggest some young people represent an opportunity for digital services due to their digital and technology skills .[83,106] However, research has found some subgroups within younger populations may also experience barriers to developing digital skills.[107]

Public engagement of younger people found positivity about DMHIs and confidence using them.[108] However, professionals working with young people were alarmed about risks of using AI without a health professional, for example in terms of traumas being triggered.[108]

There are proposals to train young people on safe use of AI and chatbots, and provide parental control functionalities.[109] Unicef suggest the need to design safeguarding in AI tools from early stages.[110]

A 2023 study found implementation of a referral chatbot led to notable but variable increases in gender, sexual and ethnic minority group referrals,[111] with AI's human-free nature giving users confidence they would not be judged.[111]

This aligns with research suggesting people can be more willing to disclose to 'virtual humans',[112] and that GenAI services show promise for addressing mental health disparities in the LGBTQ+ community.[113] However, evidence describe this community as facing unique vulnerabilities, including:[113,114]

- 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

Potential harms to individuals

Stakeholders describe the need to reduce risks of harm including mitigating against data misuse, or use of AI systems to spread misinformation (PN719) (PN708), carry out scams or manipulate people.[100,106,115–120]

Stakeholders highlight engaging with misinformation can negatively impact mental health,[121,122] for example, through exposure to mental health misinformation.[123]

Stakeholders raise concerns that if technology enables earlier risk detection or self-diagnosis of mental ill-health, such labels might negatively impact people's behaviour and exacerbate risks.[55,108,124]

Research involving academic, third sector, lived experience, and trade union perspectives identified some concerns about digitalisation potentially eroding relationships between patients and healthcare professionals.[71,74,125] This includes where people use AI apps for self-management rather than receiving in-person care,[74] and reductions in people's willingness to engage in future treatment.[126]

Stakeholders also note concern about reductions in face-to-face support,[59,82,126–128] which some suggest is important for clinical outcomes in mental healthcare.[59,129,130]

Some research has indicated that excessive automation could lead to isolation,[55] and that excessive personalisation[i] could exacerbate the breakdown of social cohesion and polarisation.[55,116,131]

Research has also identified risks to people's autonomy and dignity,[56,62,132–135] such as:

[i] Hyper-personalization in healthcare is the use of data and technology to tailor care to each patient's needs. It can include customizing health plans, communications, and care management.

- the potential for loss of clinician autonomy[135]
- people becoming institutionalised in their own homes through monitoring by smart technologies[j] [56]
- reducing patients privacy through AI-supported video monitoring in care settings,[136,137] especially if they are not offered opportunities to give informed consent about the use of such surveillance and the sensitive data it produces[137,138]

Bias

Stakeholders emphasise the need to reduce and eliminate, where possible, the perpetuation or amplification of societal bias (PN637) (PN708), especially when AI tools are used for mental health support.[54,55,60,73,74,97,134,135,139–146]

Bias in AI tools (algorithmic bias) can stem from various places (Figure.1), including AI tools being trained on biased datasets and outputting discriminatory outcomes (PN637),[99,143,147–152] or developers making biased decisions in the design or training of AI tools.[150,153]

For example, mental health Electronic health record (EHR) data is susceptible to cohort and label bias.[154] This can occur because culture-bound presentations of mental disorders, combined with a lack of transcultural literacy among clinicians, often lead to both over- and under-diagnosis.[155]

People can also exhibit bias when using AI tools, such as over-relying on[55,143,156,157] or mistrusting[55] AI outputs (Table.4). All these biases can be conscious or unconscious.[k]

The charity Money and Mental Health highlighted possible negative consequences of bias in automated systems, including AI systems, in identifying specific people as 'high risk' with implications around exclusion from financial or insurance products.[115,159] See example 2 in Table.4.

A 2024 independent review on equity in medical devices produced seven bias reduction recommendations, including public engagement, education and good data governance.[142,160] Researchers propose intersectional accountability,[54] and that the outputs of machine learning modules can be tested for fairness, to measure how they behave for across different groups.[161]

[j] See extended discussion on use of robotics in social care here: PN591

[k] "Unconscious bias (also known as implicit bias) refers to unconscious forms of discrimination and stereotyping based on race, gender, sexuality, ethnicity, ability, age, and so on."[158]

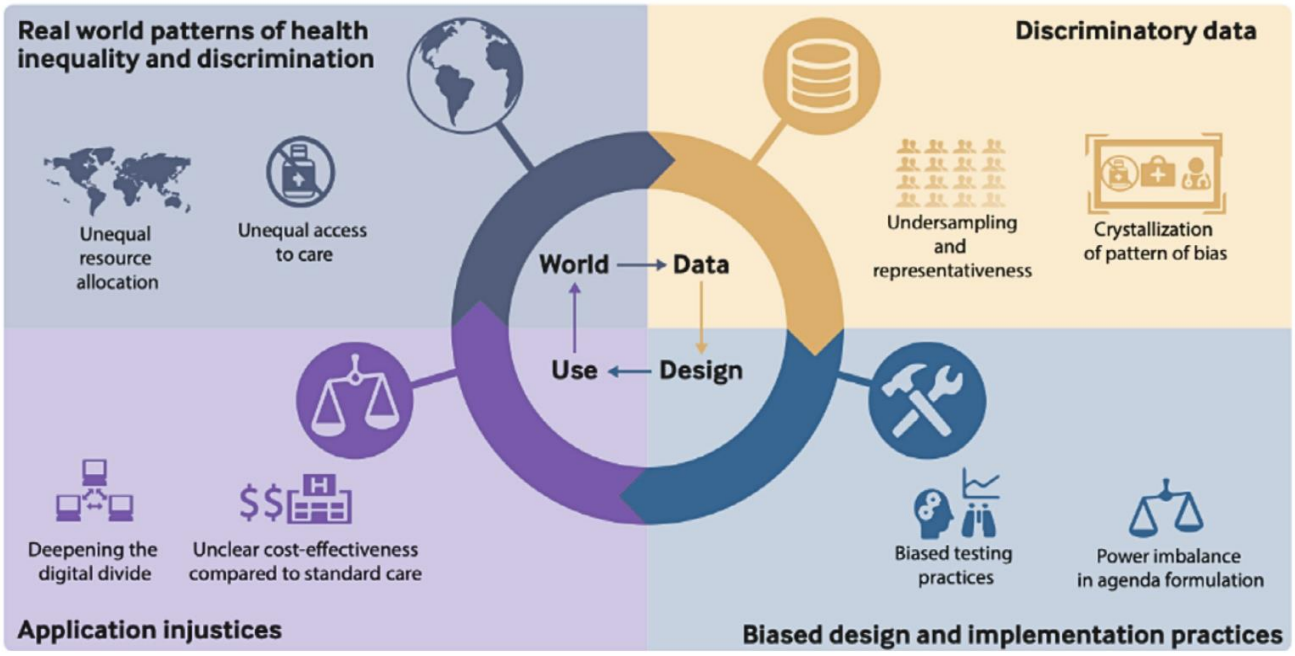

Figure 1: Sources of bias and discrimination with potential to interact with AI implementation and create cascading effects on health equities. From Fusar-Poli et al (2022, p.25)

Table 4 Implications of bias: case studies

Year	Country	Description
2024	Sweden	Amnesty International stated that Sweden's Social Insurance Agency was using a biased AI system that "disproportionately flags certain groups for further investigation regarding social benefits fraud, including women, individuals with foreign backgrounds… low incomes earners, and individuals without a university degree."[162] A similar case over 2013-18 in the Netherlands led to thousands of wrongful profiling and government officials resigning.[163,164,165–167]
2017	US	Research using existing service user datasets to evaluate the accuracy of commonly used ML algorithms (likely used by health insurers) for suicide risk prediction found the models were reasonably accurate for many ethnicities, but functioned poorly for Black and American Indian/Alaskan Native service users.[168] The authors highlighted potential effects could include unnecessary treatment and increased healthcare costs or burdens for such service users.[168]

Generative AI and unreliability

The recent development of GenAI has offered many benefits, see PN737. However, stakeholders highlight how it has created new complexities (see Table.5),[169–171] beyond the ethical concerns relevant to all AI systems.

Unreliable responses generated by GenAI models are one challenge,[171–173] including hallucinations (false information or forgetting service user details)[118,174,175] or harmful/risky responses, particularly in the case of commercial companion chatbots not designed for mental healthcare.[109,118,175–181]

Stakeholders also outline potential for use to create misinformation[120] or inadvertent misuse, such as sharing false information.[118,175]

Researchers suggest unreliability in GenAI needs technical responses to mitigate and reduce it,[141] such as controlling the resources the model draws from.[182]

Industry stakeholders suggest 'Rule-based AI'[l] might currently be more appropriate for use in mental healthcare delivery.[10]

Table 5 Examples of current unreliability in GenAI-based chatbots

Year	Chatbot provider	Description
2023	National Eating Disorders Association (US)	An AI Chatbot supporting those with eating disorders was suspended after giving weight loss advice[183]
2023-2024	Commercial companion chatbot providers	There are several documented international cases of people ending their lives after interacting with chatbots, which in one case provided encouragement.[184,185] However, the articles note the apps were not designed or regulated for mental health purposes, and describe the people in question as having recognised anxiety challenges.[184,185]

Data protection and privacy

Research has highlighted cybersecurity implications of integrating AI in healthcare (HS55, HS56, PN637, CBP9821),[73,186] including in mental health services.[187]

Stakeholders also suggest GenAI models have specific vulnerabilities to cyber-attacks, which create data protection and privacy implications.[188]

Stakeholders have noted privacy concerns from the use of AI for mental healthcare, such as companies gathering large amounts of data, with potential for the data to be used to target people with advertising.[53–56,74,109,134,144,151,157,186,189,190] This risk may be higher in wellbeing services, which are less regulated.[191] In 2023, the online therapy application BetterHelp faced large fines after selling sensitive data for use in advertising.[192]

[l] Rule-Based AI systems use a set of predetermined rules to make decisions (e.g. clinical standards), rather than learning from data to make decisions, this can make them more predictable and transparent but makes them less adaptable. See more here.

Data can be highly sensitive,[83,116] for example, people with mental health problems may be vulnerable to scams and manipulative advertising techniques, or struggle with compulsive spending.[115,159]

Some stakeholders have noted that companies can be bought out, and data transferred into new ownership, potentially overseas.[193]

Conversely, others have noted the importance of companies collecting and analysing user data to identify, report and moderate risky or criminal content disclosed through AI tools.[194]

Sometimes service users perceive DMHIs as providing greater privacy. For example, employers or family may be less likely to find out they are seeking help as the sessions are virtual and do not provide formal diagnosis.[108] However, service users may have concerns about who will own or manage the data, the risks of commercial exploitation, and effective data-sharing consent processes.[108,128]

Technical risk reduction measures proposed by researchers includes 'federated learning'[195,196] where the data remains on the user's device, or careful use of blockchain[m].[197] However, research highlights complexities related to blockchain and data protection, such as challenges around the 'right to be forgotten'.[198]

Regulatory responses on data protection and privacy

AI tools and services are subject to the Data Protection Act 2018, which is regulated by the ICO.

The ICO provided guidance in 2023[199] and a toolkit in 2022[200] on AI and Data Protection, and suggested eight questions relevant to GenAI developers and users.[201]

In December 2024, the ICO published a response to an open consultation on GenAI and data protection, resulting in updates to its position on 'legitimate interests lawful basis for web scraping to train generative AI models' and 'engineering individual rights into generative AI models'.[202] The ICO also noted a forthcoming joint statement on foundation models with the Competition and Markets Authority (CMA).[202]

The National Cyber Security Centre (NCSC) also offer guidelines and principles to embed privacy protections for AI/ML systems from development stages,[203,204] and launched a Cyber Resilience Audit (CRA) Scheme.[205]

Transparency, accountability and liability

The World Health Organisation (WHO) emphasise the importance of transparency[n] and accountability across research and implementation of AI systems for mental

[m] A blockchain is a distributed ledger with growing lists of records (blocks) that are securely linked together via cryptographic hashes (PB28).

[n] As the Alan Turing Institute explain, AI transparency means both interpretability of systems outputs (what it's saying and why it behaved the way it did), and justification for system processes and design (process transparency and outcome transparency).[206] See Figure.2. See more about proposals for improving transparency in AI model documentation here: PN708.

healthcare, including communicating failures and risk estimates.[54–56,134,135,151,207] See Figure.2.

Regulatory responses include ICO 2020 guidance on explaining decisions made with AI,[208] and MHRA guiding principles for transparency of ML-based medical devices published in 2024.[209]

Experts argue that service users should be informed on when, how and to what extent an AI is being used,[62,140,141,144,210,211] and be supported to understand the rationale behind AI decisions.[141,211,212]

Stakeholders raise concerns about who is liable if something goes wrong (PN708) (PN637),[72,120,213–216] suggesting accountability and liability should be spread proportionately across the clinical algorithm supply chain.[125,135,217–219] For example, creating a protocol if someone is misdiagnosed.[214]

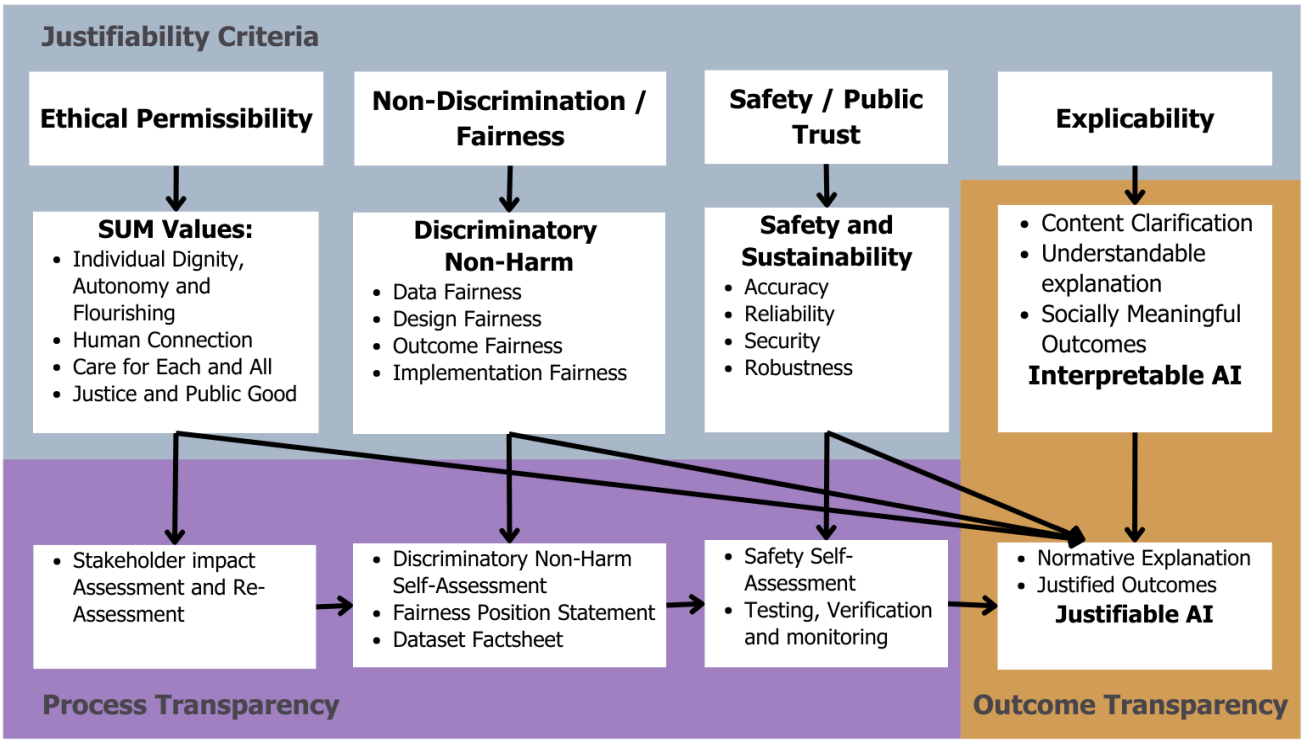

Figure 2: AI Transparency Map, illustrating Process Transparency and Outcome Transparency, taken from a 2019 Alan Turing Institute report (p.36)[206]

The quality of research and evidence

The evidence base for the various uses of AI across psychiatry and mental healthcare delivery appears to be limited. This is due to studies being small-scale or using small data sets,[102,220] and study quality, design and reporting often being insufficient.[221–224]

The MHRA and NICE have highlight difficulties in identifying what makes an AI tool effective or safe.[75]

Challenges in improving research into the clinical application of AI may include a lack of expertise. Research notes that clinicians may not be used to dealing with such depth of data or Research Ethics Committees (RECs)[o] and ethical review processes may not be fit to judge if an experiment is safe or scientifically justified.[226,227]

Similarly to other AI research,[228] stakeholders highlight unique challenges arising from the engagement of start-ups and the private sector in a large portion of AI mental health research. These include varied motivations for data generation (such as a desire to sell a product) and that products designed for user engagement could influence results.[53,229]

Data availability and quality

Similarly to the use of AI in other areas (PN637), many reviews identify challenges in accessing appropriate high quality data for research on AI/ML and mental health,[211,230–234] with social media data frequently used by researchers.[230,231]

NHS data has been found to be fragmented across NHS trusts and different health sectors with poor interoperability and integration.[50] Academic analysis highlights such data was never intended for AI research or development resulting in challenges around data quality and access.[153,235–237]

Researchers also highlight challenges around consent where data might be used for multiple purposes over time,[238] suggesting the need for transparency and ongoing dialogue with service users.[239,240] The NHS plans to increase interoperability and data collection standards (PN637), see Table.6 for more activities.

Table 6 Public sector collaborative activities to improve data availability

Name	Description
DATAMIND	A national data hub for mental health research, funded by the Medical Research Council (MRC) since 2021[241] as part of the Mental Health Platform (one of five challenge-led MRC-UKRI hubs).[242] It is working on a number of projects to unlock use of data,[243] including publishing a data collection tool so mental health data can be collected during all physical health studies,[244] and an Equity Audit tool.[245]
The National Federated Data Platform (NHS)	This software will enable NHS organisations to bring their operational data together, supporting data management and access for frontline staff as well as analysis and coordination at regional and national levels. Roll out is planned between 2024-2027.[246]
NHS England Secure Data Environment	This covers England, and supports approved researchers with ethically approved projects to safely access anonymised NHS data.[247]
'One London'	An example of a connected data ecosystem which integrates data across multiple boroughs, organisations and services.[248]

[o] Research Ethics Committees (RECs) are part of research governance. RECs are responsible for reviewing/approving all research involving human participants, and are delegated such responsibilities by the research organisation within which they operate.[225]

OpenSAFELY	This is an open-source software platform to support analysis of electronic health records (EHR), which has been deployed by two large EHR providers in the NHS.[249] It has been used to explore mental ill-health in primary care,[250] and the Wellcome Trust is currently supporting integration of NHS Talking Therapies (IAPT) data into the platform.[251]

Regulatory challenges and responses

There are an estimated 20,000+ mental health apps available.[252] Research indicates that the majority lack robust scientific evidence to support them.[253,254]

Experts state that app and software classification is a global challenge, with some classified as medical devices while others promoting themselves as offering non-medical functionality such as 'wellbeing', which are therefore less regulated.[127,191,255,256]

In the UK, qualification and classification as a medical device is based on 'intended purpose statements' (claimed functionality) by manufacturers (including statements in adverts and social media)[257]

An MHRA and NICE Partnership funded by the Wellcome Trust over 2023-2026 is exploring and addressing key regulatory challenges.[26] Early outputs of the project align with and expand on findings in academic literature such as:

- Qualification as a medical device is based on symptoms targeted as well as functionality, and considers severity of condition and clinical risk.[75,258]
- Intended statement purposes are often inadequate for DMHI, or become inaccurate when product features are updated.[75,191,259,260]
- Additional guidance on when products qualify as medical devices is needed (MHRA guidance on this will be published in 2025), along with improved post-market surveillance and reporting of 'adverse events'[p].[75,191,260–264] Academics have proposed some recommendations for identifying adverse events in clinical investigations and post-market.[265]
- Concerns about managing risks when people use tools unsupervised, and that if people have a poor DMHI experience it might influence future help-seeking behaviour.[75]
- A whole system approach being needed, across government agencies, with aspiration for international harmonisation[q] of regulation on DMHI.[75,260,263]

[p] Adverse events include any unexpected occurrence in a research trial participant, they are related to adverse reactions which are "any untoward and unintended responses to the trial intervention, at any dose administered, including all Adverse Events judged by either the reporting investigator or the sponsor as having a reasonable causal relationship to the trial intervention."[38]

[q] "Regulatory harmonization represents a process where regulatory authorities align technical requirements for the development and marketing of pharmaceutical products."[266] The UK is a member of the International Medical Device Regulators Forum, this is one route through which global harmonisation could be achieved.

- Public engagement found that many favoured regulation, endorsement or authoritative guidance, and most were unaware of (or unlikely to use) MHRA Yellow Card reporting[r].[108]

The MHRA are translating general "Software as a Medical Device" guidance into specific guidance for DMHIs, including new guidance on which products will be considered medical devices.[268,269] Future project work includes providing greater clarity for developers in the UK regarding the standard of evidence that should be provided for DMHIs.[26] Additionally, the project it is hopes international engagement may result in a harmonised approach to clinical evidence generation and a framework for global compliance recognition.[26,270]

Other core regulatory activities can be seen in Table.7.

Researchers also recommend that initiatives to improve regulatory processes should: commit to give advance warning about guidance changes to reduce market uncertainty,[191] examine the whole system implicated in deployment,[259] include civil society organisations who can represent user interests.[81]

In addition to the EU AI ACT 2024 described in Table.2, see Table.8 for international regulatory case studies.

Some stakeholders suggest that the EU AI Act is likely to impact the development of AI tools in the UK, given that developers will wish to develop products for international markets.[271,272]

An analysis from European Parliamentary staff analysis stated that the UK's AI Bill (2023-24) is closer to the US regulatory approach, and may lead UK developers to look to US markets instead.[273]

Table 7 Other regulatory activities relevant to AI and mental healthcare

Date	Lead	Aims
Launched in 2023	MRHA, NICE, CQC, and the Health Regulatory Authority (HRA).	Establishment of the AI and Digital Regulation Service (AIDRS), to support innovators navigating regulatory complexity. This is a cross agency collaboration, and was formerly known as the multi-agency advisory service (MAAS).[24,25]
2023-2024	MHRA and NICE	An 'Innovative Devices Access Pathway' (pilot), focusing on technologies addressing critical unmet needs for which there are currently no solutions[274,275]
2024-2025	MHRA	The AI Airlock pilot; a regulatory sandbox[s] which will enable industry and regulators to work together to identify and solve

[r] The Yellow Card Scheme is run by the MHRA, it allows the public, patients or healthcare professionals to report problems with any healthcare products through their website or app.[267]

[s] "A regulatory sandbox is a controlled environment that allows innovators and businesses to test and develop new AI technologies in an environment with reduced regulatory constraints. The idea behind the sandbox is to provide a safe space for businesses and regulators to work together to understand how new technologies can be developed and regulated in a responsible and ethical way."[276] They are described in the AI Bill (2023-24) see RB16. The FCA and ICO already offer sandbox services.[277,278]

	regulatory challenges surrounding regulating AI as a Medical Device.[279] The pilot cohort have been announced.[280] The AI AirLock pilot is also being supported by the Department of Health and Social Care, the NHS AI lab, Team AB, and the ICO.[281]

Table 8 International regulatory case studies

Organisation	Example of	Description
The US Food and Drug Administration (FDA)	Trial of a pre-certification scheme for wellbeing apps	In the US a pre-certification programme for start-ups was trialled.[191,282] The trial report concludes the approach to be impractical, with legislative change needed to enable flexible risk-based regulation approaches to supplement established ones.[282] Additional academic analysis found some apps exempt from pre-certification had potential for privacy and data security risks, suggesting more detailed criteria are needed.[283]
The German Federal Agency for Drugs and Medical devices (BfArM)	Establishment of a fast-track pathway to evaluate and integrate healthcare technologies	In 2019 the Digital Healthcare Act in Germany established a 'fast-track' pathway to support integration of digital Healthcare technologies.[284] The BfArM provide a structured assessment for the Digital Health Applications (known in German as DiGa), which are then listed in a permanent online directory and can be part of standard care.[284–286] Devices must meet multiple criteria including being lower risk (Class I or Class IIa),[286] and similar products are grouped together, with pricing standards applying in some cases.[284] A 2023 evaluation found health benefits from DiGas.[285] Critiques include mixed user satisfaction and insufficient evidence for some DiGas; reforms to the scheme are underway.[284]

Risk mitigation

Researchers have found that effective mitigations for ethical risks are not yet in place.[135] They have recommended that practitioners, researchers and decision-makers should put more focus on identifying and implementing mitigation solutions before and during service delivery, rather than just risk identification.[102]

Designing ethics into AI systems

Stakeholders highlight that AI systems need ethical principles and values designed in from the early stages to enhance trust and user satisfaction.[101,147,206,287–293] See principles published by the WHO in 2024 (Figure.3). Examples of guidance to support this include:

- IEEE Standards Association 2019 recommendations for 'ethically aligned design' of AI systems[294]
- MHRA 2021 guidance on developing ML-assisted medical devices[295]

Figure 3 Consensus ethical principles for use of AI for Health

Source: WHO (2024)[73]. Similar principles are described in Floridi et al. (2018)[296]

Bias and other ethical challenges can result from AI's design and development processes, therefore many social and technical solutions to reduce these challenges have been proposed.[292,297,298] For example, academic research suggests chatbots could be specifically designed to encourage users not to be too emotionally or socially reliant on them or over-reliant on AI information, thus mitigating one of the risks chatbots could create.[177,299,300]

Assessments to improve safety

Research highlights the need to mitigate risks when implementing autonomous and semi-autonomous AI systems into complex, dynamic healthcare settings. It suggests that sources of risk include structural, organisational, technological, epistemic, and cultural factors.[301] Systematic reviews note many AI and mental health studies do not include sufficient safety assessments.[302,303]

NHS stakeholders stated that simulation-based approaches[t] prior to adoption could enable safety and efficacy evaluation.[305]

In 2023, the UK government established an AI Safety Institute.[19] In 2024 it released an open-source safety testing platform called 'Inspect' to support consistency in safety evaluation approaches.[307,308] Industry stakeholders are currently developing a tool for safety assessment of GenAI tools used in mental healthcare.[309,310]

AI-human collaboration in decision-making

Most academic and clinical stakeholders, and the public appear to favour systems designed for human-AI collaboration in mental healthcare delivery, as AI models could miss contextual information crucial to decision-making.[83,101,311–314]

Others argue that service users should be co-reasoners alongside clinicians,[315,316] and service user experiential knowledge must not be undermined by AI.[317,318] Research encompassing academic and therapist perspectives expresses concerns that generalised AI tools alone might not effectively deal with unique complexities which occur in service users.[60,82,319]

Academic studies highlight potential trade-offs between ensuring a human collaboration against delivering administrative efficiency savings,[210] and between interpretability and performance.[236] Therefore studies suggest research on optimising human-AI interaction is needed.[320,321]

Evaluations, assessments and labels

Numerous assessment approaches and labels have been proposed by academic and third sector stakeholders to address challenges of bias, data security and transparency across AI research and delivery, including in mental healthcare. See examples in Table.8.

Research has concluded that existing methods of health technology assessment need adaptation to be suited for use with AI-based health technologies.[322] A 2024 Ada Lovelace report concluded that at present, no available evaluation methods are able to adequately determine that an advanced AI system is safe.[323]

Experts highlight the difficulty in evaluating GenAI using traditional scientific methods, due to it producing variable responses (even to the same prompt) and its lack of transparency.[324] The UK Government's AI Safety Institute is currently developing evaluative work.[19,325]

Industry stakeholders also suggest businesses have AI Ethics Committees to oversee decisions.[287,326]

[t] Simulation-based approaches involve simulating real-world scenarios so interaction and learning can take place within a safe contained environment.[304] For example, this could include simulated scenarios of doctors explaining an AI-tool to patients so that they gain confidence doing this, [305] or simulating cases where doctors need to diagnose or suggest prescriptions and experimenting with them using an AI or not to see what effect AI has on their decision-making.[306]

Table 8 Examples of assessment approaches to mitigate risks of AI use

Name	Description
STANDING Together	The cross-sector partnership STANDING Together published Standards for data Diversity, Inclusivity and Generalisability in 2023.[327,328]
Algorithmic Impact Assessment	Algorithmic Impact Assessments to mitigate against bias have been piloted by the NHS.[329] These are suggested to be modelled on existing Data Protection Impact Assessments (PN708). Data protection audits are already sometimes carried out by the ICO (PN708).
The D-Seal	In 2022, the Danish government launched the D-seal which provides certification for companies who meet requirements for cyber security and handling of AI data. Alongside this Denmark established an independent Data Ethics Council and provided a Data Ethics Toolbox.[330]
Trustworthy Assurance of Digital Mental Healthcare	The Trustworthy Assurance process aims to support the design, development, and deployment of responsible technology. It involves reflective discussions including multiple stakeholders about all stages of product development and delivery. The idea is to develop appropriate trustworthy goals and ethical principles which can be embedded into digital mental health products.[127]

References

1. Department of Health and Social Care *et al.* £21 million to roll out artificial intelligence across the NHS. *GOV.UK*.
2. Hoeksma, J. (2024). DHSC slashes investment in NHS AI Lab by £111m. *Digital Health*.
3. Department of Health and Social Care *et al.* (2019). Health Secretary announces £250 million investment in artificial intelligence. *GOV.UK*.
4. Department for Digital, Culture, Media & Sport *et al.* (2020). Guidelines for AI procurement. *GOV.UK*.
5. NHS Transformation Directorate (2020). A buyer's guide to AI in health and care.
6. The British Medical Association (2024). Mental health pressures data analysis.
7. The British Medical Association (2024). Mental health workforce report.
8. Lord Darzi (2024). Independent investigation of the NHS in England. Department of Health & Social Care.
9. Google Cloud Introduction to prompting Generative AI on Vertex AI. *Generative AI on Vertex AI*.
10. Darcy, A. (2023). Why Generative AI Is Not Yet Ready for Mental Healthcare. *Woebot Health*.
11. Office for Life Sciences *et al.* (2024). Mental Health Mission. *Notice*.
12. Office for Life Sciences *et al.* (2024). Life Sciences Healthcare Goals. *GOV.UK*.
13. Office for Life Sciences *et al.* (2024). Mental Health Goals. *GOV.UK*.
14. Department of Health and Social Care (2023). Suicide prevention strategy for England: 2023 to 2028. HM Government.
15. Prime Minister's Office (2024). King's Speech 2024: background briefing notes. *GOV.UK*.
16. Department of Health and Social Care *et al.* Draft Mental Health Bill 2022. *GOV.UK*.
17. Prime Minister's Office *et al.* The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. *GOV.UK*.
18. Department for Science, Innovation and Technology *et al.* (2024). Game-changing tech to reach the public faster as dedicated new unit launched to curb red tape. *Press release*.
19. Department for Science, Innovation & Technology (2024). AI Safety Institute approach to evaluations. *GOV.UK*.
20. UK Parliament (2023). Artificial Intelligence (Regulation) Bill [HL] - Parliamentary Bills.
21. European Parliament (2023). EU AI Act: first regulation on artificial intelligence.
22. Busch, F. *et al.* (2024). Navigating the European Union Artificial Intelligence Act for Healthcare. *Npj Digit. Med.*, Vol 7, 1–6. Nature Publishing Group.
23. Department for Science, Innovation & Technology (2024). Online Safety Act: explainer. *GOV.UK*.
24. AI and Digital Regulations Service for health and social care Home. *AI and Digital Regulations Service for health and social care*.
25. Boffa, R. (2024). Navigating the Future: Building the AI and Digital Regulations Service. *AI and Digital Regulations Service for health and social care*.
26. Medicines and Healthcare products Regulatory Agency *et al.* (2024). Update on pioneering initiative on regulation and

evaluation of digital mental health technologies. *GOV.UK*.
27. NHS England Digital (2024). Clinical risk management standards.
28. NHS England Digital Technology Assessment Criteria (DTAC). *NHS Transformation Directorate*.
29. NHS England Digitally enabled therapies assessment criteria.
30. National Institute for Health and Care Excellence Medical Technologies Evaluation Programme | NICE guidance. *NICE*. NICE.
31. National Institute for Health and Care Excellence Evidence standards framework (ESF) for digital health technologies. *NICE*. NICE.
32. NICE (2022). Evidence standards framework for digital health technologies. NICE.
33. Medicines and Healthcare products Regulatory Agency (2023). Approved bodies for medical devices. *GOV.UK*.
34. Silverwood, J. (2024). New UK medical device industry body Team-AB officially launches. *Medical Device Network*.
35. Medicines and Healthcare products Regulatory Agency An introductory guide to the medical device regulation (MDR) and the in vitro diagnostic medical device regulation (IVDR).
36. Medicines and Healthcare products Regulatory Agency (2022). Chapter 2: Classification.
37. Medicines and Healthcare products Regulatory Agency (2020). Medical devices: conformity assessment and the UKCA mark. *GOV.UK*.
38. Coomarasamy, A. *et al.* (2016). Definitions of adverse events, seriousness and causality. in *PROMISE: first-trimester progesterone therapy in women with a history of unexplained recurrent miscarriages – a randomised, double-blind, placebo-controlled, international multicentre trial and economic evaluation*. NIHR Journals Library.
39. AI and Digital Regulations Service for health and social care (NHS) Post-market surveillance of medical devices.
40. Medicines and Healthcare products Regulatory Agency (2024). Regulating medical devices in the UK. *GOV.UK*.
41. ISO (2024). ISO/WD 18969: Clinical evaluation of medical devices [Working Draft].
42. CEN-CENELEC (2021). CEN has published new guidelines on health and wellness apps to help to sort the best from the rest.
43. Care Quality Commission (2024). Assessing quality and performance.
44. Care Quality Commission (2024). Mental health services: evidence categories.
45. Care Quality Commission (2024). Scope of registration: Glossary of terms.
46. Information Commissioner's Office (2024). For organisations. ICO.
47. Information Commissioner's Office What are the research provisions? ICO.
48. British Standards Institute (2023). Validation framework for the use of artificial intelligence (AI) within healthcare. Specification BS 30440:2023 | 31 Jul 2023 | BSI Knowledge. *BS 30440:2023 | BSI Knowledge*.
49. Bhatnagar, A. *et al.* (2024). How is artificial intelligence affecting society? POST UK Parliament.
50. Science, Innovation and Technology Committee (2024). *Legacy – Parliament 2019–24*. House of Commons.
51. Gajjar, D. *et al.* (2024). Artificial intelligence: ethics, governance and regulation.
52. Naik, N. *et al.* (2022). Legal and Ethical Consideration in Artificial

Intelligence in Healthcare: Who Takes Responsibility? *Front. Surg.*, Vol 9,
53. Muchamore, I. *et al.* (2024). How lived experience expertise shapes research and development in digital mental health: A Review of literature and Insights. The Wellcome Trust.
54. Mestre, R. *et al.* (2024). Building Responsible AI for Mental Health: Insights from the First RAI4MH Workshop. University of Southampton, and Institute for Experiential AI (Northeastern University).
55. Fusar-Poli, P. *et al.* (2022). Ethical considerations for precision psychiatry: A roadmap for research and clinical practice. *Eur. Neuropsychopharmacol.*, Vol 63, 17–34.
56. Crowther, N. *et al.* (2022). A digital cage is still a cage. *Care Management Matters*.
57. Health and Social Care Committee (2023). Digital transformation in the NHS. House of Commons.
58. Mind Digital tech. *The influence and participation toolkit*.
59. Corradi (2022). The role of technology in mental healthcare. The Nuffield Council on Bioethics.
60. Chapman, A. *et al.* (2024). Sociotechnical Considerations for Accessibility and Equity in AI for Healthcare. in *Companion Proceedings of the ACM Web Conference 2024*. 1158–1161. ACM.
61. Yin, R. *et al.* (2024). The views and experiences of older adults regarding digital mental health interventions: a systematic review of qualitative studies. *Lancet Healthy Longev.*, Vol 5, Elsevier.
62. Fiske, A. *et al.* (2019). Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *J. Med. Internet Res.*, Vol 21, e13216.
63. Vallejos, E. P. (2024). Mindful Mortality: Digital Mental Health and the Dying Process.
64. Stix, C. (2018). 3 ways AI could help our mental health. *World Economic Forum*.
65. Allcock, J. A. *et al.* (2024). Landscape of Digital Technologies Used in the National Health Service in England: Content Analysis. *JMIR Form. Res.*, Vol 8, e51859.
66. Graham, A. K. *et al.* (2021). Resolving Key Barriers to Advancing Mental Health Equity in Rural Communities Using Digital Mental Health Interventions. *JAMA Health Forum*, Vol 2, e211149.
67. Yom-Tov, E. *et al.* (2023). Digitally filling the access gap in mental health care: An investigation of the association between rurality and online engagement with validated self-report screens across the United States. *J. Psychiatr. Res.*, Vol 157, 112–118.
68. Tudor, S. (2024). Digital exclusion in the UK: Communications and Digital Committee report. House of Lords Library.
69. Mind (2020). Digital services for people with mental health problems and digital exclusion during the coronavirus pandemic.
70. Mental Health Foundation (2021). Tackling digital exclusion in older people.
71. Studman, A. (2023). Access denied? Socioeconomic inequalities in digital health services. Ada Lovelace Institute.
72. Bélisle-Pipon, J.-C. *et al.* (2021). What Makes Artificial Intelligence Exceptional in Health Technology Assessment? *Front. Artif. Intell.*, Vol 4, Frontiers.
73. World Health Organisation (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models.

74. Hamdoun, S. *et al.* (2023). AI-Based and Digital Mental Health Apps: Balancing Need and Risk. *IEEE Technol. Soc. Mag.*, Vol 42, 25–36.
75. Hopkin, G. *et al.* (2024). Considerations for regulation and evaluation of digital mental health technologies. *Digit. Health*, Vol 10, 20552076241293313. SAGE Publications Ltd.
76. Middle, R. *et al.* (2022). Experiences of digital exclusion and the impact on health in people living with severe mental illness. *Front. Digit. Health*, Vol 4, 1004547.
77. Spanakis, P. *et al.* (2024). Measuring the digital divide among people with severe mental ill health using the essential digital skills framework. *Perspect. Public Health*, Vol 144, 21–30. SAGE Publications Ltd.
78. Greenway, F. T. *et al.* (2024). Hybrid mHealth care: Patient perspectives of blended treatments for psychosis. A systematic review. *Schizophr. Res.*, Vol 274, 1–10.
79. NHS England (2024). NHS Talking Therapies for anxiety and depression manual.
80. Gentry, E. *et al.* (2024). 'I don't trust it, but I have to trust it': The Paradox of Trust vs Use of Online Technology Across The Mental Health Spectrum. OSF.
81. Bossewitch, J. *et al.* (2021). Digital Futures in Mind: Reflecting on Technological Experiments in Mental Health & Crisis Support. University of Melbourne.
82. Brown, J. E. H. *et al.* (2021). AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM - Ment. Health*, Vol 1, 100017.
83. De Choudhury, M. *et al.* (2023). Benefits and Harms of Large Language Models in Digital Mental Health. arXiv.
84. Eisner, E. *et al.* (2023). Digital tools to support mental health: a survey study in psychosis. *BMC Psychiatry*, Vol 23, 726.
85. Gumley, A. I. *et al.* (2022). The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in Scotland and Australia. *Lancet Psychiatry*, Vol 9, 477–486.
86. Lewis, S. *et al.* (2020). Smartphone-Enhanced Symptom Management In Psychosis: Open, Randomized Controlled Trial. *J. Med. Internet Res.*, Vol 22, e17019.
87. NHS Transformation Directorate A digital recovery platform for severe mental illness.
88. Spanakis, P. *et al.* (2021). Use of the Internet and Digital Devices Among People With Severe Mental Ill Health During the COVID-19 Pandemic Restrictions. *Front. Psychiatry*, Vol 12, Frontiers.
89. Hassan, L. *et al.* (2023). User engagement in a randomised controlled trial for a digital health intervention for early psychosis (Actissist 2.0 trial). *Psychiatry Res.*, Vol 329, 115536.
90. Borghouts, J. *et al.* (2021). Barriers to and Facilitators of User Engagement With Digital Mental Health Interventions: Systematic Review. *J. Med. Internet Res.*, Vol 23, e24387.
91. Yao, R. *et al.* (2022). Inequities in Health Care Services Caused by the Adoption of Digital Health Technologies: Scoping Review. *J. Med. Internet Res.*, Vol 24, e34144.
92. NHS England (2020). Advancing mental health equalities strategy.
93. Isiwele, A. (2024). Executive Summary: Promoting Cultural Humility in Mental Healthcare for Black Youth in London: Impact Workshop. *UCL Institute of Education (IOE). Faculty of Education and Society: London, UK.* UCL Institute of Education (IOE). Faculty of Education and Society.

94. National Collaborating Centre for Mental Health (2023). Ethnic Inequalities in the Improving Access to Psychological Therapies (IAPT) programme: a policy review. NHS Race and Health Observatory.
95. Ahad, A. A. *et al.* (2023). Understanding and Addressing Mental Health Stigma Across Cultures for Improving Psychiatric Care: A Narrative Review. *Cureus*, Vol 15, e39549.
96. Barnes, A. J. *et al.* (2024). AI and culture: Culturally dependent responses to AI systems. *Curr. Opin. Psychol.*, Vol 58, 101838.
97. Middleton, S. *et al.* (2024). AI for Defence: Readiness, Resilience and Mental Health. *RUSI J.*, 52–62.
98. Martinez-Martin, N. (2024). A broader approach to ethical challenges in digital mental health. *World Psychiatry*, Vol 23, 394–395.
99. Schoene, A. M. *et al.* (2024). Automatically extracting social determinants of health for suicide: a narrative literature review. *Npj Ment. Health Res.*, Vol 3, 1–11. Nature Publishing Group.
100. Sezgin, E. *et al.* (2024). Behavioral health and generative AI: a perspective on future of therapies and patient care. *Npj Ment. Health Res.*, Vol 3, 1–6. Nature Publishing Group.
101. Robinson, A. *et al.* (2024). Equity in Digital Mental Health Interventions in the United States: Where to Next? *J. Med. Internet Res.*, Vol 26, e59939.
102. Alhuwaydi, A. M. (2024). Exploring the Role of Artificial Intelligence in Mental Healthcare: Current Trends and Future Directions – A Narrative Review for a Comprehensive Insight. *Risk Manag. Healthc. Policy*, Vol 17, 1339–1348.
103. Kuhail, M. A. *et al.* (2024). Human-Human vs Human-AI Therapy: An Empirical Study. *Int. J. Human–Computer Interact.*, Vol 0, 1–12. Taylor & Francis.
104. North, M. (2024). Generative AI is trained on just a few of the world's 7,000 languages. Here's why that's a problem – and what's being done about it. *World Economic Forum*.
105. Children's Commissioner for England (2024). Over a quarter of a million children still waiting for mental health support.
106. Gómez-González, E. *et al.* (2023). Artificial intelligence for healthcare and well-being during exceptional times - A recent landscape from a European perspective. European Commission.
107. Wies, B. *et al.* (2021). Digital Mental Health for Young People: A Scoping Review of Ethical Promises and Challenges. *Front. Digit. Health*, Vol 3, Frontiers.
108. Humphreys, J. *et al.* (2024). *Digital Mental Health Technology: User and Public Perspectives*. Medicines and Healthcare products Regulatory Agency.
109. VoiceBox (2023). Coded Companions: young people's relationships with AI chatbots.
110. Brown, I. A. *et al.* (2023). SAFEGUARDING GIRLS AND BOYS: When Chatbots answer their private questions.
111. Habicht, J. *et al.* (2024). Closing the accessibility gap to mental health treatment with a conversational AI-enabled self-referral tool. *Nat. Med.*, 595–602.
112. Lucas, G. M. *et al.* (2014). It's only a computer: Virtual humans increase willingness to disclose. *Comput. Hum. Behav.*, Vol 37, 94–100.
113. Bragazzi, N. L. *et al.* (2023). The Impact of Generative Conversational Artificial Intelligence on the Lesbian, Gay, Bisexual, Transgender, and Queer Community: Scoping Review. *J. Med. Internet Res.*, Vol 25, e52091.

114. Kerrigan, P. *et al.* (2023). Automating vulnerability: Algorithms, artificial intelligence and machine learning for gender and sexual minorities. in *Routledge Handbook of Sexuality, Gender, Health and Rights*. Routledge.
115. D'Arcy, M. and M. (2023). What could AI mean for our money and mental health? *Money and Mental Health Policy Institute*.
116. Ettman, C. K. *et al.* (2023). The Potential Influence of AI on Population Mental Health. *JMIR Ment. Health*, Vol 10, e49936.
117. Gomez-Gonzalez, E. *et al.* (2020). Artificial Intelligence in Medicine and Healthcare: applications, availability and societal impact. Publications Office of the European Union.
118. Obika, D. *et al.* (2024). Safety principles for medical summarization using generative AI. *Nat. Med.*, 1–3. Nature Publishing Group.
119. Williamson, S. M. *et al.* (2024). The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. *Information*, Vol 15, 299. Multidisciplinary Digital Publishing Institute.
120. Pataranutaporn, P. *et al.* (2021). AI-generated characters for supporting personalized learning and well-being. *Nat. Mach. Intell.*, Vol 3, 1013–1022. Nature Publishing Group.
121. Harari, Y. *et al.* (2023). Opinion | You Can Have the Blue Pill or the Red Pill, and We're Out of Blue Pills. *The New York Times*.
122. Verma, G. *et al.* (2022). Examining the impact of sharing COVID-19 misinformation online on mental health. *Sci. Rep.*, Vol 12, 8045. Nature Publishing Group.
123. Bizzotto, N. *et al.* (2023). Buffering against exposure to mental health misinformation in online communities on Facebook: the interplay of depression literacy and expert moderation. *BMC Public Health*, Vol 23, 1577.
124. Corrigan, P. W. (2007). How Clinical Diagnosis Might Exacerbate the Stigma of Mental Illness. *Soc. Work*, Vol 52, 31–39.
125. The British Medical Association (BMA) (2024). Principles for artificial intelligence (AI) and its application in healthcare.
126. British Medical Association (2024). BMA Principles for Artificial Intelligence (AI) and its application in healthcare.
127. Burr, C. *et al.* (2022). Trustworthy Assurance of Digital Mental Healthcare. Zenodo.
128. Rethink Mental Illness (2022). Summary paper: Engaging experts by experience about the role of digital technology in the future of mental health care.
129. Webber, M. *et al.* (2017). A review of social participation interventions for people with mental health problems. *Soc. Psychiatry Psychiatr. Epidemiol.*, Vol 52, 369.
130. Gelso, C. J. *et al.* (2018). The real relationship and its role in psychotherapy outcome: A meta-analysis. *Psychotherapy*, Vol 55, 434–444. Educational Publishing Foundation.
131. Whittlestone, J. *et al.* (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Nuffield Foundation.
132. Hertog, E. *et al.* Data-Driven Parenting: Robust Research and Policy Needed to Ensure that Parental Digital Monitoring Promotes a Good Digital Society. The British Academy.
133. Yunike, Y. *et al.* (2023). The Implications of Utilizing Artificial Intelligence-Based Parenting Technology on Children's Mental Health: A Literature Review.

Poltekita J. Ilmu Kesehat., Vol 17, 1083–1099.
134. Saeidnia, H. R. *et al.* (2024). Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact. *Soc. Sci.*, Vol 13, 381. Multidisciplinary Digital Publishing Institute.
135. Morley, J. *et al.* (2024). The Ethics of AI in Health Care: An Updated Mapping Review. Social Science Research Network.
136. Gooding, P. M. *et al.* (2021). Semi-Automated Care: Video-Algorithmic Patient Monitoring and Surveillance in Care Settings. *J. Bioethical Inq.*, Vol 18, 541–546.
137. Solaiman, B. *et al.* (2023). Monitoring Mental Health: Legal and Ethical Considerations of Using Artificial Intelligence in Psychiatric Wards. *Am. J. Law Med.*, Vol 49, 250–266.
138. Appenzeller, Y. E. *et al.* (2020). Ethical and Practical Issues in Video Surveillance of Psychiatric Units. *Psychiatr. Serv.*, Vol 71, 480–486.
139. Chan, S. C. C. *et al.* (2024). Bridging the equity gap towards inclusive artificial intelligence in healthcare diagnostics. *BMJ*, Vol 384, q490. British Medical Journal Publishing Group.
140. Stade, E. C. *et al.* (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Ment. Health Res.*, Vol 3, 12.
141. Lawrence, H. R. *et al.* (2024). The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Ment. Health*, Vol 11, e59479–e59479.
142. Department of Health and Social Care (2024). Government response to the report of the equity in medical devices: independent review. *GOV.UK*.
143. NHS England (2023). Artificial intelligence (AI) and machine learning.
144. Olawade, D. B. *et al.* (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *J. Med. Surg. Public Health*, Vol 3, 100099.
145. Villongco, C. *et al.* (2020). "Sorry I Didn't Hear You." The Ethics of Voice Computing and AI in High Risk Mental Health Populations. *AJOB Neurosci.*, Vol 11, 105–112. Taylor & Francis.
146. Dellen, E. van (2024). Precision psychiatry: predicting predictability. *Psychol. Med.*, Vol 54, 1500–1509.
147. Lee, E. E. *et al.* (2021). Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, Vol 6, 856–864.
148. Birk, R. H. *et al.* (2022). Digital Phenotyping for Mental Health: Reviewing the Challenges of Using Data to Monitor and Predict Mental Health Problems. *Curr. Psychiatry Rep.*, Vol 24, 523–528.
149. Galderisi, S. *et al.* (2024). Ethical challenges in contemporary psychiatry: an overview and an appraisal of possible strategies and research needs. *World Psychiatry*, Vol 23, 364–386.
150. Hildebrandt, M. (2019). The Issue of Bias. The Framing Powers of Machine Learning. Social Science Research Network.
151. Vayena, E. *et al.* (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Med.*, Vol 15, e1002689. Public Library of Science.
152. Chen, I. Y. *et al.* (2019). Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA J. Ethics*, Vol 21, 167–179. American Medical Association.
153. Verheij, R. A. *et al.* (2018). Possible Sources of Bias in Primary

Care Electronic Health Record Data Use and Reuse. *J. Med. Internet Res.*, Vol 20, e9134.
154. Rajkomar, A. *et al.* (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann. Intern. Med.*, Vol 169, 866–872.
155. Sharpley, M. *et al.* (2001). Understanding the excess of psychosis among the African-Caribbean population in England: Review of current hypotheses. *Br. J. Psychiatry*, Vol 178, s60–s68.
156. Khera, R. *et al.* (2023). Automation Bias and Assistive AI: Risk of Harm From AI-Driven Clinical Decision Support. *JAMA*, Vol 330, 2255–2257.
157. Koutsouleris, N. *et al.* (2022). From promise to practice: towards the realisation of AI-informed mental health care. *Lancet Digit. Health*, Vol 4, e829–e840.
158. Tsipursky, G. (2020). What Is Unconscious Bias (And How You Can Defeat It) | Psychology Today.
159. Money and Mental Health (2023). Policy response on artificial intelligence and machine learning. *Money and Mental Health Policy Institute*.
160. Dame Margaret Whitehead *et al.* (2024). Equity in medical devices: independent review. *GOV.UK*.
161. Verma, S. *et al.* (2018). Fairness definitions explained. in *Proceedings of the International Workshop on Software Fairness*. 1–7. Association for Computing Machinery.
162. Amnesty International (2024). Sweden: Authorities must discontinue discriminatory AI systems used by welfare agency.
163. Amnesty International (2021). Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal.
164. Heikkilä, M. (2022). Dutch scandal serves as a warning for Europe over risks of using algorithms. *POLITICO*.
165. Burgess, M. *et al.* (2023). This Algorithm Could Ruin Your Life. *Wired*.
166. Geiger, G. *et al.* (2023). Suspicion Machines. Lighthouse Reports.
167. Parliamentary question | Rotterdam 'fraud prediction' algorithms automating injustice: Dutch Government violating fundamental rights and the rule of law.
168. Coley, R. Y. *et al.* (2021). Racial/Ethnic Disparities in the Performance of Prediction Models for Death by Suicide After Mental Health Visits. *JAMA Psychiatry*, Vol 78, 726–734.
169. King, D. R. *et al.* (2023). An Introduction to Generative Artificial Intelligence in Mental Health Care: Considerations and Guidance. *Curr. Psychiatry Rep.*, Vol 25, 839–846.
170. David C *et al.* ChatGPT and large language models: what's the risk?
171. Harvey, H. (2024). Are LLM-based ambient scribes and clinical summarisers medical devices? *Hardian Health*.
172. Laestadius, L. *et al.* (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media Soc.*, Vol 26, 5923–5941.
173. Thirunavukarasu, A. J. (2023). Large language models will not replace healthcare professionals: curbing popular fears and hype. *J. R. Soc. Med.*, Vol 116, 181–182. SAGE Publications.
174. Brocki, L. *et al.* (2023). Deep Learning Mental Health Dialogue System. arXiv.
175. Blease, C. *et al.* (2023). ChatGPT and mental healthcare:

balancing benefits with risks of harms. *BMJ Ment Health*, Vol 26, Royal College of Psychiatrists.
176. Crawford, J. *et al.* (2024). 1 in 3 people are lonely. Will AI help, or make things worse? *The Conversation*.
177. Ma, Z. *et al.* (2024). Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. *AMIA. Annu. Symp. Proc.*, Vol 2023, 1105–1114.
178. Maples, B. *et al.* (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *Npj Ment. Health Res.*, Vol 3, 1–6. Nature Publishing Group.
179. Laestadius, L. *et al.* (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media Soc.*, Vol 26, 5923–5941. SAGE Publications.
180. VoiceBox (2024). AI Companion Bots and Explicit Images: Understanding the Risks Posed to Young Users | VoiceBox.
181. De Freitas, J. *et al.* (2024). Chatbots and mental health: Insights into the safety of generative AI. *J. Consum. Psychol.*, Vol 34, 481–491.
182. Asbach, M. *et al.* (2024). AI in Psychiatry: Changing the Landscape of Mental Health Care. *Psychiatr. Times*, Vol 14, 15–17. MJH Life Sciences.
183. (2023). NEDA Suspends AI Chatbot for Giving Harmful Eating Disorder Advice %. *Psychiatrist.com*.
184. Xiang, C. (2023). 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. *VICE*.
185. King, J. (2024). Boy, 14, killed himself after 'falling in love' with AI chatbot. *Metro*.
186. Murdoch, B. (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med. Ethics*, Vol 22, 122.
187. Inkster, B. *et al.* (2023). Cybersecurity: a critical priority for digital mental health. *Front. Digit. Health*, Vol 5,
188. Sai, S. *et al.* (2024). Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies, and Limitations. *IEEE Access*, Vol 12, 31078–31106.
189. Martinez-Martin, N. *et al.* (2021). Ethical Development of Digital Phenotyping Tools for Mental Health Applications: Delphi Study. *JMIR MHealth UHealth*, Vol 9, e27343.
190. Mozilla Foundation (2024). Creepy.exe: Mozilla Urges Public to Swipe Left on Romantic AI Chatbots Due to Major Privacy Red Flags.
191. Simon, D. A. *et al.* (2022). Skating the line between general wellness products and regulated devices: strategies and implications. *J. Law Biosci.*, Vol 9, lsac015.
192. Neporent, L. (2023). BetterHelp Mental Health App Faces $7.8M FTC Fine For Sharing Private User Data. *Psychiatrist.com*.
193. Wyatt, J. *et al.* (2024). Personal Communication.
194. Coghlan, S. *et al.* (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digit. Health*, Vol 9, 20552076231183542.
195. Kumar, A. *et al.* (2024). FTL-Emo: Federated Transfer Learning for Privacy Preserved Biomarker-Based Automatic Emotion Recognition. in *Proceedings of Data Analytics and Management*. (eds. Swaroop, A. et al.) 449–460. Springer Nature.
196. Fauzi, M. A. *et al.* (2022). Comparative Analysis between Individual, Centralized, and

Federated Learning for Smartwatch Based Stress Detection. *J. Pers. Med.*, Vol 12, 1584.
197. Rachakonda, L. *et al.* (2020). SaYoPillow: A Blockchain-Enabled, Privacy-Assured Framework for Stress Detection, Prediction and Control Considering Sleeping Habits in the IoMT. arXiv.
198. Jimenez-Gomez, B. S. (2019). Risks of Blockchain for Data Protection: A European Approach. *St. Clara High Technol. Law J.*, Vol 36, 281–344.
199. Information Commissioner's Office (2023). Guidance on AI and data protection. ICO.
200. Information Commissioner's Office (2022). AI and data protection risk toolkit. ICO.
201. Almond, S. (2023). Generative AI: eight questions that developers and users need to ask. ICO.
202. Information Commissioner's Office (2024). Information Commissioner's Office response to the consultation series on generative AI. ICO.
203. National Cyber Security Centre (2023). Guidelines for secure AI system development.
204. Martin R (2024). Machine learning security principles updated.
205. Catherine H (2024). Cyber Resilience Audit (CRA) scheme launches for assured CAF-based audits.
206. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. Zenodo.
207. WHO (2023). Artificial intelligence in mental health research: new WHO study on applications and challenges.
208. Information Commissioner's Office *et al.* (2020). Explaining decisions made with AI. ICO.
209. Medicines and Healthcare products Regulatory Agency (2024). Machine learning medical devices: transparency principles. *GOV.UK*.
210. Thompson, M. (2024). National Digital Conference 2024 – Chair's Blog. *Digital Leaders*.
211. Sun, J. *et al.* (2023). Artificial intelligence in psychiatry research, diagnosis, and therapy. *Asian J. Psychiatry*, Vol 87, 103705.
212. Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Sci. Eng. Ethics*, Vol 26, 2051–2068.
213. Health Innovation Network (2023). Preventing clinician burnout: could Ambient Voice Technology (AVT) be key?
214. Pham, K. T. *et al.* (2022). Artificial Intelligence and Chatbots in Psychiatry. *Psychiatr. Q.*, Vol 93, 249–253.
215. Jones, C. *et al.* (2023). Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. *Med. Law Rev.*, Vol 31, 501–520.
216. Prictor, M. (2022). WHERE DOES RESPONSIBILITY LIE? ANALYSING LEGAL AND REGULATORY RESPONSES TO FLAWED CLINICAL DECISION SUPPORT SYSTEMS WHEN PATIENTS SUFFER HARM. *Med. Law Rev.*, Vol 31, 1–24.
217. Floridi, L. (2016). Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, Vol 374, 20160112. Royal Society.
218. Goetze, T. S. (2022). Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 390–400. Association for Computing Machinery.
219. Smith, H. *et al.* (2020). Artificial intelligence in clinical decision-making: Rethinking liability. *Med.*

Law Int., Vol 20, 131–154. SAGE Publications Ltd.
220. Milne-Ives, M. *et al.* (2022). Artificial intelligence and machine learning in mobile apps for mental health: A scoping review. *PLOS Digit. Health*, Vol 1, e0000079.
221. Tornero-Costa, R. *et al.* (2023). Methodological and Quality Flaws in the Use of Artificial Intelligence in Mental Health Research: Systematic Review. *JMIR Ment. Health*, Vol 10, e42045.
222. Milne-Ives, M. *et al.* (2020). The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *J. Med. Internet Res.*, Vol 22, e20346.
223. Nagendran, M. *et al.* (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, Vol 368, British Medical Journal Publishing Group.
224. Abd-Alrazaq, A. A. *et al.* (2020). Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J. Med. Internet Res.*, Vol 22, e16021.
225. Governance arrangements for research ethics committees.
226. Graham, S. *et al.* (2019). Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Curr. Psychiatry Rep.*, Vol 21, 116.
227. Nebeker, C. *et al.* (2019). Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Med.*, Vol 17, 137.
228. The Royal Society Science in the age of AI: How artificial intelligence is changing the nature and method of scientific research. The Royal Society.
229. Espie, C. A. *et al.* (2022). Evidence-informed is not enough: digital therapeutics also need to be evidence-based. *World Psychiatry*, Vol 21, 320–321.
230. Arji, G. *et al.* (2023). A systematic literature review and analysis of deep learning algorithms in mental disorders. *Inform. Med. Unlocked*, Vol 40, 101284.
231. Ahmed, A. *et al.* (2022). Overview of the role of big data in mental health: A scoping review. *Comput. Methods Programs Biomed. Update*, Vol 2, 100076.
232. Zhang, T. *et al.* (2022). Natural language processing applied to mental illness detection: a narrative review. *Npj Digit. Med.*, Vol 5, 1–13.
233. Iyortsuun, N. K. *et al.* (2023). A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. *Healthcare*, Vol 11, 285.
234. Xian, X. *et al.* (2024). Debate and Dilemmas Regarding Generative AI in Mental Health Care: Scoping Review. *Interact. J. Med. Res.*, Vol 13, e53672.
235. Morley, J. (2024). AI and the NHS: is it the silver bullet that will improve the health service's productivity? *Nuffield Trust*.
236. Newby, D. *et al.* (2024). Optimising the use of electronic medical records for large scale research in psychiatry. *Transl. Psychiatry*, Vol 14, 1–10. Nature Publishing Group.
237. Herrett, E. *et al.* (2015). Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int. J. Epidemiol.*, Vol 44, 827.
238. Andreotta, A. J. *et al.* (2022). AI, big data, and the future of consent. *AI Soc.*, Vol 37, 1715–1728.
239. Agnew, T. (2023). AI and data: Consent to use of data 'a sticking plaster for broken systems'. *Digital Health*.
240. Evans, B. J. *et al.* (2024). Co-creating Consent for Data Use — AI-Powered Ethics for Biomedical AI.

NEJM AI, Vol 1, AIpc2400237. Massachusetts Medical Society.
241. UK Research and Innovation (UKRI) (2021). New data hub for mental health research.
242. DATAMIND (2024). What's Next for DATAMIND: Shaping the Future of Mental Health Together. *Datamind*.
243. DATAMIND Road Builders. *About Us > Our Focus*.
244. Pauline Whelan *et al.* (2023). A Data Collection Tool to capture a Core Mental Health Dataset within Physical Health Clinical Trials. *figshare*. University of Manchester.
245. Datamind Developing an equity audit tool to understand the representativeness of participants in clinical trials.
246. NHS England (2024). NHS Federated Data Platform (FDP).
247. Health Data Research Innovation Gateway The NHS Research Secure Data Environment (SDE) Network.
248. (2024). London Care Record partners. *OneLondon*.
249. OpenSAFELY About OpenSAFELY.
250. Walker, V. M. *et al.* (2024). COVID-19 and Mental Illnesses in Vaccinated and Unvaccinated People. *JAMA Psychiatry*, Vol 81, 1071–1080.
251. Wellcome Trust (2024). Personal Communication.
252. Schueller, S. M. *et al.* (2018). Discovery of and Interest in Health Apps Among Those With Mental Health Needs: Survey and Focus Group Study. *J. Med. Internet Res.*, Vol 20, e10141.
253. Eis, S. *et al.* (2022). Mobile Applications in Mood Disorders and Mental Health: Systematic Search in Apple App Store and Google Play Store and Review of the Literature. *Int. J. Environ. Res. Public. Health*, Vol 19, 2186. Multidisciplinary Digital Publishing Institute.
254. Larsen, M. E. *et al.* (2019). Using science to sell apps: Evaluation of mental health app store quality claims. *Npj Digit. Med.*, Vol 2, 1–6. Nature Publishing Group.
255. Witte, T. (2022). Is it a wellness app or medical device? A critical boundary issue in the smart wearables sector. *Insights*.
256. Torous, J. *et al.* (2021). The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, Vol 20, 318–335.
257. Medicines and Healthcare products Regulatory Agency (2023). Medical devices: software applications (apps). HM Government.
258. Lewis, T. L. *et al.* (2014). mHealth and mobile medical Apps: a framework to assess risk and promote safer use. *J. Med. Internet Res.*, Vol 16, e210.
259. Gerke, S. *et al.* (2020). The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *Npj Digit. Med.*, Vol 3, 1–4.
260. Hopkin, G. *et al.* (2024). Building robust, proportionate, and timely approaches to regulating and evaluating digital mental health technologies. *Lancet Digit. Health*,
261. NICE (2023). Evidence gaps | Tools and resources | Digitally enabled therapies for adults with anxiety disorders: early value assessment. NICE.
262. Bergin, A. D. G. *et al.* (2023). Identifying and Categorizing Adverse Events in Trials of Digital Mental Health Interventions: Narrative Scoping Review of Trials in the International Standard Randomized Controlled Trial Number Registry. *JMIR Ment. Health*, Vol 10, e42501.
263. Taher, R. *et al.* (2024). Bridging the gap from medical to psychological safety assessment:

consensus study in a digital mental health context. *BJPsych Open*, Vol 10, e126.
264. Taher, R. *et al.* (2023). The Safety of Digital Mental Health Interventions: Systematic Review and Recommendations. *JMIR Ment. Health*, Vol 10, e47433.
265. Taher, R. *et al.* (2024). Developing a process for assessing the safety of a digital mental health intervention and gaining regulatory approval: a case study and academic's guide. *Trials*, Vol 25, 604.
266. Center for Drug Evaluation and Research (2024). International Regulatory Harmonization. *FDA*, FDA.
267. Medicines and Healthcare products Regulatory Agency How to report to the Yellow Card scheme | Making medicines and medical devices safer.
268. Medicines and Healthcare products Regulatory Agency (2023). Software and AI as a Medical Device Change Programme - Roadmap. *GOV.UK*.
269. Medicines and Healthcare products Regulatory Agency (2024). Regulatory Roadmap points the way ahead for new measures to support safe access to medical technology including AI and diagnostics. *GOV.UK*.
270. Members of the Software Team, Healthcare Quality and Access Group at the Medicines and Healthcare products Regulatory Agency (MHRA) (2024). Personal Communication.
271. Beveridge, C. (2024). The EU AI Act: The Key Takeaways.
272. Regan, J. (2024). The EU AI Act: A Beginner's Guide for UK and International Businesses using AI. *CRS*.
273. Szczepański, M. *et al.* (2024). The United Kingdom and artificial intelligence. European Parliamentary Research Service.
274. Office for Digital Health | Digital health | What we do | About. *NICE*. NICE.
275. Medicines and Healthcare products Regulatory Agency (2024). The Innovative Devices Access Pathway (IDAP). *GOV.UK*.
276. Ring, W.-G. (2023). Why We Need a Regulatory Sandbox For AI. *Oxford Law Blogs*.
277. Financial Conduct Authority (FCA) (2022). Regulatory Sandbox. *FCA*.
278. Information Commissioner's Office (ICO) (2024). Regulatory Sandbox. ICO.
279. Medicines and Healthcare products Regulatory Agency (2024). AI Airlock: the regulatory sandbox for AIaMD. *GOV.UK*.
280. Medicines and Healthcare products Regulatory Agency (2024). MHRA trials five innovative AI technologies as part of pilot scheme to change regulatory approach. *GOV.UK*.
281. Medicines and Healthcare products Regulatory Agency (2024). AI Airlock pilot cohort. *GOV.UK*.
282. Center for Devices and Radiological Health (2022). Digital Health Software Precertification (Pre-Cert) Pilot Program. *FDA*. FDA.
283. Alon, N. *et al.* (2020). Assessing the Food and Drug Administration's Risk-Based Framework for Software Precertification With Top Health Apps in the United States: Quality Improvement Study. *JMIR MHealth UHealth*, Vol 8, e20482.
284. Schmidt, L. *et al.* (2024). The three-year evolution of Germany's Digital Therapeutics reimbursement program and its path forward. *Npj Digit. Med.*, Vol 7, 1–8. Nature Publishing Group.
285. Mäder, M. *et al.* (2023). Evidence requirements of permanently listed digital health applications (DiGA) and their implementation in the German DiGA

directory: an analysis. *BMC Health Serv. Res.*, Vol 23, 369.
286. Giebel, G. D. *et al.* (2024). Integration of digital health applications into the German healthcare system: development of "The DiGA-Care Path". *Front. Health Serv.*, Vol 4, Frontiers.
287. Bernard, A. (2024). Responsible Innovation in Practice: NLP for Safeguarding Assistance (Kooth PLC).
288. World Economic Forum (2022). A Blueprint for Equity and Inclusion in Artificial Intelligence.
289. IBM (2022). Value alignment. *IBM Design for AI*.
290. Kaur, D. *et al.* (2022). Trustworthy Artificial Intelligence: A Review. *ACM Comput Surv*, Vol 55, 39:1-39:38.
291. Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Hum.-Comput. Stud.*, Vol 146, 102551.
292. Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, Vol 90, Elsevier.
293. Department for Science, Innovation and Technology *et al.* (2019). Understanding artificial intelligence ethics and safety.
294. IEEE (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition). The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems.
295. Medicines and Healthcare products Regulatory Agency (2021). Good Machine Learning Practice for Medical Device Development: Guiding Principles. *GOV.UK*.
296. Floridi, L. *et al.* (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.*, Vol 28, 689–707.
297. Ktena, I. *et al.* (2024). Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.*, Vol 30, 1166–1173. Nature Publishing Group.
298. Solaiman, I. *et al.* (2021). Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. in *Advances in Neural Information Processing Systems*. Vol 34, 5861–5873. Curran Associates, Inc.
299. Kretzschmar, K. *et al.* (2019). Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support. *Biomed. Inform. Insights*, Vol 11, 1178222619829083. SAGE Publications Ltd STM.
300. Goddard, K. *et al.* (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc. JAMIA*, Vol 19, 121–127.
301. Macrae, C. (2024). Managing risk and resilience in autonomous and intelligent systems: Exploring safety in the development, deployment, and use of artificial intelligence in healthcare. *Risk Anal.*, Vol n/a,
302. Li, H. *et al.* (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *Npj Digit. Med.*, Vol 6, 1–14.
303. Taher, R. *et al.* (2023). The Safety of Digital Mental Health Interventions: Systematic Review and Recommendations. *JMIR Ment. Health*, Vol 10, e47433.
304. Collett, D. (2024). Simulation-based learning. *Learning Environments*.

305. Woollard, J. (2023). Simulation Labs: Creating a Space for Constructive Failure. *Health Innovation Network*.
306. Dhesi, A. S. *et al.* (2022). Insights from developing and evaluating the NHS blood choices transfusion app to support junior and middle-grade doctor decision making against guidelines. *Transfus. Med. Oxf. Engl.*, Vol 32, 318–326.
307. Department for Science, Innovation, and Technology *et al.* (2024). AI Safety Institute releases new AI safety evaluations platform. *GOV.UK*.
308. Jackson, F. (2024). U.K.'s AI Safety Institute Launches Open-Source Testing Platform. *TechRepublic*.
309. Wysa (2024). Wysa Launches Multilingual AI Safety Initiative to Evaluate Large Language Models for Mental Health Support.
310. Wysa Multilingual AI mental health safety.
311. American Hospitals Association (2024). Will AI Help Address Our Behavioral Health Crisis?
312. Balcombe, L. *et al.* (2022). Human-Computer Interaction in Digital Mental Health. *Informatics*, Vol 9, 14.
313. Lederman, R. *et al.* (2021). The Digital Therapeutic Alliance: Prospects and Considerations. *JMIR Ment. Health*, Vol 8, e31385.
314. Thornton, N. *et al.* (2024). AI in health care: what do the public and NHS staff think? The Health Foundation.
315. Salloch, S. *et al.* (2024). What Are Humans Doing in the Loop? Co-Reasoning and Practical Judgment When Using Machine Learning-Driven Decision Aids. *Am. J. Bioeth.*, Vol 24, 67–78. Taylor & Francis.
316. Ho, A. *et al.* (2024). A Holistic, Multi-Level, and Integrative Ethical Approach to Developing Machine Learning-Driven Decision Aids. *Am. J. Bioeth.*, Vol 24, 110–113. Taylor & Francis.
317. McCradden, M. *et al.* (2023). Evidence, ethics and the promise of artificial intelligence in psychiatry. *J. Med. Ethics*, Vol 49, 573–579. Institute of Medical Ethics.
318. Slack, S. K. *et al.* (2023). First-person disavowals of digital phenotyping and epistemic injustice in psychiatry. *Med. Health Care Philos.*, Vol 26, 605–614.
319. Grodniewicz, J. P. *et al.* (2023). Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. *Front. Psychiatry*, Vol 14,
320. Lai, Y. *et al.* (2021). Human-AI Collaboration in Healthcare: A Review and Research Agenda. in
321. Dvijotham, K. (Dj) *et al.* (2023). Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. *Nat. Med.*, Vol 29, 1814–1820. Nature Publishing Group.
322. Hendrix, N. *et al.* (2022). Assessing the Economic Value of Clinical Artificial Intelligence: Challenges and Opportunities. *Value Health*, Vol 25, 331–339.
323. Elliot Jones *et al.* (2024). Under the radar? Examining the evaluation of foundation models. Ada Lovelace Institute.
324. Coiera, E. *et al.* (2024). AI as an Ecosystem — Ensuring Generative AI Is Safe and Effective. *NEJM AI*, Vol 1, AIp2400611. Massachusetts Medical Society.
325. The AI Safety Institute (AISI) About.
326. Blackman, R. (2022). Why You Need an AI Ethics Committee. *Business Ethics*.
327. STANDING Together working group (2023). The Recommendations.
328. Ganapathi, S. *et al.* (2022). Tackling bias in AI health datasets through the STANDING Together

initiative. *Nat. Med.*, Vol 28, 2232–2233.
329. Department of Health and Social Care (2022). UK to pilot world-leading approach to improve ethical adoption of AI in healthcare. *GOV.UK*.
330. Weiergang, F. L. (2020). Denmark: an independent council and a labelling scheme to promote the ethical use of data.

Contributors

POST is grateful to Hannah Gardiner for researching this briefing, to the Nuffield Foundation for funding their parliamentary fellowship, and to all contributors and reviewers. For further information on this subject, please contact the co-author, Natasha Mutebi.

Members of the POST board*

Aynsley Bernard, Kooth

Dr Graham Blackman, University of Oxford

Professor Adriane Chapman, The Governance in AI Research Group (GAIRG)*

Claudia Corradi, The Nuffield Council on Bioethics

Dr David Crepaz-Keay, the Mental Health Foundation*

Fiona Dawson, Mayden

Zoe Devereux, University of Birmingham*

Dr Piers Gooding, La Trobe University

Dr Caroline Green, University of Oxford

Lara Groves, Ada Lovelace Institute

James Heard, The Governance in AI Research Group (GAIRG)*

Dr Gareth Hopkin, Science Policy and Research Programme Team, National Institute for Health and Care Excellence (NICE)*

Dr Becky Inkster, Cambridge University

Dr Grace Jacobs, Kings College London*

Lauren Jerome, Queen Mary University of London

Dr Indra Joshi, Trustee for Lift Schools

Dr Andrey Kormilitzin, University of Oxford
Associate Professor Akshi Kumar, Goldsmiths, University of London

Professor Agata Lapedriza, Northeastern University; Universitat Oberta de Catalunya

Dr Paris Alexandros Lalousis, Kings College London*

Dr Sophia McCully, The Nuffield Council on Bioethics

Dr Rafael Mestre, Southampton University*

Dr Thomas Mitchell, The Governance in AI Research Group (GAIRG)*

Dr Max Rollwage, Limbic*

Dr Annika Marie Schoene, Northeastern University

Julia Smakman. Ada Lovelace Institute*

John Tench, Wysa

Associate Professor Stuart Middleton, Southampton University

Alli Smith, Office for Life Sciences

Mona Stylianou, Everyturn Mental Health*

Dr James Thornton, The Governance in AI Research Group (GAIRG)*

Dr Pauline Whelan, CareLoop*

Dr Gwydion Williams, Wellcome Trust*

Dr James Woollard, Oxleas NHS Foundation Trust, NHS England*

Andy Wright, Everyturn Mental Health

Emeritus Professor Jeremy Wyatt, The Governance in AI Research Group (GAIRG)*

Information Commissioners Office*

Members of the Software Team, Healthcare Quality and Access Group at the Medicines and Healthcare products Regulatory Agency (MHRA)*

MHRA AI Airlock programme team

The Joint Digital Policy Unit (a joint unit between the Transformation Directorate in NHS England, and DHSC)

*denotes people and organisations who acted as external reviewers of the briefing. Some of them were also part of the interview contribution process. Note that contributors are listed in alphabetical order by surname.

The Parliamentary Office of Science and Technology (POST) is an office of both Houses of Parliament. It produces impartial briefings designed to make research evidence accessible to the UK Parliament. Stakeholders contribute to and review POSTnotes. POST is grateful to these contributors.

If you have any comments on our briefings please email post@parliament.uk. Please note that we are not always able to engage in discussions with members of the public who express opinions about the content of our research, although we will carefully consider and correct any factual errors.

If you have general questions about the work of the House of Commons email hcenquiries@parliament.uk or the House of Lords email hlinfo@parliament.uk.

DOI: https://doi.org/10.58248/PN738

Image Credit: Glenn Carstens-Peters

post@parliament.uk
parliament.uk/post
@POST_UK