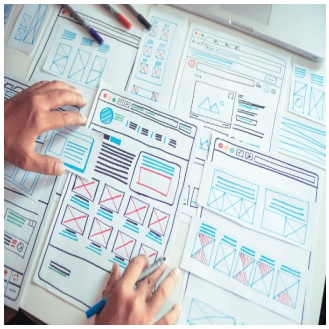
New Framework Addressing Bias, Safety, and Patient Engagement for Conversational AI



Artificial Intelligence (AI) is increasingly utilised in healthcare, offering potential for improved patient outcomes and streamlined clinical processes. Patient-facing conversational AI agents, such as chatbots, engage directly with patients for tasks like symptom diagnosis and appointment scheduling. Despite their benefits, concerns persist regarding their effectiveness, safety, and potential to exacerbate health disparities. Existing frameworks for evaluating healthcare technologies lack specific guidance for addressing the unique challenges of conversational AI. There's a critical need for a tailored framework to ensure these AI tools are designed and implemented equitably, considering cultural competence and inclusivity. Such a framework would identify biases, safety issues, and performance gaps early on, prioritising equity, transparency, and accountability. A recent study published in PLOS Digital Health aims to develop a comprehensive framework guiding developers on equity, diversity, and inclusion issues in healthcare AI implementation.

Roadmap for Equitable Implementation of Conversational AI

A scoping exercise conducted in July 2022 aimed to identify guidelines for implementing patient-facing conversational AI agents in healthcare, emphasising equity, diversity, and inclusion. Initially searching PubMed yielded only three relevant papers, so the search expanded to include grey literature and consultations with key stakeholders, resulting in 17 policy articles. Recommendations from these articles were extracted and categorised into domains such as safety and user involvement, then organised into phases of conversational AI deployment. Subsequent stakeholder consultations involved 33 semi-structured interviews with a diverse range of participants from various sectors and regions. These interviews provided feedback on the recommendations and discussed activities to ensure equity in conversational AI systems. A Public and Patient Involvement (PPI) group also contributed throughout the research process, providing real-world perspectives and validating the findings. Their insights helped refine the roadmap for equitable conversational AI implementation in healthcare. Data were gathered from 33 interviews with key stakeholders from diverse backgrounds in terms of sex, ethnicity and sexual orientation. There were ten community members and 23 industry experts and healthcare professionals (S3 Table for demographic information). The analysis revealed ten significant stages and activities to achieve equity in conversational AI.

Stage 1: Conception and Planning

Stakeholders emphasise the importance of designing conversational AI with equity in mind by identifying public health disparities and tailoring solutions to mitigate them. This includes conducting a 'needs assessment' to recognise health inequalities and determine how conversational AI can address specific conditions affecting marginalised populations. Setting clear behavioural and health outcomes from the outset allows for effective evaluation of the AI's impact on reducing disparities. Defining the role of conversational AI in clinical or administrative tasks helps optimise its functionality for targeted intervention, such as providing tailored health education or facilitating appointment scheduling. Understanding the intended users' characteristics, including aspects of marginalisation and intersectionality, ensures that conversational AI is accessible and engaging for diverse populations. Incorporating culturally relevant information and grounding conversational AI in established behaviour change theories further enhances its effectiveness in influencing users' behaviours and improving health outcomes.

Stage 2: Diversity and Collaboration

Stakeholders stress the importance of involving diverse communities in designing conversational AI tools to mitigate bias and ensure acceptability and usability. They emphasise the need for diverse representation in both design teams and community input to identify potential blind spots and ensure accessibility for disadvantaged groups. Patient and public involvement and engagement (PPIE) groups are essential for influencing the design process and catering to the needs of specific patient groups. Conversational AI should be culturally competent, creating a comfortable environment for discussing sensitive issues and avoiding language that perpetuates stereotypes. Developers must consider the relevance and acceptability of conversational AI to target communities and involve community champions or leaders to address potential mistrust or scepticism. Collaboration with frontline clinicians and patient advocacy groups ensures accuracy, relevance, and safety considerations, enhancing the integration of conversational AI with existing care pathways and services. Involving health professionals in the design and implementation phases enhances their understanding of conversational AI and demonstrates its potential benefits in improving clinical outcomes.

Stage 3: Preliminary Research

Stakeholders highlight the importance of exploring existing conversational AI interventions before developing new ones, as expanding successful and evidence-backed interventions may be more equitable and cost-effective. Feasibility estimation and technical exploration are necessary if an

© For personal and private use only. Reproduction must be permitted by the copyright holder. Email to copyright@mindbyte.eu.

appropriate intervention is unavailable, considering factors such as platform selection and technical complexities. Access to technology is crucial for the success of conversational AI interventions, necessitating an understanding of users' device accessibility. Identifying high-quality training datasets representing diverse communities is essential to avoid biases in healthcare delivery. Developers must stay informed about biases in AI medical devices and adopt mitigation strategies. Preliminary psychosocial and behavioural research helps understand conversational AI desirability, acceptability, and potential barriers to engagement among diverse communities. Addressing "AI hesitancy" and reluctance to disclose health-related information requires tailored approaches. Mixed-methods research may be necessary to ensure a user-centred and community-centred approach to conversational AI development.

Stage 4: Co-Production

Involving specific patient and public groups in conversational AI design and development is crucial, particularly for ethnic minority communities with diverse language needs. Conversational AI tools should accommodate slang, colloquialisms, and diverse grammar and spelling, with reliable translation capabilities for non-English speakers. Native speakers should participate in co-producing and translating scripts to ensure accuracy and comprehension. Co-production should include individuals from diverse social, cultural, and religious backgrounds to understand cultural complexities and social norms. In deciding conversational AI's human-like qualities, diverse user input is essential for determining conversational styles, physical appearance, and persona establishment. Co-production helps refine the chatbot's role and capabilities, making it more accessible for users with physical health problems or disabilities. Alternative delivery modes, such as voice activation, should be considered for users with visual impairment or physical difficulties. Overall, involving diverse groups ensures that conversational AI tools are inclusive, accessible, and culturally sensitive.

Stage 5: Safety Measures

Developing conversational AI for medical and health advice requires careful consideration of ethical considerations to prevent harm or unintended consequences. Accuracy of information is paramount, with regular updates to ensure alignment with medical recommendations. Limitations should be clearly communicated, and users should be encouraged to seek professional medical advice for urgent conditions. Safety protocols should be implemented, including recognising phrases indicating crisis and pathways to appropriate resources. Accountability for safety breaches rests with developers and host organisations, and collaboration with clinical safety officers is crucial for safe implementation. Data security is a significant concern, with developers needing to consider the type of information collected, secure user authentication, and comprehensive privacy policies to protect user data. Additional security measures may be necessary if the conversational AI is hosted on external platforms or apps.

Stage 6: Preliminary Testing

After developing the chatbot's prototype, proof of concept testing is crucial to ensure equity, safety, and efficacy. This testing involves functionality, usability, and conversationality assessments, with input from diverse patient groups. It verifies content, detects technical issues, and identifies potential biases. User feedback is essential for refining the prototype and optimising performance, with emotional responses assessed to gauge engagement. Pilot studies with equity metrics assess behaviour change aligned with defined outcomes, especially for culturally sensitive AI targeting minority groups. Cultural and language sensitivity testing may be necessary, along with strategies to address user hesitancy through collaboration and qualitative research. Ultimately, developers must demonstrate that their conversational AI is effective, safe, and acceptable to users from diverse backgrounds before integrating into healthcare services.

Stage 7: Healthcare Integration

When integrating conversational AI into healthcare organisations, alignment with broader healthcare objectives is crucial. This includes reducing disparities, improving outcomes, and increasing access. Integration testing ensures seamless data exchange, especially with electronic health records. Collaboration with regulatory bodies ensures compliance and necessary approvals. Legal agreements outline responsibilities and procedures for system failure. Staff training is vital to address reservations and ensure effective use. Flexibility is needed for regional adaptation, and scalability requires effective communication and shared learning among healthcare providers.

Stage 8: Service Evaluation and Auditing

Assessing the impact of health conversational AI tools requires thorough evaluation and auditing to promote equity and address health inequalities. Collaboration with academic health science networks can provide valuable support in understanding the broader implications of conversational AI for public health outcomes. Implementers must plan their evaluation approach meticulously, considering factors like time constraints, financial investments, and ethical considerations. Evaluations should focus on specific health and behavioural outcomes, potentially involving factors such as feasibility, patient satisfaction, behaviour change, or influence on health outcomes. Monitoring frequently used features and incorporating user feedback mechanisms are crucial for enhancing utility and effectiveness. Longitudinal data collection may be necessary for measuring patient outcomes accurately. Service-level metrics and demographic data can provide insights into patient uptake and effectiveness in reducing health inequalities. Pre- and post-test studies using anonymous user identifiers can offer valuable insights, and comparative evaluation approaches may help determine acceptability, reach, cost-effectiveness, and equity compared to existing service modalities.

Stage 9: Maintenance

The sustainability of health conversational AI relies on user satisfaction, particularly among minoritised communities, and alleviating healthcare staff workload while maintaining cost-effectiveness. Initial financial investments must be justified by demonstrating potential cost savings and other benefits like enhanced safety and improved staff well-being. Regular updates are crucial to keep conversational AI current with advancements in healthcare and technology, requiring clear guidelines and protocols for execution. Interoperability with other systems in the clinical pathway is essential to ensure smooth operation within the broader healthcare technology ecosystem. Evaluating the impact of conversational AI on healthcare services and staff workload is necessary to address emerging issues and enhance long-term viability. Increasing

© For personal and private use only. Reproduction must be permitted by the copyright holder. Email to copyright@mindbyte.eu.

visibility and awareness of conversational AI among diverse communities, addressing ethical concerns, and involving community champions in promotion efforts are vital for widespread adoption and effectiveness, particularly in addressing health inequalities.

Stage 10: Termination

Developing a health conversational AI requires consideration of scenarios where discontinuation becomes necessary, such as threats to patient health or technology obsolescence. Clear instructions for removal must be provided to host organisations, especially considering integration with patient care pathways. Developers should establish a removal policy, including advance notice periods, to mitigate disruptions and allow users to prepare for alternative arrangements. Communication of termination, guidance on removing the AI from devices, and availability of substitute services are essential. Strategies to address health inequalities post-termination should be devised, and data management plans must ensure compliance with regulations and maintain user trust. Gathering feedback on termination's impact is crucial, particularly among underserved communities.

Key Components for an Equitable Framework for Integrating Conversational AI in Healthcare

The study aimed to develop a comprehensive framework for the equitable integration of health-focused conversational AI into healthcare settings. It emphasised the importance of minimising biases related to language and expressions in conversational AI by involving diverse communities in the design, implementation, and discontinuation processes. Stakeholders stressed the significance of community-driven approaches to AI design, which could enhance inclusivity, acceptability, and engagement, thereby reducing social health inequalities. Key steps outlined in the framework included:

- Inclusive Design Process: Encouraging the active participation of underrepresented communities in the co-production, co-development, and collaboration processes to ensure that conversational AI tools are tailored to diverse user needs and preferences.
- Clear Guidance and Regulation: Emphasizing the need for clear guidance, regulation, and evidence at every stage of conversational AI development and implementation, from conception to maintenance or termination, to ensure fairness, safety, and effectiveness.
- Thorough Evaluation and Validation: Advocating for rigorous evaluation and validation of conversational AI technologies to ensure their safety, effectiveness, and equity before and after deployment in real-world healthcare settings.
- Consideration of Unique Contexts: Acknowledging the diversity of healthcare systems globally and advocating for flexibility and adaptation of the framework's activities to suit the unique needs and constraints of different healthcare ecosystems.
- Collaborative Approach: Highlighting the importance of collaboration between developers, healthcare organisations, communities, regulators, and policymakers to navigate the evolving landscape of AI technologies and ensure their fair and beneficial impact on healthcare.

The framework also addressed challenges and limitations associated with the implementation of conversational AI in healthcare, such as resource constraints, varying levels of tech expertise, and the need for ongoing adaptation and refinement of the framework in response to evolving technologies and healthcare contexts. Overall, the study underscored the importance of fostering transparency, shared learning, and collaboration to maximise the efficacy and impact of conversational AI in addressing health inequities and improving patient outcomes.

Source: **PLOS Digital Health**

Image Credit: **iStock**

Published on : Wed, 15 May 2024

© For personal and private use only. Reproduction must be permitted by the copyright holder. Email to copyright@mindbyte.eu.