Ethical Gaps in Virtual Health Assistants

Ananya Ramesh

Virtual health assistants (VHAs) are AI tools that communicate with patients and clinicians using text or talking. As they become increasingly accepted and used, VHAs can provide patients and clinicians with health information and reminders, along with a foundation for preliminary assessment. These tools have the potential to add significant value to health systems by improving efficiency and access. However, the increasing use of VHAs raises various concerns about privacy, bias, transparency, explainability, and control.

Virtual health assistants are rapidly transforming healthcare delivery by offering continuous, personalised support through the automation of routine queries. With round-the-clock service, VHAs can make healthcare accessible to various people, especially in places with limited resources. However, the deployment also brings in multiple complex ethical, privacy, and technical challenges. This includes protecting sensitive health data, ensuring fairness across diverse populations, maintaining transparency, explainability of AI's decisions, and ensuring there are no data leaks.

This lack of transparency makes it difficult for end-users to trust the AI system they are interacting with, potentially leading to doubt or even dismissal. As the use of AI is increasing exponentially, it is crucial to examine and address these risks. In response, this article proposes a layered, actionable framework that integrates technical, organisational, and ethical strategies to make VHAs both trustworthy and truly inclusive.

The sensitive data collected about patients by VHAs creates vulnerabilities to cyberattack and misuse. The varying jurisdictions and regulations about patient privacy such as HIPAA in the United States and the General Data Protection Regulation (GDPR) in the European Union can complicate compliance and security practices. Additionally, many VHAs are built on datasets that are not fully representative or sufficiently governed and can contribute to the continuation of social biases and discrimination, thereby magnifying inequities in health care, especially for marginalized populations.

Many VHAs primarily act as "black boxes" because they hide their decision-making processes from users and clinicians, which creates a barrier to trust and accountability. Often, patients can have no or limited control over their data or how the recommendations are made and, therefore, provide no informed consent or autonomy over their care. These concerns highlight the urgent need for key stakeholders to work and design VHAs based on ethical justification in transparency and patient-centered care. This helps guarantee safe, equitable, and trustworthy use.

A range of technical and policy strategies have been adopted to address the risks of VHAs. For privacy, most of them rely on conventional security, such as password protection or basic user authentication. In terms of transparency, efforts to enhance it have traditionally involved releasing restricted model documentation or technical white papers. However, these rarely explain the AI's reasoning in a way that end-users or patients can understand. To tackle bias, efforts usually involve just adjusting the data sometimes or checking fairness after the fact. User consent practices are usually static, limited to lengthy terms and conditions at sign-up, and rarely revisited. Even with existing rules, their implementation is inconsistent and not paired with immediate monitoring or public involvement.

Major real-world failures highlight continuous shortcomings in VHA design in the current approaches. IBM Watson for Oncology, for instance, was revealed to have recommended unsafe or incorrect cancer treatments, including suggestions that could have proven fatal, due to reliance on synthetic training data rather than real patient cases and a lack of clinical transparency. Babylon Health's chatbot misdiagnosed or triaged patients inconsistently across demographic groups, sometimes underestimating the severity of female heart attack symptoms, which exposed algorithmic bias and fairness gaps. Similar issues have surfaced with Ada Health and other VHAs, where limited

explainability, insufficient clinical validation, or non-representative training data led to unreliable or misleading outcomes. These incidents show how incomplete technical solutions can seriously compromise patient safety, equity, and public trust.

Infrastructural barriers also limit the success of existing solutions. There is a natural trade-off between how well a model performs and how easily its decisions can be understood. Deeper neural networks tend to be more accurate but are much harder to interpret. Additionally, protecting data privacy while complying with evolving regulations across different regions requires significant effort and resources. Bias mitigation is limited by the high cost and complexity of collecting diverse, high-quality data. Additionally, achieving consensus among clinicians, developers, patients, and regulators about priorities or standards remains complicated. Many VHA deployments lack real-time feedback or adaptive management, resulting in stagnant solutions unable to keep up with evolving societal needs. These ongoing issues underscore the critical importance of comprehensive approaches to ethics and governance in VHAs.

To overcome these complex challenges, such as ethical, privacy, and technical challenges, VHAs need a multi-layered framework that integrates stakeholder input, transparent AI, strong privacy safeguards, and inclusive design to create trustworthy and effective systems. VHAs work best through an inclusive approach by all stakeholders. Some of the stakeholders include patients, healthcare professionals, ethicists, AI developers, and regulatory bodies. Continuous collaboration through co-design, advisory groups, and feedback helps build tools that reflect real needs and values. This process ensures that the development of VHAs is not only driven by technological capabilities but also by ethical considerations, user requirements, and clinical efficacy. It is essential to include marginalised communities to avoid bias. Clearly defining roles and collective accountability promotes transparency and builds trust, and allows VHAs to operate as reliable, user-centric health care tools rather than simple “black boxes”.

Explainable AI (XAI) is critical for promoting trust and accountability in VHA systems by making their decision-making processes transparent and understandable to users, clinicians, and regulators. The proposed solution employs integrated AI systems that combine high-performance models with understandable components. Techniques such as attention mapping, decision trees, and post-hoc explanation methods will help in the translation of complex outputs into understandable information for both patients and clinicians. This also helps healthcare professionals and patients to be assisted by comprehension, which also brings confident utilization of VHA-generated insights. The natural opacity of some AI algorithms, while sometimes unavoidable, necessitates more care in ensuring AI's decisions are correct and linking AI decisions to explicit explanations, especially in high stakes healthcare settings. Regular interdisciplinary evaluations further ensure that the explanations remain accurate and meaningful, bridging the gap between technical complexity and clinical applicability.

To protect patient privacy while maximising data usage, we can use federated learning as an integrated component of the approach. This decentralized training paradigm enables AI models to be trained locally on institutional datasets. Then, only the encrypted model is updated and shared to a central server for aggregation. This reduces exposure of private health information and mitigates the risk of possible breaches. Furthermore, federated learning promotes model generalizability and reduces bias by providing multiple institutions datasets to train on, while still protecting specific data privacy considerations. This method allows for collaborative model training across various healthcare organisations without direct raw data sharing, addressing significant privacy and security concerns associated with centralized data collection. This approach is particularly beneficial for large-sized medical data, such as high-resolution images or 3D scans, as it significantly reduces the overhead of collaborative learning. This method significantly enhances data security by keeping sensitive patient information within local institutional boundaries.

This framework can also incorporate a dynamic consent mechanism that will allow patients to exercise

control and flexibly share their personal health data. This is different from the usual way of getting consent, this approach allows for real-time modification or withdrawal of consent through accessible digital interfaces. Patients can receive timely notifications regarding changes in data usage policies and can tailor their consent preferences for various applications or research purposes. This dynamic process enhances patient reliance, increases transparency, and aligns data governance with changing legal and ethical standards.

As VHAs should work irrespective of who is using them, they must be designed with inclusivity at their core, considerate of diverse linguistic, cultural, and accessibility needs. This involves employing natural language processing models capable of understanding various dialects and accents, offering multimodal interaction options for individuals with disabilities, and interfaces accommodating varied digital literacy levels from the outset. User testing involves individuals across demographic groups, including those with disabilities, to identify and mitigate usability barriers. The ultimate objective is to develop VHAs capable of delivering equitable healthcare experiences across a broad spectrum of users.

Policies and security form the backbone of any trustworthy VHA framework, ensuring that ethical and technical safeguards translate into reliable systems. The security protocol and governance architecture support the proposed approach. Privacy-by-design principles are applied using methods such as data encryption, access control, and privacy protection techniques. Security audits and risk management practices are regular practices to ensure responsiveness to emerging threats. Modular compliance processes allow for compliance with various international regulatory requirements. Clear communication of data policies empowers users and reinforces trust.

Implementing VHAs with integrated XAI, federated learning, dynamic consent, and inclusive design requires a phased, resource-conscious strategy. Stakeholder engagement can operate in three layers: end-users, including patients and clinicians, developers, and policymakers, ensuring a wide range of input from design through deployment. Although VHAs are restricted or banned in some high-risk domains, like healthcare, explainable AI methods (like SHAP or LIME) help reduce concerns, clarify decision logic, and support adoption even in sensitive settings. Federated learning is paired with differential privacy and cryptographic techniques to protect data, with frameworks like PySyft supporting collaboration. Although this integration requires a lot of initial capital, phased implementations that prioritize the ways of doing privacy and explainability can use open-source tools, cloud services and partnerships to reduce these upfront costs. Over time, shared datasets and modular compliance enable these systems to be scalable, efficient, and adaptable, with continuous improvements in XAI. Ongoing feedback and regular evaluation are critical to ensure each phase (in fig. 1) remains effective and relevant.

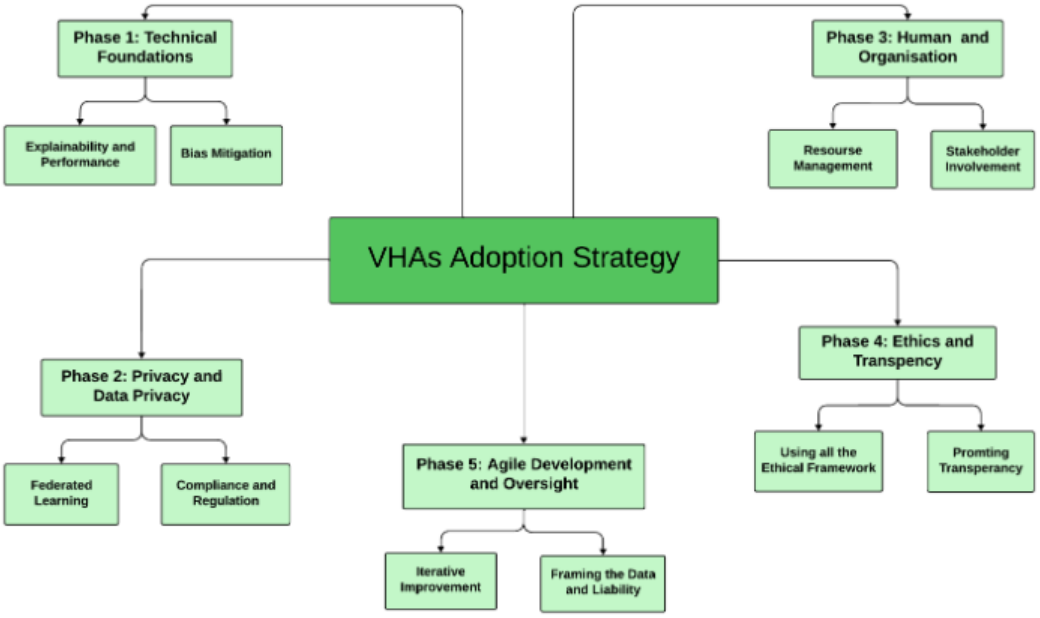

Figure 1: Five-phase strategy for adopting VHAs, integrating technical, privacy, and ethical for effective implementation. It presents a roadmap for VHA adoption.

Phase 1 builds technical foundations addressing bias and explainability; Phase 2 ensures privacy and data governance via federated learning and regulatory compliance; Phase 3 develops human and organisational readiness, resource management, and stakeholder engagement; Phase 4 adopts ethics and transparency through ethical frameworks and transparency; Phase 5 enables agile development and oversight with iterative improvement, data governance, and clarified liability.

Implementing a layered framework for virtual health assistants is expected to lead to significant benefits at multiple levels. For patients, these include higher levels of trust, as they gain greater control over their health data and receive recommendations that are transparent, fair, and tailored to their needs. Clinicians benefit from decision support systems that explain their reasoning, facilitating safer, and more collaborative care. Health systems will have better compliance with existing standards, less bias and error rates, and improved public confidence in digital healthcare solutions. By prioritizing inclusivity and continuous stakeholder inputs, the VHA ecosystem will adapt more rapidly to changing challenges and demographics, especially supporting more equitable health outcomes. By integrating technological advancements with human-centricity, this approach ensures that the potential of artificial intelligence in healthcare is achieved ethically and securely.

Even when the framework exists, there are still some real limitations. First of all, building and maintaining these systems takes time, money, and commitment from many diverse individuals, which can be quite difficult in a limited resource environment. Federated learning and explainable AI offer exciting new possibilities but are also technically challenging and seldom improve system performance compared to simpler models. True stakeholder engagement, especially with stakeholder groups like marginalized voices, can be difficult. Regulations also always seem to evolve, and keeping up with changes is an ongoing task. And, of course, like any configuration of health care technology, there can never be assurances of complete security or being free of all bias. This framework takes us forward, but it won't solve everything, as there will always be new issues as technology changes, and as the human world changes.

Bringing ethics, transparency, and inclusivity into the core of virtual health assistant design is not just a good idea, but it is the key to trust and impact in the real world. This framework offers an approach that mixes technology with real-world experience and governance that is practical and ethical in

healthcare. But establishing and sustaining improvements to VHAs is not a one-time project. Future research should focus on developing dynamic ethical frameworks that can adapt to rapid technological advancements and emerging societal values.

Ultimately, making VHAs truly trustworthy demands not only strong initial frameworks, but continuous real-world monitoring, stakeholder feedback, and a readiness to evolve with advances in both technology and society. The future of digital health is not in short-cuts, but in solutions that are safe, equitable, and truly human-centered.