COMMUNITY PAPER

TRUSTWORTHY AI IN HEALTH: GOVERNING A DOUBLE-EDGED SWORD

A Policy Brief of the Global Health Hub Germany Working Group on AI in Health

Background of this Paper

This policy brief provides a set of timely, strategic recommendations on key governance actions for Germany to ensure trustworthy deployment of artificial intelligence (AI) in health now that the EU AI Act has come into force. The brief draws on the expertise of the Global Health Hub Germany (GHHG) Community's Annual Theme Working Group for 2025/2026, focusing on Artificial Intelligence in Global Health. It is supported by evidence-based insights and best practices from academic research, field reports, regulatory analysis, and stakeholder consultations to inform strategic decision-making and foster inclusive, responsible, and effective use of Artificial Intelligence in health and public health.

About the authors
Written by: Allison Colbert, Ertila Druga, Jana Fehr, Rachel Firestone.
Additional contributions were made by: Michael Bayerlein, Vladimir Choi, Felix Holl, Zahra Karimian, Felipe Mejia Medina, Meenu Singh, Kemna Solveig, Elias Staatz, Franziska Laporte Uribe, Sonali Wayal.

The Working Group was co-hosted by the Global Mental Health and Global Women's Health Hub Communities. Support from the Global Health Hub Germany Secretariate comes from Katrin Würfel and Anya Abanto Graffman.

Executive Summary

Artificial Intelligence (AI) is rapidly transforming healthcare and public health, from diagnostics and clinical decision support to surveillance and health information systems, drug discovery, and patient engagement. Its ability to process vast amounts of data and detect patterns beyond human capacity is significant.

At the same time, these capabilities introduce risks that remain insufficiently understood and governed. Biased or opaque decision-making, algorithmic discrimination, and overreliance on automated systems can cause serious harm in high-stakes clinical and public health contexts. As AI becomes more deeply embedded in health systems worldwide, the challenge is to ensure that implementation is safe, accountable, explainable, and equitable.

AI in healthcare therefore increasingly represents **a double-edged sword**: the same technologies that can expand access and improve outcomes can also amplify harm at scale when governance mechanisms are weak or absent. These tensions are particularly acute in healthcare, where failures in oversight directly affect patient and public safety, public trust, and equitable access to care.

The EU AI Act, which Germany is beholden to and is the first comprehensive horizontal legal framework for AI, establishes whether an AI system is legally permissible. Yet it does not ensure that benefits are distributed equitably across demographic groups or that patients are meaningfully informed of AI involvement in their care.

This policy brief leverages analysis of two health AI deployments in high-income clinical contexts - AI in mammography screening in Germany and mental health triage in the United Kingdom - to identify key gaps between legal compliance and trustworthiness that persist in highly regulated settings and to derive lessons on how AI in healthcare should be assessed, regulated, and adapted to specific contexts.

We then take these lessons and offer strategic, time-sensitive recommendations for the EU AI Act's implementation for healthcare, in Germany and for adaptation in broader health contexts.

Taken together, these recommendations aim to support accountable and equitable AI governance in healthcare. By examining governance gaps in two real-world high-income deployments, the brief identifies practical actions to bridge the gap between legal compliance and trustworthy implementation, particularly as Germany operationalizes the EU AI Act for health.

The brief's recommendations focus on transparency, traceability, evaluation, and evidence infrastructure, and center three priorities:

1. **Keeping patients and clinicians at the center** of AI-supported care
2. **Ensuring meaningful human oversight and public accountability**
3. **Building adaptive governance structures** that maintain safety and performance over time

Introduction

Artificial intelligence is increasingly embedded in everyday decision-making: drafting communications, recommending content, and supporting routine choices. Even in these low-stakes contexts, we recognize its limitations: outputs can be incomplete, biased, or incorrect.

In clinical and public health settings, these risks carry far greater consequences. Small errors can translate into missed diagnoses, inappropriate treatment, or systematic exclusion of patient groups. At the same time, these systems offer substantial promise: automating routine tasks, improving early detection, expanding access to care, and alleviating pressure on overstretched health systems.

This tension defines **AI in health as a double-edged sword:** a technology capable of both advancing and undermining health outcomes, depending on how it is governed, implemented, and monitored.

While AI tools can expand access for marginalized populations – including migrants, refugees, and underserved communities – weak oversight can still create significant governance risks, even in highly-regulated or high-income systems.

This brief analyzes two high-income deployments to identify cross-cutting governance challenges and implementation lessons that become even more important when AI tools are adapted or transferred across contexts. These challenges are relevant to the EU AI Act's implementation in healthcare, both in Germany and its engagements internationally, and flag how potential harm persists even in highly regulated settings in the absence of clear, sector-specific implementation guidance.

The EU AI Act establishes an important legal baseline, requiring documentation, human oversight, and conformity assessments for high-risk systems. However, it does not ensure equitable performance across populations, independent evidence generation, or meaningful patient awareness of AI involvement in care. Across both case studies, important governance challenges persist despite compliance with regulatory requirements.

Where such gaps persist in Europe, they can also be found in other global settings with stronger or lighter regulatory infrastructure and different degrees of available financial resources. Germany therefore cannot credibly co-develop or export trustworthy tools internationally without also working to ensure robust domestic implementation.

To structure the analysis, the brief draws on case studies in women's health and mental health, examining AI deployment across diagnostic and triage pathways, respectively. The cases reflect areas of active engagement within the Global Health Hub Germany community, grounding the analysis in ongoing interdisciplinary dialogue and real-world perspectives. The analysis applies the FUTURE-AI framework, discussed further in the following section, as the primary analytical lens.

Grounded in case-study evidence, the brief offers recommendations to strengthen transparency, oversight, accountability, and implementation capacity in health AI governance under the EU AI Act. Rather than presenting a comprehensive solution, the recommendations offer a pragmatic set of actions to help shape government decision-making and contribute to ongoing strengthening of AI governance in Germany and across its global health partnerships.

Analytical Framework

This brief examines health AI through three complementary frameworks that operate at different levels of analysis.

First, it draws on the WHO guidance on Ethics and Governance of Artificial Intelligence for Health (WHO, 2021), which provides the ethical foundation for trustworthy AI in healthcare and emphasizes autonomy, safety, transparency, accountability, equity, and sustainability.

Second, it applies the FUTURE-AI framework as the primary analytical lens for evaluating the case studies (Lekadir et al., 2025). Developed through international expert consensus, the FUTURE-AI guideline translates broad ethical principles into practical operational criteria for assessing whether health AI systems are trustworthy and deployable in real-world settings. The framework evaluates AI systems across six dimensions of the FUTURE acronym:

- Fairness: equitable performance across different patient groups and populations
- Universality: ability to generalize across settings, populations, and healthcare environments
- Traceability: transparency, documentation, monitoring, and accountability throughout the AI lifecycle
- Usability: safe and effective integration into clinical workflows and user needs
- Robustness: reliable performance under varying real-world conditions
- Explainability: ability for stakeholders to understand how systems function and generate outputs

The analysis then considers the EU AI Act as its regulatory implementation context. While the Act establishes important legal requirements for high-risk AI systems, this brief focuses on how trustworthy AI governance can be operationalized in practice within Germany's healthcare system and global health engagements.

Responsible use of AI in healthcare does not emerge from technology alone. It requires ethical foundations, operational standards, and regulatory implementation. Numerous ethical frameworks have sought to provide that foundation (Jobin et al., 2019) (Beauchamp & Childress, 1979) Additional information on these frameworks is provided in Annex I.

Case Studies

Case Study I – AI in Mammography Screening in Germany

Germany's national mammography screening program[1] is the largest in Europe, inviting approximately 14 million women aged 50–75 for biennial screening.[2] The program uses a double-reading protocol with independent radiological

[1]As of 2025 see BIPS - Mammography screening significantly reduces breast cancer mortality, retrieved from: https://www.bips-institut.de

[2] Bundesamt für Strahlenschutz (BfS). (2022) Brustkrebsfrüherkennung mittels Röntgenmammographie bei Frauen ab 70 Jahren: Wissenschaftliche Bewertung des Bundesamtes für Strahlenschutz gemäß § 84 Absatz 3 Strahlenschutzgesetz https://doris.bfs.de/jspui/

review and escalation procedures for discrepancies.[3]

It is within this double-reading setting that Vara AI is deployed. Vara AI is a CE-certified (Class IIb) AI-based medical device that integrates into existing radiology workflows. Mammogram images are automatically routed to the Vara platform, which returns a risk classification (normal/abnormal) and visual annotations highlighting regions of concern directly to the radiologist's workstation. The radiologist reviews their own independent read alongside the AI output before making the final clinical decision. Vara AI received CE certification as an independent "second reader" in October 2025.

Vara AI certification approval was based largely on the PRAIM study (2021-2023) [4], a prospective, multi-center observational study evaluating performance and productivity effects (Eiseman et al., 2025). The study compared AI-supported double reading with standard double reading and reported higher detection rates, reduced review time for normal cases, and additional cancers identified through AI–human discrepancies, alongside some missed cases by the AI system.

Since July 2025, Vara AI has also been integrated into the QuaMaDi high risk[5] screening program in Schleswig-Holstein.[6] However, deployment details, including audit mechanisms, patient communication and performance monitoring plans, remain largely undisclosed.

Case Study Analysis

F — Fairness: There is currently limited publicly available evidence on Vara's performance across subgroups beyond age and breast density. Evidence across different ethnicities or socioeconomic groups is missing, limiting assessment of equity across populations. Independent fairness assessments are infeasible, due to Vara AI's proprietary architecture.

U — Universality: Evidence is based on European screening contexts, which support some degree of generalizability across clinical sites. Performance evidence is currently being expanded to mammography screening in Egypt and India[7], though deployment details or results are not yet publicly available. The PRAIM study was also co-funded and co-authored by the developer, introducing a potential conflict of interest in the interpretation and reporting of findings. Additionally, deployment remains largely within structured screening environments, and it is unclear how well performance transfers to different health system designs, population risk profiles (e.g., high-risk vs. population-based screening), or lower-resource settings.

T — Traceability: The Vara AI use case does not provide sufficient publicly available information to demonstrate compliance with the traceability principle across the full AI lifecycle. While clinical performance outcomes are published, there is limited transparency on key

[3] Institute for Quality and Efficiency in Health Care (IQWiG9) (2026) The breast cancer screening program in Germany https://www.informedhealth.org/

[4] The study was also registered retrospectively in the German Clinical Trials Register in March 2022, eight months after its start (see German Clinical Trials Register DRKS00027322, retrieved May 15th, 2026 https://drks.de/)

[5] Providing more intensive, often personalized monitoring for asymptomatic individuals with a substantially elevated chance of developing cancer (e.g., those with BRCA gene mutations). In contrast to a population-based screening program which is a broad, public-health initiative that systematically invites average-risk demographics (e.g., based purely on age or sex) for standardized testing. https://www.ncbi.nlm.nih.gov/books/NBK605831/

[6] QuaMaDi (Qualitätsgesicherte Mamma-Diagnostik) is a high-risk breast cancer diagnostic program, distinct from standard population-based screening. It targets women who are referred due to elevated individual risk, including those with a familial history of breast cancer, known high-risk gene variants (e.g., BRCA1/2) or findings such as high mammographic breast density.

[7] For Egypt, 2026; for India, 2024

elements such as comprehensive risk management processes, detailed technical documentation, continuous quality control mechanisms, periodic auditing and model updating, AI logging practices, and post-deployment governance structures. In practice beyond the study, Vara's model evaluation is conducted in-house without an independent audit trail. This lack of accessible information constrains the ability of healthcare providers and researchers to monitor system behavior, assess risks over time, and ensure accountability. In addition, women participating in the screening were not informed of AI involvement in the examination or how it supported their diagnostic results.

U — Usability: Vara demonstrated strong usability within clinical workflows, as it integrates into routine screening, supports radiologist decision-making without replacing it, and reduces reading time for normal cases. In the PRAIM study, radiologists spent 43% less time for AI-tagged normal cases than in standard reading.[8] Vara AI's decision-referral design also aims to limit automation bias by presenting AI alerts only after initial human interpretation, although some residual influence may remain. However these findings have not yet been replicated in lower-volume settings or to less experienced or more burdened radiologists, where the risk of automation bias could be higher. Patient-facing acceptability and end-user experience were not assessed.

R — Robustness: The PRAIM study provides real-world performance data across diverse sites, radiologists, and equipment. However, adequate robustness testing for clinical AI systems of this type remains undefined. There are no established standards for evaluating sensitivity to input variation, scanner differences, or workflow transparency. As the study was industry-funded and co-authored[9] and lacks independent replication, confidence in robustness remains partial.

E — Explainability: Vara displays suspicious regions to the user. There are many layers to explainability[10] and an AI tool should also explain how and why it reached such a conclusion. Vara does not share information on this decision-making with clinicians.

Without this, radiologists cannot meaningfully evaluate the basis of an AI recommendation, which limits the clinical utility of the explanation provided.

Summary:

Vara AI demonstrates promising clinical benefits, but important gaps remain in independent validation, transparency, and long-term performance monitoring.

Key governance lessons:

- **Independent validation remains limited**
- **Patient-facing transparency remains weak**
- **Post-deployment accountability mechanisms remain underdeveloped**

[8] Average reading time per image examination was measured in the AI group only where study states technical impossibility to measure this in the standard reading group (Eisemann, et al., 2025).
[9] Though pharmaceutical companies often use their own studies to justify regulatory approval, that should not be the aspirational model.
[10] Explainability in clinical AI should be understood as multi-layered: (1) visual output - does the system show what it flagged; (2) clinical validity - are flagged regions diagnostically correct; (3) stability - do explanations remain consistent under input variation; and (4) mechanistic transparency - does the system convey why it reached a conclusion. Current regulatory and evaluation frameworks for medical AI largely address only the first layer.

Case Study II: AI deployment in mental health in the UK

The NHS Talking Therapies program provides structured treatment for anxiety and depression across more than 200 services in England. Referrals are voluntary and can be initiated by patients or clinicians. While demand has steadily increased, with referrals nearly doubling between 2012-13 and 2021-22 (Nuffield Trust, 2025), access rates in 2021 were still 22% below expected levels, largely due to workforce shortages constraining intake capacities (The King's Fund, 2024). Evidence indicates that individuals who complete treatment experience better outcomes than those who remain untreated (NHS Digital, 2022).

Limbic Access is a hybrid AI-system that launched on the Talking Therapies website in 2021 to help speed referrals and fill access gaps (Digitalhealth,1). The tool is a rule-based chatbot supported by an all-purpose large language model (LLM) and a cognitive layer architecture[11] to prevent hallucinations. Users interact through a text-based conversational interface, similar to a messaging app, with no avatar or virtual therapist, answering structured questions about their mental health and circumstances. Responses are used to collect intake information and assist clinical classification (referral) before any human clinician is involved.

Evidence suggests improved referral completion and efficiency across multiple observational studies. (Rollwage et al., 2023; Habicht et al., 2024; Rollwage et al., 2024). However, independent evaluation frameworks and standardized transparency mechanisms are not publicly documented.

Case Study Analysis

F — Fairness: Evidence from observational NHS studies suggests that Limbic Access is associated with increased referral completion among nonbinary users (179%) and ethnic minority users (29%), compared to their respective reference groups (Habicht et al., 2024; Rollwage et al., 2024). Qualitative findings also indicate that reduced stigma and the non-human interface may particularly support access for groups facing barriers to help-seeking. However, key evidence needed to substantiate these equity claims is missing. There is no publicly available information on training data representativeness and no formal fairness auditing framework. Though Limbic states it uses a bias detection architecture and structured mitigation procedures, these are proprietary and not shared publicly. It remains unclear whether the observed disparities reflect sustained improvements in equity or context-specific effects of deployment. Notably, offline functionality addresses connectivity barriers relevant even in high-income countries supporting equitable reach in rural or deprived areas.

U — Universality: Evidence from large-scale NHS deployment across multiple services provides real-world validation within a highly resourced, English-speaking health system. While within-system validation may suffice for UK-only use, Limbic is actively seeking international markets.[12] This makes the absence of cross-context validation a material concern, particularly given that language models are sensitive to variations in how distress is expressed across cultural and socioeconomic groups and can be optimized for specific populations, leaving others out. These sensitivities can affect classification accuracy and triage quality and often require adaptations, (Desai & Chaturvedi, 2017), which question the tool's trustworthy and sustainable transferability to other contexts.

T — Traceability: Limbic Access is certified as Class IIa UK medical device (UKCA), which mandates traceable development and risk management processes. However, publicly available information is missing on system architecture, training

[11] https://limbic.ai/research/limbic-layer
[12] Such as the US

data characteristics, model updates or known failure modes. Under frameworks such as the EU AI Act, conversational AI used as a gateway to psychological support would likely be classified as high-risk. This would ideally trigger more transparency obligations than are currently visible. At the same time, requirements on "information to users" under the EU AI Act are still vague and do not specify how traceability and transparency should extend to stakeholders beyond the AI deployers—such as clinicians, patients, and independent evaluators.

U — Usability: Limbic's chatbot-based interface, designed to minimize direct human involvement in the intake process,[13] reduces stigma, increases referral completion, and integrates smoothly into existing NHS intake workflows. Evidence also suggests downstream benefits, including improved attendance and recovery rates when embedded within care pathways (Rollwage et al., 2026). However, usability gains are primarily measured through system-level efficiency and engagement metrics, with limited insight into differential user experience across vulnerable groups.

R — Robustness risks are significant in the context of LLM-based mental health triage. Beyond general vulnerabilities such as hallucination, bias, data leakage, and adversarial prompting (Bernini et al., 2026), Limbic's hybrid rule-based and LLM architecture reduces but does not eliminate risks such as sycophancy, where responses reinforce user beliefs rather than provide clinically grounded guidance. More pressing are failure modes with direct safety implications: misclassification of risk severity, inappropriate triage decisions, and inadequate escalation of crisis presentations While proposed safeguards such as cognitive safety architectures, suicidality detection, and human oversight could mitigate these risks, there is limited evidence that such systems were in place at the time of deployment across the NHS (Rollwage et al., 2026).

E — Explainability represents the most fundamental gap. As an LLM-based conversational system, Limbic Access produces fluent responses without exposing the evidential basis, uncertainty levels, or decision logic behind outputs. This opacity is particularly problematic given Limbic Access's role at the referral entry point, where user disengagement or misdirection may directly affect access to care. At present, no published mechanism allows outcomes, such as drop-off or delayed help-seeking, to be reliably traced back to specific AI interactions.

Summary:

Limbic Access demonstrates promising gains in efficiency and access in NHS Talking Therapies, but important questions remain regarding transparency, independent evaluation, and explainability.

Key Governance Lessons:

- **Access and efficiency gains are promising**
- **Transparency and explainability remain limited**
- **Independent evaluation infrastructure remains insufficient**

[13] Limbic incorporates a human-in-the-loop architecture, meaning clinician oversight is retained at key decision points. The interface minimizes direct human interaction during the patient-facing intake process but is not fully autonomous.

Synthesis

Across both case studies, AI systems demonstrate measurable benefits alongside persistent governance gaps. While the technologies differ substantially in purpose and design, three common findings emerge.

Finding 1: Evidence generation remains heavily dependent on developers.

In both cases, much of the available evidence was generated, funded, or co-authored by system developers. Although such studies provide important insights, limited independent validation and transparency make it difficult to assess performance, fairness, and safety across diverse populations and real-world settings. Stronger public evidence infrastructure is needed to support trustworthy adoption and oversight.

Finding 2: Deployment is occurring faster than governance infrastructure.

Both systems have achieved real-world deployment at scale despite significant gaps in transparency, independent evaluation, and post-market monitoring, including context-specific validation for new deployments. This reflects a broader pattern in health AI: implementation is advancing faster than the governance mechanisms needed to assess long-term performance, identify unintended consequences, and ensure accountability over time.

Finding 3: The EU AI Act establishes legality, but not necessarily trustworthiness.

The EU AI Act provides a meaningful legal baseline for high-risk AI systems through requirements related to documentation, risk management, and human oversight. However, compliance with the EU AI Act's current requirements alone does not guarantee ongoing equitable performance, meaningful transparency, or public trust. The most consequential governance challenges identified in both case studies emerge in this space between legal compliance and trustworthy implementation.

These gaps offer implementation lessons on independent evaluation, transparency, and context-specific validation with implications for Germany's own health system and its role in both global health and international health technology markets. Germany, as one of the WHO's largest governmental contributors and a consistent advocate for stronger multilateral health governance, particularly for digital health and health AI within its G7 presidency and its participation in subsequent G7 health ministerial meetings, can help move international discussions toward shared norms and standards for AI in health.

In addition, as both an adopter and exporter of AI-enabled health tools, Germany operates within transnational technology flows that include deployment in lower-resource settings, some of which have heavier regulatory infrastructure than Europe while others' are more sparse, and among vulnerable populations within its own health system.[14] Particularly when AI systems are transferred to different socioeconomic and cultural contexts, performance cannot simply be transposed, making testing, calibration, and adaptation to the specific context essential, along with solution transparency and oversight beyond market entry approval.

[14] Particularly in migrant and refugee care contexts, AI systems require governance through rights-based safeguards to avoid unequal risk exposure (Council of Europe; 2025).

Why This Matters Now:

- **Patient safety and equity:** Poorly governed AI systems can amplify errors, biases, and unequal outcomes at scale, particularly among vulnerable populations.
- **Public trust:** Trust in health AI depends not only on technical capacity but also on the transparency, accountability, and meaningful patient engagement necessary for good performance long term.
- **European competitiveness:** As AI becomes increasingly central to healthcare innovation, trustworthy governance can become a strategic advantage that strengthens public confidence, supports adoption, and reinforces Germany's leadership in responsible health AI development, in the region and internationally.

Policy Recommendations

The mammography and mental health triage cases showcase complex challenges and opportunities that cut across country contexts. The recommendations below are guided by a simple principle: trustworthy governance enables sustainable innovation. The objective is not to slow the adoption of health AI, but to establish the evidence, transparency, and accountability mechanisms necessary for safe and equitable scaling to different populations. Germany has the regulatory, clinical, and research capacity to lead in this area. The following recommendations translate that capacity to specific governance actions that support operationalizing the EU AI Act for health by strengthening trust and supporting innovation.

Domestic Health AI Governance

Recommendation 1 - Establish standards for patient-facing AI disclosure

Rationale: Both case studies reveal a common governance gap: patients are rarely informed when AI is involved in their care. While the EU AI Act establishes transparency obligations between developers and deployers, it provides limited guidance on patient-facing disclosure. This weakens informed consent, patient autonomy, and public trust. Establishing consistent standards for patient-facing transparency would strengthen accountability while helping ensure that patients remain active participants in AI-supported care.

Operational Actions: The BMG should establish mandatory transparency standards for patient-facing AI disclosure:

- Requiring clear, accessible disclosure to patients when AI is used in their diagnosis, treatment, or care pathway
- Standardizing patient-facing communication formats from updated consent forms to meaningful opt-out options in accessible formats (e.g., digital interfaces and information leaflets)
- Guaranteeing patients a meaningful right to opt-out of AI-supported care where clinically feasible

Recommendation 2: Establish a national health AI registry and evidence infrastructure

Rationale: Effective governance depends on knowing which AI systems are being used, where they are deployed, how they perform, and whether evidence supporting their use remains valid over time. At present, no comprehensive infrastructure exists in Germany to systematically track deployment, monitor outcomes, or support independent evaluation of health AI systems. A national registry and evidence infrastructure would transform the currently fragmented "black box of local deployment" into a transparent and auditable public resource, enabling continuous oversight, post-market learning, and evidence-based decision-making (Fehr, et al., 2024).

Operational actions: The BMG is well positioned to establish a health AI registry infrastructure and open reporting requirements to address this, including:

- Establishing a national public registry (in the absence of an EU-wide register) of clinically deployed AI systems.
- Requiring standardized public reporting for deployers on their system type, intended use, and regulatory status, deployment sites (e.g. hospitals, screening programs, primary care), population coverage, scale of use, updates and version changes over time, and links to evaluation evidence.
- Designating and ideally funding an independent evaluation body (e.g., within existing HTA structures), potentially in collaboration with academic institutions, to conduct validation studies where evidence is predominantly industry generated.
- Establishing post-market surveillance requirements to monitor performance, drift, and unintended consequences over time.

Recommendation 3 – Link public funding and reimbursement to trustworthy AI practices

Rationale: Public funding and reimbursement create powerful incentives that shape AI development and deployment. Yet current frameworks often reward market entry without requiring transparent evidence generation or independent assessment of performance across populations. Selecting public funding and reimbursement beneficiaries based on trustworthy AI practices would encourage stronger evidence standards while building on existing assessment and reimbursement mechanisms rather than creating new regulatory structures.

Operational Actions: As Germany implements the EU AI Act, the BMG has a unique opportunity to embed FUTURE-AI traceability and explainability principles into funding, evaluation, and reimbursement pathways. These could include the following prerequisites to select AI tools for funding and reimbursements:

- Requiring prospective protocol registration as a non-negotiable condition of public funding for AI health pilots and before any data collection begins.
- Mandating publication of dataset documentation and demographic characteristics.
- Requiring independent fairness analyses with publicly available results.
- Integrating these requirements into existing HTA, reimbursement, and funding processes, building on existing structures such as AMNOG rather than creating parallel pathways.

International Health AI Governance

Recommendation 4: Advance international quality standards for health AI deployment
Rationale: Evidence generated in one health system cannot automatically be assumed to transfer to substantially different systems and structures. Differences in language, cultural, health literacy, disease prevalence, clinical workflows, infrastructure, and regulatory capacity can significantly affect AI performance and safety. Yet no widely adopted international mechanism currently requires context-specific validation before health AI systems are deployed in new settings. Without such safeguards, AI tools risk reinforcing inequities, reducing effectiveness, or creating new harms when they transfer across populations and health systems. By establishes context-specific validation standards and fostering quality labels as a market, Germany would help ensure AI systems remain safe, effective, and equitable across its international partnerships and exports. Trustworthy AI quality being offered in German and European markets will encourage similar quality across international contexts.

Operational Actions: the BMG and the BMZ could collaborate to establish inclusive, fairness-based standards for international health AI deployment, particularly in less regulated or lower-resourced settings, including:

- Making published disaggregated performance data across demographic subgroups a condition of international evidence recognition through WHO, bilateral partnerships, and development cooperation agreements
- Developing and requiring a specific safety framework that addresses language robustness, hallucinations in clinical contexts, and cultural calibration of symptom expression before any LLM-based health triage/routing tool is promoted globally.

Recommendation 5: Establish a responsible health AI export framework

Rationale: Health AI systems developed and validated in Europe are increasingly deployed internationally, often in settings with different patient populations, health system structures, languages, and regulatory capacities. Performance demonstrated in one context cannot automatically be assumed to transfer to another. A responsible export framework that systematically incorporates international expertise from the start, including perspectives from Low- and Middle-Income Countries (LMICs), would help ensure that health AI systems remain safe, effective, and equitable when deployed internationally while supporting Germany's broader global health objectives.

Operational Actions: Germany is well placed to leverage collaborative efforts between the BMG, BMZ, and the Federal Ministry of Research, Technology, and Space (BMFTR) and produce an International Health Governance and Responsible Export Framework to ensure that trustworthy health AI deployment in international contexts is commensurate with its domestic policies. This would include actions such as:

- Requiring context-specific validation before international deployments of health AI systems developed in Germany
- Incorporating structured capacity strengthening and technical partnership components into international deployment agreements.
- Requiring documented local adaptation where language, population, or healthcare context differs substantially.

Purpose of this Paper

The work of the Global Health Hub Germany covers various global health issues such as antimicrobic resistance, digital health, climate change, migration and global mental health. This policy paper was developed by the Global Health Hub Germany (GHHG) Annual Theme Working Group 2025/2026, an interdisciplinary group of professionals spanning policy, academia, technology development, health-system strengthening, clinical care, and public health who consider, evaluate, develop, or use health AI within their work.

This paper examines the dual nature of AI in healthcare and public health, offering actionable, grounded recommendations for policymakers and regulators on ensuring the trustworthy deployment of health AI. It does so in explicit dialogue with recent and emerging regulatory developments in Europe and beyond and identifies key building blocks of an evidence-based governance infrastructure for health AI — one that is equitable, accountable, and fit for the complexity of global health systems.

About the Global Health Hub Germany

The Global Health Hub Germany offers all individuals and institutions active in the field of global health the opportunity to connect in an independent network across eight different stakeholder groups: International organisations, youth, politics, foundations, think tanks, business, science, and civil society. The members of the Hub work together on current issues of global health. The interdisciplinary ex- change generates themes, issues and solutions that the Hub brings to policymak- ers to support informed policy-making and advance in global health. Founded in 2019, the Hub now has around 2.500 members. For more information visit www.globalhealthhub.de.

About the Hub Communities

The Hub Communities are working groups led by the members of the Global Health Hub Germany themselves. They meet regularly to exchange ideas, share expertise and work together on global health issues. If you would like to join a Hub Community or learn more about their work, contact Katrin Lea Würfel, Head of Community Management: katrin.wuerfel@globalhealthhub.de.

Published by: Global Health Hub Germany

c/o Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH
Köthener Str. 2-3, 10963 Berlin, Deutschland
Phone: +49 30 59 00 20 210
info@globalhealthhub.de
www.globalhealthhub.de

Version: July 2026

The Global Health Hub Germany is funded by the

Federal Ministry of Health

References

Ethical Framework

- Beauchamp, T. L., & Childress, J. F. (1979). Principles of Biomedical Ethics. Oxford: Oxford University Press.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
- John, S., & Wu, J. (2022). "First, Do No Harm"? Non-Maleficence, Population Health, and the Ethics of Risk. *Social Theory and Practice*. 48(3), 525–551; https://aihcp.net/2024/09/10/understanding-non-maleficence-in-health-care-ethics/
- Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., … Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, *388*, e081554. https://doi.org/10.1136/bmj-2024-081554
- World Health Organization (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. https://iris.who.int/handle/10665/341996. License: CC BY-NC-SA 3.0 IGO

Mammography Case Study

- Bundesamt für Strahlenschutz. (2025, July). Mammography screening considerably reduces breast cancer mortality. Retrieved from https://www.bfs.de/SharedDocs/Pressemitteilungen/BfS/EN/2025/010.html
- Buschmann, L., Bonberg, D. S. N., Karch, A., Brenner, H., Harth, V., Heise, J. K., et al. (2025). Participation in the German mammography screening program: an analysis of data from the NAKO health study. *Deutsches Ärzteblatt International*, 122(24), 655. https://doi.org/10.3238/arztebl.m2025.0156
- Cuocolo, R., Bernardini, D., Pinto dos Santos, D., Klontzas, M. E., Akinci D'Antonoli, T., Semedo, L. C., ... & Williams, M. C. (2025). AI medical device post-market surveillance regulations: consensus recommendations by the European Society of Radiology. *Insights into Imaging*, *16*(1), 275.
- Eisemann, N., et al. (2025). Prospective multicenter observational study of an integrated AI system with live monitoring in mammography screening. *Nature Medicine*, 31, 188–196. https://doi.org/10.1038/s41591-024-03408-6
- Eisemann, N., & Katalinic, A. (2025). Research briefing. *Nature Medicine*, 31, 1422–1423. https://doi.org/10.1038/s41591-025-03714-7
- European Commission. (2025). European Health Data Space Regulation (EHDS). https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en
- Johner Institute. (2025). EU AI Act and medical devices. https://blog.johner-institute.com/iec-62304-medical-software/ai-act-eu-ai-regulation/
- Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: a mapping review. *Social Science & Medicine*, 260, 113172. https://doi.org/10.1016/j.socscimed.2020.113172
- Press releases: Egypt, 2026 (LinkedIn); India, 2024
- Reed Smith. (2025, June). The EU AI Act and medical devices: Navigating high-risk compliance. https://www.reedsmith.com/our-insights/blogs/viewpoints/102kq35/the-eu-ai-act-and-medical-devices-navigating-high-risk-compliance/
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689.

- Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices (Medical Device Regulation). *Official Journal of the European Union*, L 117/1.
- Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space (EHDS). *Official Journal of the European Union*, L 2025/327
- Vara. (2025). PRAIM Study and Company Publications. https://www.vara.ai/
- World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*.

Mental Health Triage Case Study

- AI.gov.uk Knowledge Hub. (2025) Limbic Access: Streamlining Access to Therapy through a chatbot. https://ai.gov.uk/knowledge-hub/tools/limbic-access/
- Aymen Dia Eddine Berini, Norziana Jamil, Ala-Eddine Benrazek, Abderrahmane Lakas, Leila Ismail, Mohamed Amine Ferrag, Kwok-Yan Lam. Security and privacy in LLMs: A comprehensive survey of threats and mitigation strategies, Information Fusion, Volume 132, 2026, 104241, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2026.104241. (https://www.sciencedirect.com/science/article/pii/S156625352600120X)
- Desai G, Chaturvedi SK. Idioms of Distress. J Neurosci Rural Pract. 2017 Aug;8(Suppl 1):S94-S97. https://doi.org/10.4103/jnrp.jnrp_235_17. PMID: 28936079; PMCID: PMC5602270.
- Digital Health London. (2022) Limbic. https://digitalhealth.london/innovation-directory/profile/limbic
- Habicht, J., Viswanathan, S., Carrington, B., Hauser, T. U., Harper, R., & Rollwage, M. (2024). Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nature medicine*, *30*(2), 595–602. https://doi.org/10.1038/s41591-023-02766-x
- Limbic. (2026). The Most Proven AI in Mental Healthcare. https://limbic.ai/
- Limbic. (2026). Using AI to Make Care More Human Not Less: Meet Limbic Layer. https://limbic.ai/layer
- Naddaf M. AI chatbots are sycophants - researchers say it's harming science. Nature. 2025 Nov;647(8088):13-14. https://doi.org/10.1038/d41586-025-03390-0. PMID: 41136779.
- Rollwage M, Habicht J, Juechems K, Carrington B, Viswanathan S, Stylianou M, Hauser TU, Harper R. Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study. JMIR AI. 2023 Dec 13;2:e44358. Https://doi.org/10.2196/44358. PMID: 38875569; PMCID: PMC11041479.
- Rollwage, M., McFadyen, J., Juchems, K., Balogh, A., Pisupati, S., Mircea, M. T., Hauser, T. U., Prichard, G., & Harper, R. (2026). A cognitive layer architecture to support large-language model performance in psychotherapy interactions. *Nature medicine*, 10.1038/s41591-026-04278-w. Advance online publication. https://doi.org/10.1038/s41591-026-04278-w
- Rollwage, M. (2024). Our Commitment to Safety. *Resources: Blog / News,* 18 July 2024. https://limbic.ai/blog/our-commitment-to-safety. Accessed 05 May 2026

Synthesis & Recommendations

- Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: agar transparency of artificial intelligence products in healthcare. Front Digit Health. 2024 Feb 20;6:1267290. Https://doi.org/10.3389/fdgth.2024.1267290. PMID: 38455991; PMCID: PMC10919164.
- https://pmc.ncbi.nlm.nih.gov/articles/PMC3240751/ - Automation bias: Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012 Jan-Feb;19(1):121-7. Https://doi.org/10.1136/amiajnl-2011-000089 Epub 2011 Jun 16. PMID: 21685142; PMCID: PMC3240751.

- Parliamentary Assembly of the Council of Europe. Committee on Migration, Refugees and Displaced Persons. Artificial intelligence and migration: report (Doc. 15952). Strasbourg: Council of Europe; 2025
- Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev. 2017 Feb 16;2(2):MR000033. Https://doi.org/10.1002/14651858.MR000033.pub3. PMID: 28207928; PMCID: PMC8132492.
- Governance and structure of HTA bodies: https://toolbox.eupati.eu/resources/governance-and-structure-of-hta-bodies/#:~:text=In%20general%2C%20HTA%20bodies%20are%20established%20in,to%20administrative%20structures%20for%20a%20health%20system.
- https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality
- https://artificialintelligenceact.eu/high-level-summary/
- https://artificialintelligenceact.eu/national-implementation-plans/
- https://www.gleisslutz.com/en/know-how/federal-government-draft-bill-implement-eu-artificial-intelligence-act
- https://artificialintelligenceact.eu/article/50/
- https://www.technologyslegaledge.com/2025/11/state-of-the-act-eu-ai-act-implementation-in-key-member-states/
- https://go.pharmazie.com/en/pharma-pricing-germany-lt/
- https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf
- https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf
- https://gdpr-info.eu/art-9-gdpr/
- https://healthai.agency/app/uploads/2025/05/UNITE-HealthAIPositionPaper_March2025.pdf
- https://healthai.agency/app/uploads/2025/05/HealthAI_GlobalLandscapeReport_Oct.2024.pdf

Annex I: Analytical Frameworks

Table 1: Analytical frameworks informing the policy brief

Framework/ Instrument	Core Focus	Key Principles/Features	Role in the Policy Brief
FUTURE-AI International Consensus Guideline (Lekadir et al., 2025)	Operational guidance for trustworthy AI in healthcare	Usability, robustness, fairness, universality, traceability, and explainability	Functions as the operational analytical lens for the case study analysis and shapes the recommendations
World Health Organization Guidance on AI for Health (2021)	Ethical governance foundation for AI in global health	Six ethical principles covering autonomy, accountability, equity, transparency, human rights, and governance of AI in health	Serves as an ethical governance guidance in the context of global health, as a well-respected and open-ended resource designed for an audience of international ministries, regulators, companies, and healthcare institutions, many of whom operate in and govern low-resourced health settings
Principles of Biomedical Ethics (Beauchamp & Childress, 1979)	Normative framework for clinical ethics	Autonomy, beneficence, non-maleficence, and justice	Defines cross-cutting ethical obligations healthcare systems owe patients, grounding ethical AI debates including the FUTURE-AI guidance document in a long tradition of biomedical ethics

Annex II: Presentation of the FUTURE-AI and WHO Principles

Two international ethical guidelines for trustworthy / responsible AI in Healthcare:

The WHO guidance establishes what must be upheld in healthcare regarding AI, while the FUTURE-AI framework specifies how AI systems must be built and managed to move from principle to practice.

FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare[15]

Six core principles

1. Fairness

Fairness means that AI tools in healthcare should perform equitably across all individuals and groups, including under-represented and disadvantaged populations, without systematic bias. It calls for identifying, measuring, and mitigating biases in data and model design so that care informed by AI does not worsen existing health disparities. Ensuring fairness enhances trust and supports ethical, non-discriminatory access to AI-enabled healthcare.

2. Universality

Universality emphasizes that healthcare AI should be generalizable and interoperable across different clinical settings, patient populations, and healthcare environments. This principle requires developers to define intended use settings early and validate tools with external and diverse datasets to ensure wide applicability. By supporting adaptability and transferability, universality helps ensure that AI benefits are not limited to narrow clinical scenarios.

3. Traceability

Traceability involves maintaining robust documentation, monitoring, and oversight throughout the AI lifecycle so that development, deployment, outcomes, and potential errors are trackable and auditable. It supports accountability by specifying roles and responsibilities of developers, clinicians, and institutions, and by enabling clear processes to investigate and address problems or harms associated with AI use. Effective traceability fosters trust in AI systems and ensures compliance with ethical and regulatory standards.

4. Usability

Usability means that AI tools must be designed with real users in mind—clinicians, patients, and other stakeholders—ensuring that they can be safely and effectively integrated into clinical workflows. This includes engaging end users early to understand needs, designing intuitive interfaces, and evaluating tools in real-world settings. High

[15] Lekadir K, Frangi A F, Porras A R, Glocker B, Cintas C, Langlotz C P et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare *BMJ* 2025; 388 :e081554 doi:10.1136/bmj-2024-081554

usability enhances adoption, reduces errors, and improves the clinical utility of AI technologies.

5. Robustness

Robustness focuses on the reliability, stability, and resilience of AI systems under varied real-world conditions, including variations in data, equipment, and clinical practices. A robust AI tool should maintain performance across different contexts and be tested rigorously to detect vulnerabilities or failure modes. This principle helps ensure that AI systems are safe, dependable, and resilient over time and across diverse healthcare environments.

6. Explainability

Explainability requires that AI systems be transparent and interpretable to developers, clinicians, regulators, and sometimes patients, so that their functioning and decisions can be understood and questioned. It includes clear documentation of data sources, model logic, assumptions, and limitations, and communication of these in ways suited to different audiences. Explainability supports trust, informed use, and effective oversight of AI in health care.

The FUTURE-AI framework also provides general recommendations:

- **Engage stakeholders continuously –** AI developers should involve diverse stakeholders throughout the AI lifecycle to anticipate needs, identify risks, and support acceptance and responsible adoption.
- **Ensure data protection –** Strong privacy, security, governance, and cybersecurity measures must be applied across the AI lifecycle to protect health data and prevent misuse or reidentification.
- **Implement measures to address AI risks –** Developers should proactively identify and mitigate risks such as bias, lack of robustness, and poor generalizability through appropriate modelling and validation strategies.
- **Define an adequate AI evaluation plan –** AI tools should be evaluated using independent test data, appropriate metrics, and benchmarking against standards or existing solutions to ensure trustworthy performance.
- **Comply with AI regulations –** Applicable AI laws and regulatory requirements should be identified early and followed throughout development and deployment to ensure legal and ethical compliance.
- **Investigate application-specific ethical issues –** Developers should systematically identify and address ethical, social, and societal issues unique to each AI application in healthcare.
- **Investigate social and environmental issues –** The broader social, workforce, and environmental impacts of healthcare AI, including sustainability and carbon footprint, should be assessed and mitigated to ensure positive societal outcomes.

WHO Guidance - Ethics and governance of AI for health[16]

Six key ethical principles to guide the development and use of AI technology for health. While ethical principles are universal, their implementation may differ according to the cultural, religious and other social context.

- **Protect autonomy:** This principle requires that AI systems support, rather than replace, human decision-making in health care, ensuring that clinicians and patients remain in control of medical decisions. Meaningful human oversight must always be present, allowing AI decisions to be monitored, questioned, overridden, or reversed when necessary. Protecting autonomy also includes safeguarding privacy, confidentiality, and informed consent, with AI never used to manipulate or experiment on individuals without valid consent or to restrict access to essential health services.
- **Promote human well-being, human safety, and the public interest:** AI technologies in health care must not cause harm and should meet rigorous standards for safety, accuracy, and effectiveness before and after deployment. Developers, funders, and users have an ongoing responsibility to monitor performance and identify unintended physical, mental, or social harms. The use of AI should prioritize patient welfare and the public interest, particularly by preventing discrimination, stigmatization, or harm resulting from diagnoses or information that patients cannot reasonably act upon.
- **Ensure transparency, explainability and intelligibility:** AI systems should be understandable to developers, users, regulators, and affected individuals through transparency about their design, data, assumptions, and limitations. Explainability should be tailored to different audiences so that people can meaningfully understand and question AI-supported decisions, even when technical complexity creates challenges. Transparent testing, evaluation, and independent oversight are essential to ensure safety, fairness, accountability, and trust in real-world health-care settings.
- **Foster responsibility and accountability:** Although AI systems perform tasks, responsibility for their use and outcomes always rests with human actors and institutions. Clear accountability mechanisms must exist to assign responsibility, provide remedies, and ensure redress when AI systems cause harm. To avoid diffusion of responsibility, all stakeholders involved in the development, deployment, and use of AI should share collective responsibility for minimizing harm and upholding ethical standards.
- **Ensure inclusiveness and equity:** AI in health care should be designed and deployed to promote equitable access and benefit across populations, regardless of age, gender, income, ability, or geographic location. Developers must actively identify, prevent, and mitigate bias in data and algorithms to avoid reinforcing health disparities or discrimination. Inclusive participation, diverse datasets, and continuous monitoring are necessary to ensure AI technologies benefit everyone, especially marginalized and vulnerable groups.
- **Promote artificial intelligence that is responsive and sustainable:** AI systems must be continuously evaluated to ensure they respond appropriately to real health needs and perform as intended in their specific contexts of use. Responsiveness includes the ability to modify, improve, or discontinue AI technologies when they are ineffective, harmful, or unsustainable. Sustainability requires alignment with long-term health system capacity, workforce adaptation, and environmental responsibility, ensuring AI strengthens rather than burdens health systems over time.

[16] World Health Organization (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. https://iris.who.int/handle/10665/341996. License: CC BY-NC-SA 3.0 IGO