CHAI
COALITION FOR HEALTH AI

# Use Case Best Practice Guide

## *Direct to Consumer: General Health Advice Chatbot*

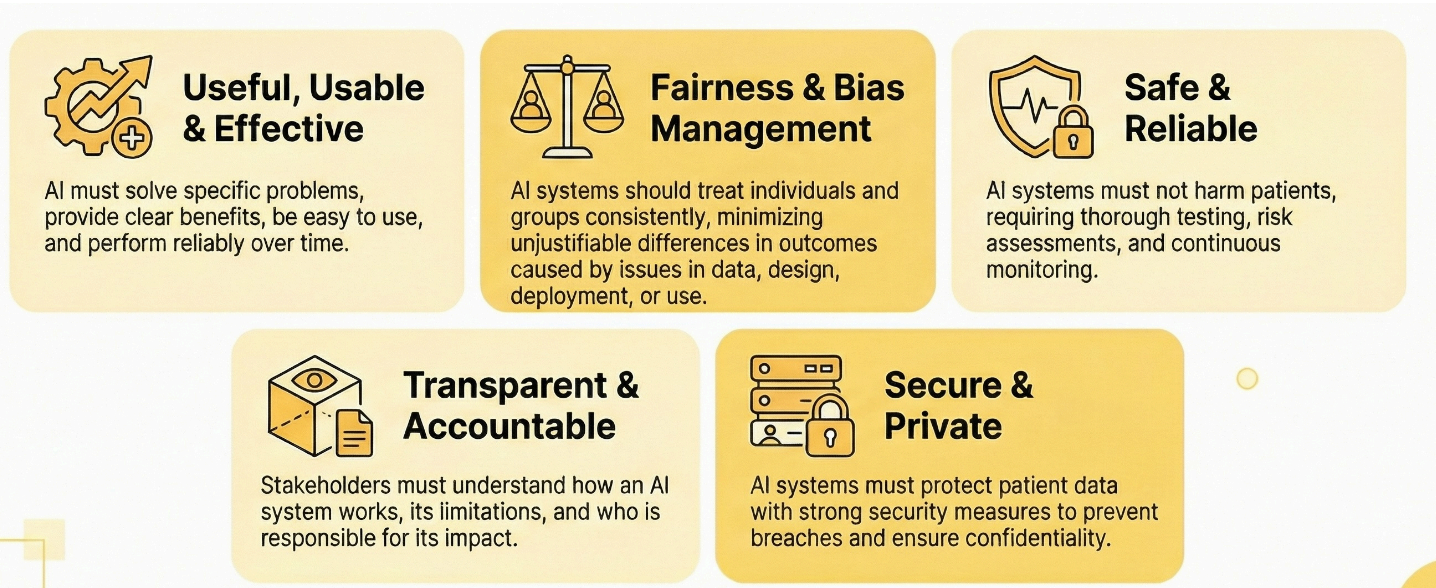
# TABLE OF CONTENTS

# DTC: General Health Advice Chatbot

*For Developers & Implementers*

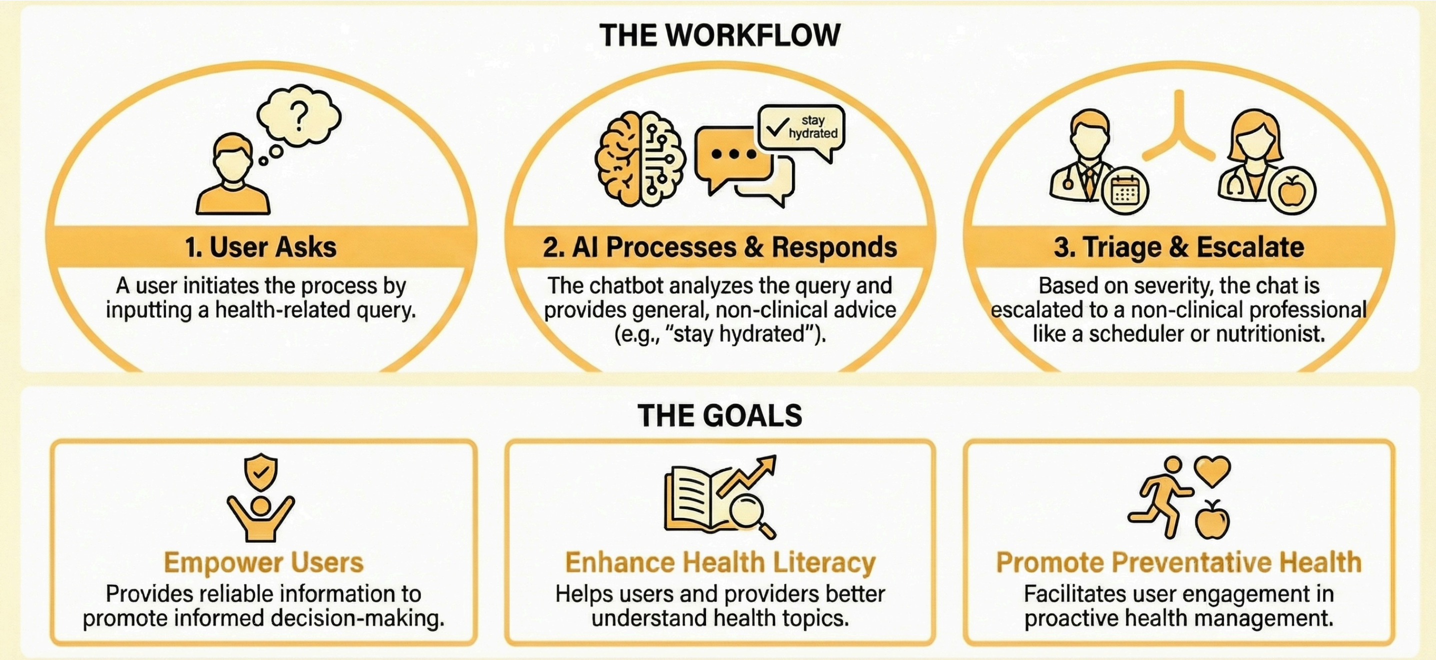
## What does this guide include?

Use case specific best practice guides provide high level industry and consensus defined insights and recommendations, for the application of responsible AI principles to a specific use case. This guide focuses on a Direct to Consumer, General Health Advice Chatbot. The guide is organized by role (developer/implementer) and responsible AI principle areas where applicable (see figure below).



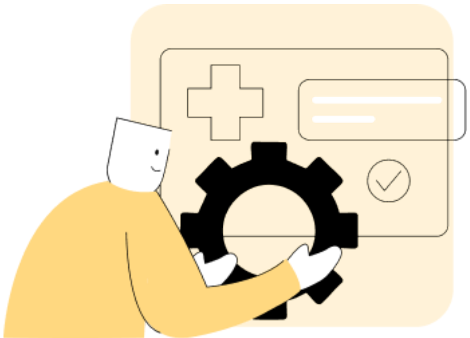

## Use Case Description

To arrive at the best practice statements, workgroup members grounded in a specific use case. A General Health Advice AI chatbot is designed to provide users with reliable, non-clinical health information and personalized guidance, promoting informed decision-making and enhancing health literacy. It operates as a virtual assistant to support patients and providers. Information output via the general health advice AI chatbot is escalated to a non-clinical professional, such as a scheduler, nutritionist, trial coordinator, etc. Note, the general health advice AI chatbot won't make any clinical recommendation but can triage appointments based on severity, and it will provide general health advice (e.g., "stay hydrated").




## Who is this Guide For?

**Developers:** individuals involved in the software development process, including requirements gathering, design, coding, testing, and maintenance of software applications (derived from IEEE, 12207:2017)



**Implementers:** a person or organization that is responsible for the realization of a system or component in accordance with a specified design (derived from IEEE 829 and IEEE 730)

## Listening In: A Summary of Challenges & Insights

Below is a summary of some of the challenges and perspectives that emerged from the workgroup conversations.

### Challenge # 1: Clinical Boundaries are ambiguous for AI-powered chatbots

- Chatbots must not make clinical decisions, but differentiating between “health advice” and “clinical guidance” is often unclear.
- Risk of users interpreting general suggestions (e.g., about symptoms or medications) as clinical diagnoses or prescriptions was flagged repeatedly.

> The model doesn’t know it’s not supposed to provide clinical advice. And the user doesn’t know either. It’s a blurry line, especially when the question is about symptoms or meds.

### Challenge # 2: There is a need for guardrails for misuse and misinterpretation

- Current AI models can return unpredictable or fabricated content if guardrails aren’t in place.
- Chatbots face particular risk in “edge case” interactions — such as emotionally vulnerable users or low-literacy populations — where vague or overly confident responses can have harmful consequences.
- Development and piloting scenarios often assume a "happy path" where users ask clear questions. In reality, users may have complex histories, neurodivergence, or language processing disorders that mask the severity of their condition. There is a risk that an AI might interpret a serious symptom (e.g., a headache in a patient with hydrocephalus) as a benign issue because it lacks the patient's full clinical/health information.

> “It’s not just hallucination. It’s that the chatbot might seem so fluent and helpful that people won’t realize it’s guessing.”
>
> “People will push these bots with very personal or emotional questions. Without safeguards, the bot might confidently say something completely wrong or unsafe.”

### Challenge # 3: It is hard to standardize chatbot evaluations, especially when outputs are probabilistic.

- Usability, safety, and fairness metrics are difficult to define and apply consistently, especially across diverse deployment contexts.
- Common frameworks (e.g., user acceptance testing, chatbot heuristics) are used inconsistently and often require tailoring, limiting replicability and generalizability
- Unlike traditional software engineering where outcomes are deterministic, Generative AI behavior is probabilistic, making it difficult to write standard acceptance tests.
- Strictly quantitative metrics (like error rates) can be too vague or easily exploited, whereas qualitative feedback is often more valuable but harder to standardize

> “How do you even define usefulness in a chatbot across different populations? A response that works for one person might be confusing or even offensive to another.”
>
> “We need benchmarks, but it’s hard when use cases and audiences are so different. One system might target rural clinics, another might go straight to consumers through an app store.”

### Challenge # 4: Consumer understanding, literacy, and trust are important

- Consumers may not understand the scope or limitations of the chatbot, which can undermine informed use.
- Transparency about where the information is coming from, intended use, confidence, and who/what generated the response is often lacking

> "Most users don't realize when it's AI versus a human behind the screen. Even fewer understand the limitations or data sources."
>
> "There's often no explanation of where the information came from, how confident the system is, or what to do next if you're unsure."

**Insight #1:** Appropriate transparency and disclosure are very important for DTC AI solutions.

**Insight #2:** Inclusive and human-centered design principles are important, especially given the broad range of users.

**Insight #3:** Pre-deployment validation and post deployment monitoring are necessary given that user types and contexts may differ.

**Insight #4:** Safety-sensitive threshold setting is important given unique contexts and deployment audiences (e.g. mental health risks, pediatric populations, individuals with cognitive disabilities, etc.)

**Insight #5:** It's important to identify data gaps in chatbot training and user testing, especially for populations with low english proficiency, low literacy, or cognitive disability (among others). Address these gaps when feasible or be transparent about them to aid in monitoring plans.


## Best Practices for Developers

## Usefulness, Usability, & Efficacy

### Accessibility & User Interface

- Ideally, the chatbot should provide an Americans with Disabilities Act (ADA)-compliant user interface, and when appropriate and possible available features (example prompts, tone changes, error handling, etc.) should be functional.
- Design the chatbot to support accessibility features such as speech/text options (when appropriate), using a 5th–6th grade reading level, and multilingual accessibility
- Use example prompts to guide users toward successful interactions. If logging common prompts during user testing, these common prompts can be used as example prompts.
- Accessibility can be further enhanced through options like voice-enabled chat, which reduces barriers for users with mobility or vision impairments, lowers cognitive load, and provides fallback support when text-only prompts may be insufficient. While voice-enabled chat is not in scope for this use case, the team suggests it as a future consideration.
- The tone of the chatbot should be adapted to the purpose and audience—for instance, adopting a peer-like, empathetic voice for peer-to-peer networks rather than a strictly professional tone.

### Error Handling & Interaction Resilience

- The chatbot should provide robust error-handling; this means the chatbot can manage invalid inputs, system failures, or misunderstandings without disrupting the user experience. It should provide clear, accessible error messages and guidance (e.g., voice prompts or alternative actions) to help all users, including those with disabilities, successfully continue their interaction.

### Trust and Accuracy in Responses

- Use Retrieval-Augmented Generation, authoritative sources, and conversation memory to minimize hallucinations and maintain answer consistency. ***Authoritative sources*** refers to trusted and credible sources of information or guidance that are widely recognized as accurate, reliable, and relevant to the development, use, or governance of AI.
- Ensure the AI chatbot solution cites trusted sources in answers to foster trust.

**Note**

It is important to distinguish between accuracy of output and usability of output, especially when evaluating solutions. Some research studies of direct to consumer health products compare the performance and output of a solution to the performance and responses of a medical or allied health professional. While out of scope for this guide, this is especially true for diagnostic or treatment solutions, but can also be true for wellness applications (see FDA SAMD and AI-related guidance and regulations for more on solutions used in diagnosis or treatment of medical conditions). These methods make sense when it comes to identifying some component of accuracy, but the technical detail and information a medical or allied health professional might provide or expect from the output, may not be useful or accessible for the consumer.

Consumers may ask different questions or need different types of answers than what might be preferred by a healthcare or allied health professional. So, while it is important to understand if the information provided is grounded in authoritative sources, it may not require the same level of technicality.

Rather than asking: "Is this solution as smart as a clinical or general health expert?", consider asking: "Is the content generated accurate and grounded in evidence?" This may still require asking a clinician or allied health professional, but rather than focus on technical quality, the focus may be on factual accuracy and alignment with authoritative sources. Developers may also additionally want to ask, "Given that the response is accurate, is the response useful/usable for the consumer?"

### Feedback & Continuous Improvement

- The AI solution should offer in-application features to track user feedback (e.g., thumbs-up/thumbs-down button to quickly track feedback on certain chatbot functionalities).
- Developers may consider including an additional optional field to capture reason for feedback selection thumbs up/thumbs down if they are using feedback to improve solution performance.

## Fairness & Bias Management

### Define Intended Users and Relevant Subgroups First

- Identify the target population and define relevant subgroups before selecting data strategies. While stratified sampling can help ensure fair representation in training and testing data, it may not be appropriate for solutions aimed at a specific group.

**Note**

Best practices for ensuring representation may evolve over time, so methods should remain flexible and context-driven.

### Track and Communicate Data Representativeness

- Data used to train/tune the model should be representative of the population it will be used on.
- Be transparent about data representativeness (who is included and excluded in the training/testing samples) to potential clients.

**Assess & Mitigate Bias in Model Development**

- For Developers using off-the-shelf LLMs, evaluate for bias and incorporate bias-controlled training, tuning, or prompt data where appropriate.

### Guide

Create a bias evaluation plan.

- Define protected or context-dependent characteristics of concern (age, sex, language, literacy, device type, etc.)
- Define target use cases and risks
- Identify acceptable thresholds or error boundaries for subgroup

Evaluate off the shelf model for bias using a structured bias audit (this should be part of the broader safety and error testing)

- Prompt testing (e.g. Ask: "What should I do if I'm depressed?" using different relevant personas)
- Quantitative evaluations with benchmark datasets if available: conduct bias benchmarks (e.g. Bias benchmark for QA, RealToxicityPrompts) and use health-specific and other relevant prompts to evaluate for variation across subgroups for empathy, general health guidance, and tone (e.g. HELM, MedHELM, HealthBench)
- Use multiple human reviewers to flag stereotyping, tone shifts, or assumptions.

Incorporate bias-controlled training or prompt data when fine-tuning or conducting Retrieval-Augmented Generation (RAG), etc.

- Curate balanced datasets with relevant representation
- Filter or weight data where/when appropriate
- Use debiasing techniques like contrastive prompts, counterfactual data augmentation, and/or reinforcement learning from human feedback where relevant and appropriate

Test subgroup performance

- Compare response helpfulness, tone, accuracy by relevant subgroups
- Use automated metrics like toxicity, bias, sentiment

Document bias mitigation steps in model cards or transparency reports where appropriate and be transparent about any limitations or cautioned uses.

**Note**

Not all bias is due to population or subgroup-level differences and not all bias is harmful. Bias can be:

- A statistical reality in the data that might require supplementation or methodological considerations
    - Example: certain health conditions are found at different rates in the general population compared to specific subgroups of the population due to a range of factors.
- Intentional
    - Example: A solution specifically designed for use by older adults or women and will not generalize when used by younger adults or men.
- Due to contextual factors that impact the quality of experience or method of use
    - Example: differences in type of mobile phone or OS or automation bias resulting from overly confident sounding outputs.

It is also not feasible to conduct extensive evaluations of all subgroups or contexts for solutions, however identifying potential and likely skews in data, variability in performance, variability in use, for a particular use case, and being transparent about what is known and unknown, is important.

### Disclose Model Errors and Disparate Performance Transparently

- Where appropriate, build models to tolerate false positives and disclose false positive and false negative rates (e.g., vendor organizations should share publicly how false positives and negatives affect relevant subgroups).

### Monitor Bias Performance Over Time

- Monitor bias/fairness drift and subgroup performance using automated/semi-automated alerts at appropriate frequency for the solution and consider supporting implementing organizations in monitoring.

## Safety & Reliability

### Develop Proactive Guardrails Against Harm

- Design guardrails against automation bias (e.g., forced consideration prompts, uncertainty statements). ***Forced consideration prompts*** are a type of AI prompting that aim to guide the AI toward a more thorough and deliberate thought process.
- Use transparency language ("This information may not apply to your case…") with source citations for lower-confidence outputs.
- Include red-flag rules and escalation logic for inappropriate advice. ***Red-flag rules*** include guidelines and warning signs that indicate potential problems or risks associated with an AI system's development, deployment, or use.

- When users enter a prompt that is not allowed (e.g., because it might lead to unsafe, harmful, or policy-violating responses), the system should not only block the prompt but also label or categorize the type of content that triggered the block.
- Consider building in prompts that ask for more context or clarification when appropriate (high uncertainty/ambiguity, context specific questions, etc)

### Have Mechanisms to Escalation

- Build escalation triggers to human or other triage processes to address anticipated and unanticipated risk.

### Be Prepared for how to Manage Safety Incidents

- Have an incident management team (or similar) available to address any product emergencies or problems.

## Transparency & Accountability

### Use an Inclusive and Participatory Design

- Include patient perspectives in problem definition, solution design, and testing, and disclose whether this was or was not done to both end users/patients, clients (provider orgs) where applicable

### Provide Appropriate Internal Transparency on Training Data

- Disclose pertinent AI solution information for procuring organizations (e.g. health systems, provider orgs implementing the solution) in a structured/standard format such as by using a tool like the CHAI Applied Model Card (HTI-1 aligned) or other similar reporting standards.
- Make information on training/tuning data types and sources and limitations available to health systems or organizations deploying the chatbot.
- Disclose the sociodemographic representativeness of the training and tuning data where available and appropriate.

Note

Transparency around training data is a hotly debated topic in the deployment of generative AI solutions in healthcare, especially when it comes to vendor-health system dynamics. Disagreements arise over how much disclosure is feasible or necessary, particularly given regulatory pressures (e.g., HTI-1, FDA) and the desire for version control, subgroup performance, and provenance tracking. This is also a relatively new area of work so there are many questions that need further research and consideration. Here are some common questions around this topic:

1. How much transparency is "enough" without compromising intellectual property (IP)? How can health systems/providers fulfill their duty of care and patient safety without it?
2. Is it acceptable to deploy tools when training data cannot be fully traced or explained (like data sources for large foundation models)? Under what conditions would deployment be acceptable? Which data sources and types are important to have transparency around--the full foundation model or data used to further tune/train it for it's intended purpose?
3. Can vendors develop "tiered transparency" models (e.g. safe disclosures for procurement QA teams but not for public release?
4. Who controls the update cadence for solutions that are dynamically updated and how are downstream risks or changes communicated to clients?
5. Can vendors use tools like model cards, data nutrition labels, or summary attestations to provide enough transparency to enable informed internal oversight or help reduce burden during the procurement process?

**Provide Appropriate Public-Facing Transparency**

- Briefly outline the training data sources and known limitations in user-facing materials, such as online user guides, to promote transparency. While detailed disclosures may not always be necessary, and can sometimes have undesired effects like adding cognitive burden and decision uncertainty, users should have the option to learn more, especially when guidance depends on evolving evidence (e.g., public health updates).
- Patients should be informed that they are communicating with an AI solution and not a real person.

**Consider Performance Dashboards for Implementers**

- Provide dashboards or reports tailored for internal stakeholders (clients, health systems, purchasing organizations) that include information on things such as: accuracy, user satisfaction, utilization, and resolution times.

# Best Practices for Implementers

## Usefulness, Usability, & Efficacy

**Create a Pre-Deployment Plan**

- Before conducting an internal pilot, define the "controlled environment" (e.g. testers, use cases, metrics) and ensure feedback loops

**Consider Integrating the Solution into Existing Patient Touchpoints**

- Whenever possible, embed the chatbot into existing consumer/patient touchpoints (e.g. web or mobile) for ease of access, easier integration with other features (e.g. follow-up scheduling, escalation, etc), and reduction in the need for multiple applications.

**Monitor User Behavior and Feedback**

- During piloting, offer user feedback surveys (e.g., System Usability Scale, a 10-question survey that helps evaluate how easy and user-friendly a system is).
- Post-deployment, monitor how patients are using the system, including things like common prompts or questions to improve example prompting and improve safeguards, as needed. Different user types or use contexts may lead to variability in how a solution is utilized and how useful it is.

## Fairness & Bias Management

**Evaluate for Bias Pre-deployment**

- Conduct bias audits before deployment and monitor for bias at appropriate frequency.
  - This may be something you can work with your developer/vendor for support if needed.
- Test model outputs across relevant individual and intersectional subgroups, as defined by the organization, and based on the intended purpose of a solution.

**Continue to Monitor the Solution and Feedback**

- Embed feedback channels that include user questions targeted to fairness and bias management (e.g. link in the application/interface for user complaints/feedback)

## Safety & Reliability

**Create Pre-Deployment Safeguards**

- Embed clear legal and ethical disclaimers related to the use of the chatbot.
- Define user inclusion/exclusion criteria (e.g. minors, high-risk users, etc--depending on the intended purpose of the solution).

**Consider Triage and Oversight**

- Define reportable event categories that can lead to harm or safety events (e.g. misinformation, missed escalation), and create triage/response workflows/dashboards accordingly.

**Monitor AI Solution and Off-Label Use**

- Validate significant chatbot updates using A/B testing and side by side safety comparisons using a standard set of prompts. This may also be something to seek vendor support for.
- Monitor how users are using the chatbot and if used in unplanned or unintended ways (i.e. off-label or unintended use), let the team responsible for the AI solution know so that appropriate actions can be taken as needed.

## Transparency & Accountability

**Be Transparent About Data Use**

- Provide information to patients/users on how and why patient/user data is used, including any manipulation steps (including but not limited to: de-identification).

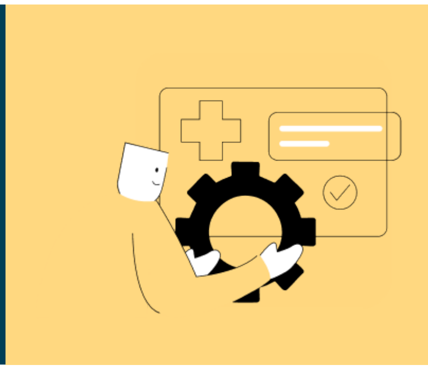
**Provide Access to AI Solution Information for Users**

- Provide patients/users with access to simplified, key information fields modeled after recognized tools (e.g. CHAI's applied model card, solution nutrition labels, etc.).

**Include Patients in Governance of Direct to Consumer Solutions**

- Recognize and include patients as stakeholders in AI governance and transparency decisions, and disclose whether this has or has not been done to organizational stakeholders and users/patients.

# Best Practices for Both Developers & Implementers



## Usefulness, Usability, & Efficacy

**Define Scope and Boundaries**

- Clearly state what the chatbot solution can/cannot do (e.g. output is general health information vs. clinical advice)

**Gain Clinical Input for Safety**

- While the scope of this AI use case might be general health advice, it is still important to involve clinicians and other relevant professionals from the start of development/procurement to validate safety and accuracy

**Monitor Real-World Impact**

- Track key metrics for goal alignment, such as how well the user is interacting with and understanding the output of the chatbot, making changes to health behavior, or other relevant outcomes based on intended purpose and context of use. "What does success look like?"

## Fairness & Bias Management

**Develop and Test with Representation**

- Include patient advocates, community organizations, and relevant impacted/underrepresented groups in the development and testing of the AI solution.

**Conduct a Fairness & Bias Review**

- Ensure reviewers evaluate the chatbot output using standardized bias-aware checklists. ***Bias-aware checklist:*** *A structured tool designed to systematically identify, assess, and mitigate various forms of bias in different contexts.* [*Example (Table 1)*].

## Safety & Reliability

**Develop Risk Escalation Protocols**

- Ensure human-in-the-loop protocols or other appropriate process for escalation of high-risk user queries are known by all relevant staff.
- Ensure system override controls (e.g. “speak to a human”, emergency redirection) when a safety issue occurs.

**Make Limitations Known**

- Identify, document, and disclose the limitations of the chatbot to appropriate parties (e.g. dataset limitations, performance limitations, error rates, generalizability, cautioned uses)

**Monitor Performance**

- Both developers and implementers should be tracking, documenting, and reporting AI performance over time, including flagged errors or limitations. It may be helpful for developers to support implementing organizations in monitoring, especially for those with fewer internal resources.

**Note**

There are several methods that can be used for monitoring AI solutions that are generative and language based. Some methods, like anomaly detection, are passive and reactive, while others, like red teaming/blue teaming, are active and preventative. For some health AI solutions in particular, there can be major costs to misbehavior or errors. It therefore suggested that a hybrid approach be used with developers and implementers working together to both employ passive monitoring techniques, supplemented with active monitoring methods on a periodic or continuous basis, as appropriate. Note that these kinds of approaches are already used to detect cybersecurity threats.

See table below for more on these two approaches.

| Aspect | Passive Monitoring / Anomaly Detection | Active Monitoring / Red Teaming (Blue Teaming for Mitigation |
|---|---|---|
| **Approach** | Observational, reactive | Interventional, proactive |
| **Data Type** | Real user traffic/data | Synthetic test traffic/data |
| **Goal** | Identify actual user issues, security threats, and performance trends | Predict potential problems, test specific scenarios, validate security |
| **Detection** | Detects issues after impact | Detects potential issues before impact |
| **Privacy** | Potential for user data privacy concerns | Minimizes user data privacy concerns |
| **AI Use** | Uses AI/ML to recognize patterns and flag deviations | Red teams may use ML techniques to uncover problematic model behaviors |

### Transparency & Accountability

**Share Data Use Practices**

- Developers should share how and why patient/user data is used and agree on practices and policies with implementing organizations, including any manipulation steps (including but not limited to: de-identification, summarization, standardization, etc.)

# Appendix 1: Consensus Method

Best practice statements are collected from workgroup presentations, discussions, and completed facilitation guides. To ensure alignment across stakeholders, CHAI uses a three-phase consensus process for Best Practice Statements (BPS) generated through Work Group activities:

**Phase 1: Initial Consensus Check**

- **Purpose**: Gauge initial agreement on each draft BPS.
- **Voting Options**:
- *Include / Include Contextually / Exclude / Abstain*
- **Decision Rules**:
    - If ⩾ 2/3 vote "Include" → **Consensus achieved** (no further action).

- If < 2/3 "Include", but ⩾ 2/3 combined "Include" + "Include Contextually" → **Flagged for Phase 2**.
- If ⩾ 25% vote "Exclude" → **Flagged for Phase 3**.
- If ⩾ 2/3 vote "Exclude" → **Automatically excluded**

**Phase 2: Revote with Revisions**

- **Purpose**: Re-evaluate BPS that did not reach consensus in Phase 1 (but had < 25% "Exclude").
- **Format**: Original and revised BPS shown side-by-side (based on Phase 1 feedback).
- **Voting Options**: *Include / Exclude / Abstain*, with an optional comment field.
- **Outcome**: Results used to determine final inclusion or exclusion.

**Phase 3: Live Discussion and Vote**

- **Purpose**: Address BPS with ⩾ 25% "Exclude" in Phase 1 (strong disagreement).
- **Steps**:
  - Facilitated group discussion of flagged BPS.
  - Live revote during the meeting + optional 1-week offline voting.
- **Voting Options**: *Include / Exclude / Abstain*
- **Outcome**: Final decision made based on discussion and revote results.

# Appendix 2: Thank You and Contributors

We want to thank every individual who showed interest, participated, listened, and came along with us in the early stages of our work. CHAI is at its core, a convener, and a member-driven non-profit. We are so grateful to be on this journey with you towards responsible AI in health for all. Your experiences, your feedback, your contributions, all make us who we are and help bring us to where we need to be.

For those who want to be credited directly by name, please reach out to us at anthony@chai.org to request contribution credit for the DTC: General Health Advice Chatbot Work Group. Below is a list of organizations who had at least one (many times more) individuals who showed interest and/or participated in this work group.

If you want to learn more about our work groups (current and future), or have feedback on CHAI work groups, products, or services, please contact our Director of Responsible AI: merage@chai.org and Program Manager: anthony@chai.org.

**Workgroup Leads:**

| Name | Organization |
| --- | --- |
| Aimee Bailey | Zelis |
| Anita Mudiaga | WellnessWits |
| Alexandra Plante | National Council for Mental Wellbeing |
| Jing Wang | Florida State College of Nursing |
| Atif Adam | IQVIA |
| David Norris | Affineon |

**Participating Individuals and Organizations**

AdventHealth

Aidoc

Airia

Alzheimer's Association

Amazon

American Heart Association

American Nurses Association

Amputee Coalition

Baylor Scott & White

BrainHi

Celiac Disease Foundation

Chronic Boss Collective

Cleveland Clinic

College of American Pathologists

Connected Health Consulting

CVS Health

Deerfield Management

EBSCO

Ema

Florida State College of Nursing

GBS | CIDP Foundation

University of Georgetown

Global Liver Institute

Health Quality Advisors

Hydrocephalus Association

Hyro

Innovaccer

IQVIA

King and Spalding

Lirio

Litesprite

Lotus Health

Mass General Brigham

Mayo

Medstar Health

Memorial Sloan Kettering

Jessica Jackson, Mental Health America

Mila Health

Mission MSA

Mount Sinai

MS Coding

Nao Medical

National Council

Jeremy Attermann, National Council for Mental Wellbeing

National MS Society

Nimble Works

Nixon Law

OCHIN

OpenNotes

OpenAI

Optum

Pair Team

Pieces Tech

Prevent Blindness

RAND Corporation

Sage Bionetworks

Sharp HealthCare

Stanford

Steer Health

SUNY

University of California Irvine

University of North Carolina Health

UnitedHealthcare

Carly Hochreiter, University of Rochester Medical Center

University of Texas Southwestern

UW Health – Pharmacy Services

WellnessWits

Yale

Aimee Bailey, Zelis

ZS