Mental Health and AI Field Guide

Scaling evidence-based mental health task-sharing programs with AI

Grand Challenges Canada® Grands Défis Canada x McKinsey Health Institute x Google

Chapter 1: Getting started

What is this field guide?

Access to mental healthcare is limited worldwide, partly because there is a shortage of trained mental health professionals, especially in low-resource settings. However, according to analysis from the McKinsey Health Institute, closing this gap could result in more years of life for people around the world, as well as significant economic gains. Task-sharing models, which allow trained non-mental health professionals to deliver evidence-based mental health services, can be a powerful solution to help bridge this access gap for mental healthcare.

To maintain the quality and sustainability of these models, they require structured processes, robust supervision, and an intensive investment of talent, which can be challenging. Solutions that use AI can help address these challenges. By standardizing training and avoiding the need for a human to be involved at every phase of the process, AI can help mental health task-sharing programs effectively scale evidence-based interventions throughout communities, maintaining a high standard of psychological support. Designing systems with appropriate oversight and diverse perspectives can impact the lives of many more people worldwide.

This field guide introduces a model by which AI could help mental health task-sharing programs scale by supporting specialists and nonspecialists performing their respective roles. It provides foundational knowledge, actionable strategies, and real-world examples to responsibly use AI to effectively scale their work.

The field guide covers several areas, including the following:

- an introduction to AI and large language models (LLMs)
- potential use cases of AI in mental health task-sharing programs and real-world examples
- a capability assessment to determine how ready a task-sharing program is to adopt AI technologies
- examples of technology, quality, safety, trust, and regulatory considerations, as well as of governance that can help reinforce the financial and operational sustainability of the AI solutions
- additional resources for task-sharing programs looking to adopt AI

It is important to acknowledge that the application of AI in task-sharing models is new and only a few pilots have been conducted. Many of the ideas outlined in this field guide are theoretical and have not yet been widely tested in real-world settings. Rather than offering a prescriptive guide, this field guide aims to introduce possibilities and inspire mental health programs to explore AI's role in expanding access to evidence-based interventions.

We invite you to use this resource to inspire opportunities for how AI can be used as a mechanism to accelerate the impact of your programs. We acknowledge that some of the use cases and examples

presented may require additional consideration when it comes to how they compare with the specifics of other countries or circumstances. Our aspiration is to collaborate with others iterating and building out this field guide out to align with the needs and realities of diverse contexts.

Therefore, this field guide is ***not*** intended to do the following:

- provide a step-by-step guide or technical toolkit for developing and implementing specific AI solutions
- advocate for the effectiveness or feasibility of AI solutions in mental health task sharing programs
- advocate for or provide examples of clinical uses of AI
- analyze or document the clinical evidence of particular AI solutions or models

Who is this field guide for?

This field guide is designed for mental health programs that train and deploy individuals to provide support or care in community-based settings. These individuals include volunteers, peer support specialists, laypeople, and licensed clinicians, such as doctors and nurses. The guide is also intended to show that there are viable opportunities requiring funding that could better unlock the potential of task sharing.

This field guide was primarily developed in collaboration with task-sharing programs and, therefore, uses them as the primary use-case example. These programs are specifically designed to address the global shortage in the mental health workforce by expanding access to evidence-based care, particularly in underserved communities; as such, these programs are a good starting point for exploring how AI can aid in scaling the delivery of mental health services.

That said, many of the AI use cases and topics explored throughout the field guide may also be relevant across the broader spectrum of health skill-building programs. For example, they may be relevant for programs that train peer support specialists and volunteers to provide support or that upskill licensed professionals. The first few chapters provide more general background and guidance, while the last few dive deeper into a more specific and specialized set of steps and considerations for those implementing task-sharing programs.

Whether your program is just beginning to explore AI or is already piloting AI solutions, we hope this field guide provides insights to support your journey.

How can you use this field guide?

Although we encourage you to read the entire document, this field guide is structured to be flexible and modular, allowing mental health programs to use it based on their specific needs. You can do any of the following:

- **explore use cases and real-world examples** to identify solutions that could enhance your program.
- **dive into specific chapters for guidance** on technical considerations, safety, and regulatory compliance, governance, and sustainability.
- **apply the frameworks and checklists** to assess your program’s readiness to implement AI, ensure responsible AI use, and develop implementation plans

How was the field guide developed?

The content of the field guide was developed and validated through a number of means:

- **10+ discovery interviews** with experts from health task sharing programs and global institutions
- **an in-person summit that featured panel discussions with more than 20+ experts** from mental health programs—including task-sharing and crisis line or peer support programs—and global institutions
- **5+ real-life case studies** from organizations using AI to improve training, supervision, and intervention quality in mental healthcare
- **a review of 100+ peer-reviewed articles** on mental health task-sharing and the use of AI in healthcare and mental health
- **a review of existing AI frameworks and guidelines in healthcare and mental health** published by leading global organizations (including the World Health Organization, the World Economic Forum, and the Coalition for Health AI)
- **input from 15+ mental health and AI experts** to ensure the applicability and relevance of the content

Who developed this field guide?

This field guide was developed by Google for Health, Google.org, Grand Challenges Canada and McKinsey Health Institute, with contributions from experts at organizations such as Born This Way Foundation, CETA, Child Mind Institute, Children's Hospital of Philadelphia, Fountain House, Friendship Bench, Georgia Tech, Global Mental Health Lab at Columbia University, Harvard T.H. Chan School of Public Health, Jack.org, Jed Foundation, Mental Health America, Mental Health for All Lab at Harvard Medical School, National Council for Mental Wellbeing, Partnership to End Addiction, ReflexAI, Stanford Medicine, StrongMinds, Substance Abuse and Mental Health Services Administration, The Trevor Project, UNICEF, Wellcome, World Health Organization, and 7 Cups.

Acknowledgments

This field guide would not have been possible without the invaluable contributions of leading professionals in mental health, task sharing, and AI. We extend our deepest gratitude to the individuals who participated in discovery interviews, the summit, and feedback rounds to help define main focus areas and refine the field guide's content.

We sincerely thank the following experts:

- **Alex Aide, Esq.,** Born This Way Foundation
- **Anita Everett, MD, DFAPA**, Center for Mental Health Services
- **Arno Klein, PhD**, Child Mind Institute
- **Arvind Sooknanan**, Fountain House
- **Bryan Cheng, PhD**, Global Mental Health Lab, Teachers College Columbia University
- **Daniel Lonnerdal, MS, FACHE**, Substance Abuse and Mental Health Services Administration
- **Dixon Chibanda, MD, PhD**, Friendship Bench, London School of Hygiene and Tropical Medicine

- **Emmett Troxel**, The Trevor Project
- **Elena Netsi, PhD**, Wellcome
- **Emily Feinstein, J.D.**, Partnership to End Addiction
- **Gwydion Williams, PhD,** PATH
- **Helen Verdeli, PhD, MSc**, Teachers College Columbia University
- **Jeff Pennington**, Children's Hospital of Philadelphia
- **John MacPhee**, The Jed Foundation
- **Kenneth Carswell, DClinPsy**, World Health Organization
- **Laura Bond, PhD, MA,** Harvard Medical School, Mental Health for All Lab
- **Laura Murray, PhD**, Johns Hopkins Bloomberg School of Public Health, CETA Global
- **Lauren Fox,** Jack.org
- **Mahmoud Khedr**, Mental Health America
- **Mark van Ommeren**, World Health Organization
- **Melani O'Leary**, Grand Challenges Canada
- **Michael Milham, MD, PhD**, Child Mind Institute
- **Miranda Wolpert, MBE**, Wellcome
- **Munmun De Choudhury, PhD**, Georgia Institute of Technology
- **Nicole Bardikoff**, Grand Challenges Canada
- **Rachel Chernick, PhD**, Partnership to End Addiction
- **Sam Dorison**, ReflexAI
- **Sahil Chopra**, Grand Challenges Canada
- **Sean Mayberry**, StrongMinds
- **Shekhar Saxena, MD**, Harvard T H Chan School of Public Health
- **Stephanie Sasser**, National Council for Mental Wellbeing
- **Trina Dutta**, Substance Abuse and Mental Health Services Administration
- **Vicki Harrison, MSW**, Stanford Center for Youth Mental Health & Wellbeing
- **Zeinab Hijazi, MSc, PsyD**, UNICEF

This field guide is a testament to the collaborative spirit of the mental health and AI communities, and we are grateful to be part of the collective effort to advance responsible AI use in mental health

Chapter 2: Background

Care provider: For the purposes of this material, an individual who completed structured training as part of a task sharing program to deliver mental health support.

Crisis counseling: Services performed by trained volunteers or paid counselors via multiple modalities (such as phone, text, or chat), which may include crisis screening, safety assessments, emotional support, and safety planning. Online or telephonic crisis programs may also connect individuals to follow-up care, referrals, crisis response teams, and emergency services, if needed.

Evidence base: Integration of the best available research with clinical expertise in the context of patient characteristics, culture and preferences.

Mental-health skill-building program: Structured initiative designed to enhance the mental well-being and resilience of individuals, aimed at equipping participants to maintain or improve overall mental health.

Peer support: Programs in which individuals with lived experience of mental illness or substance use, assist others in similar situations, foster engagement, and encourage healthful behaviors through shared understanding and mutual empowerment. This also includes individuals in similar communities. Their services aid in treatment beyond clinical settings to meet people where they are, supporting sustained recovery.

Task sharing: Models that redistribute health tasks to optimize the use of limited resources by delegating specific tasks often performed by clinical specialists to appropriately trained nonspecialists or lay providers. It is a promising way to address global health workforce shortages and insufficient access to care for critical health problems. This strategy has been successfully used for care related to mental health, maternal health, HIV, and immunizations.

Trainee (care provider in training): For the purposes of this material, a trainee is an individual undergoing structured training to deliver mental health support, typically under supervision, as part of a task sharing program.

There are a variety of task-sharing approaches utilized in different contexts. We include general background information in this chapter to capture some of the ways task sharing is implemented; however, we recognize that the use cases outlined do not reflect the reality of all countries or contexts. We hope that these materials can help identify opportunities for additional testing and validation of AI in other types of task-sharing programs across diverse contexts.

An ever-growing shortage of mental health providers exists globally, despite their services being more necessary than ever. More than half of all people will experience a mental health condition in their lifetimes, but there are only 13 mental health providers for every 100,000 people worldwide. Access is even more limited for people in low- and middle-income countries (LMICs), for people in conflict settings, for people with marginalized identities, and for those in rural and low-income communities.

To address this gap, a spectrum of mental health skill-building programs has been developed to train individuals to provide support or care to those in need. These efforts include training volunteers and laypeople to provide care and support and upskilling licensed professionals. For example, there are programs in which individuals with lived experience are trained to provide peer support (that is, peer support programs), programs in which licensed counselors provide support to individuals in crisis (that is, crisis counseling), and programs in which laypeople, nonspecialist health professionals, or community workers are trained to provide evidence-based care (that is, task-sharing programs).

Since task-sharing programs focus on addressing workforce shortages by expanding access to evidence-based care, the following chapters and sections in this field guide focus on these programs and how AI could scale their efforts. At the same time, we acknowledge that many of the AI use cases and topics explored in this field guide may also apply across the broader spectrum of mental health skill-building programs (such as programs that upskill mental health professionals).

Task sharing

Task sharing is an evidence-based solution that increases access to care. In this approach, specialist healthcare professionals delegate specific tasks to trained nonspecialist providers (such as teachers, community health workers, nurses, doctors, auxiliary health staff, and community members) to deliver direct mental health services to the public. This strategy has also been successfully used to deliver care related to maternal health, HIV, and noncommunicable diseases.

There is also strong evidence that task sharing is effective in the mental health domain. Lay providers have been trained in multiple aspects of mental healthcare—from assessment, triage, and engagement to treatment. Empirically supported treatments such as cognitive behavioral therapy (CBT) and interpersonal psychotherapy, for example, have been successfully adapted for delivery by nonspecialist providers by nonprofit organizations that operate across different contexts. Task sharing in the mental health space has been shown to improve clinical outcomes, reduce costs, and extend the reach of a limited mental health clinical workforce. Research finds that mental health task sharing is cost-effective and can both increase the number of people treated for mental health concerns and reduce their mental health symptoms. There also is evidence that task sharing for mental health service delivery is effective in low-resource settings such as LMICs.

Several mental health task-sharing programs have implemented evidence-based methods across different contexts. However, getting lay providers fully trained and confident, maintaining quality (that is, ensuring providers adhere to the program's protocols and evidence-based practices when delivering services), and ensuring sustainability (that is, achieving meaningful scale through long-term financial support) requires structured processes and support of many different kinds.

The opportunity for AI

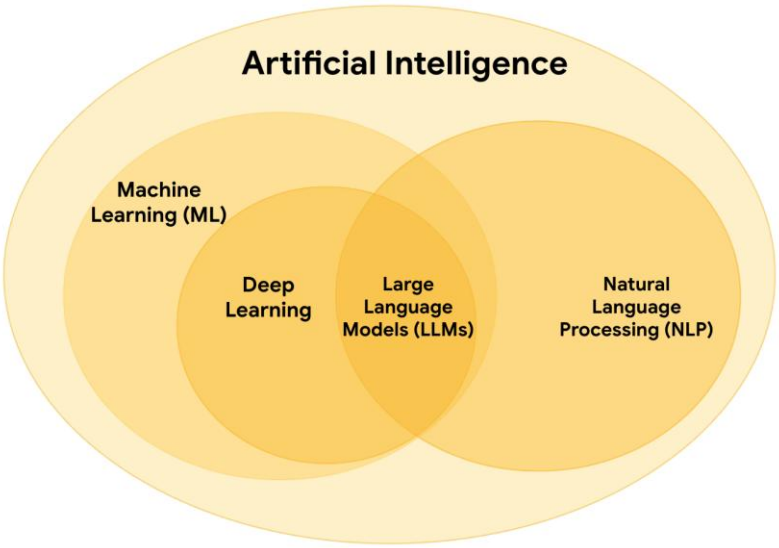
AI offers novel opportunities that can address implementers' challenges while maintaining quality and sustainability and enhancing efficiencies across the implementation cycle of task-sharing programs. By streamlining training, supervision, and decision-making, AI could help task-sharing programs scale up evidence-based interventions while maintaining high standards of model fidelity. Other mental health skill-building programs not focused on task-sharing might also find these AI solutions applicable to their contexts, enabling them to expand their reach and impact.

What is AI?

AI refers to the capability of computer systems to perform complex tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, decision-making, and creating. We recognize that this is a field that is seeing multiple innovations at a fast pace, and that definitions and guidance related to it may rapidly evolve and change.

Gen AI is a subset of AI that focuses on generating new content, including text, images, music, and other media. It uses machine learning, deep learning, and natural language processing to generate outputs that resemble human-created content. Large language models (LLMs) exemplify how gen AI can produce humanlike content by predicting and assembling words in context. For simplicity, the term “AI” is used throughout this field guide and refers to these tools broadly (Exhibit 1).

Exhibit 1



Foundational definitions

- **Machine learning:** A subset of artificial intelligence that focuses on computer algorithms that automatically learn and improve through experience without being explicitly programmed.
- **Deep learning:** A type of machine learning method that uses neural networks (that is, computation models) with multiple processing layers to learn complex patterns in data.
- **Natural language processing:** A field of AI that focuses on providing machines with the ability to understand and communicate using written and spoken human languages.
- **Large language models:** A type of deep learning model trained on massive text data sets to understand and generate natural language—or human-like text—often used in applications such as chatbots and summarization.

How does an LLM work?

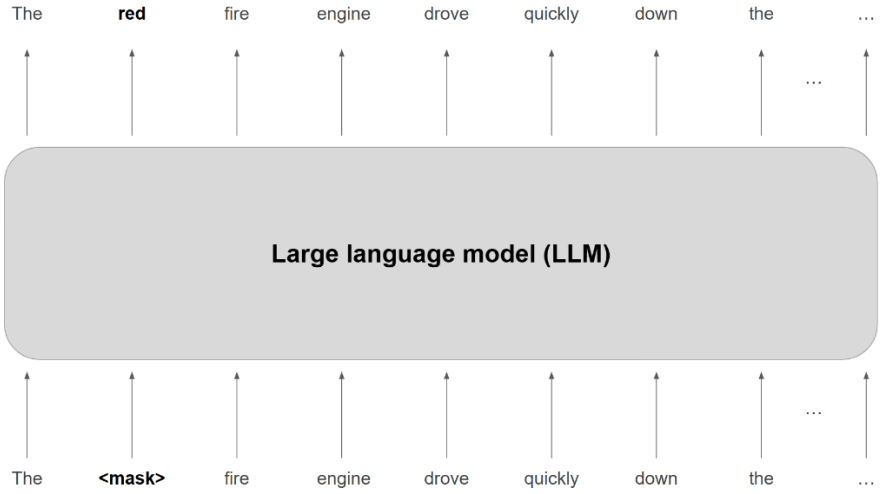
At their core, LLMs are models trained to predict the next word in language sequences or “fill in the blanks.”

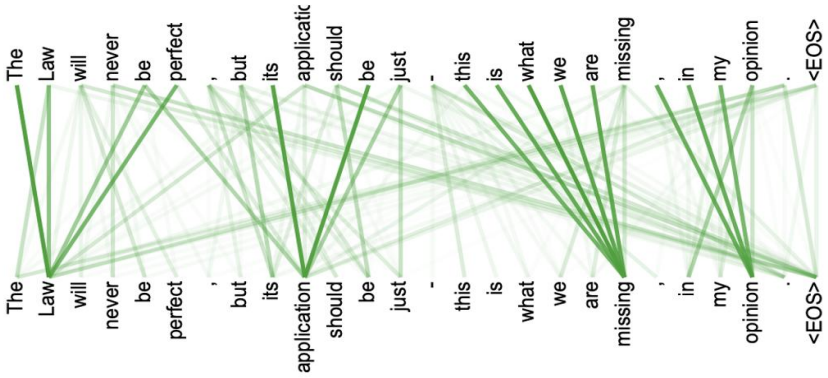
Example

Transformers allow the model LLMs to weigh the importance of different parts of a text and capture context and meaning. For example, as shown in the exhibit below, the transformer model analyzes the sentence

"The red fire engine drove quickly down the" (Exhibit 2). It uses attention mechanisms to connect words that are relevant to each other. This process helps the model understand the relationships between words and generate a more coherent and meaningful output.

Exhibit 2





Source for second image: Ashish Vaswani et al., "*Attention is all you need*," arXiv, updated August 2, 2023

For an LLM to predict the next word in a sequence, it must be trained on vast amounts of data, specifically language-based data. This makes models particularly good at language-based tasks and therefore suitable to support mental health services, which are largely delivered verbally. More recently, AI models have been trained on multimodal data (such as images and audio) and to predict multimodal outputs, expanding their potential mental health use cases. For example, they could eventually be used to design images for imaginal exposures or provide audio-based guidance to clinicians through gentle suggestions.

Using LLMs in mental health settings

The data that LLM models are trained on can come from general sources (such as the internet) or from more tailored sources (such as psychology textbooks or therapy session transcripts). When models are trained on data from general sources, the data set is likely to include some information about mental health,

but that information may not be evidence-based, may be biased, and may perpetuate mental health stigma, which affects the output that the model generates. As such, it is recommended that models be trained specifically to the contexts they will operate in, especially if they are being used in high-risk contexts such as healthcare and mental health.

There are multiple ways to prepare models to be trained to perform mental health tasks. One approach is to train the model using tailored data that is highly curated, evidence-based, representative, and specific to the mental health task the model will perform. Another approach is to take a model that was trained using data from general sources and fine-tune that model for a specific mental health use case using additional, tailored mental health data. Additional approaches—which can be used for models with data from either tailored or general sources—include giving the model "if-then" instructions for the types of outputs they should generate when presented with different mental health scenarios (in other words, prompt engineering) or having human mental health experts rate model output for quality and safety. The ratings would effectively teach the model how best to generate outputs (in other words, reinforcement learning with human feedback).

Models must be thoroughly tested before they are used with real people. This testing should involve a thorough evaluation to ensure that models are safe and unbiased prior to an initial launch. Additionally, safety and bias should be monitored on an ongoing basis.

When thoughtfully developed and responsibly deployed, AI could become a powerful ally in scaling mental health programs. In the next chapter, we explore how AI could enhance task sharing programs, highlighting illustrative AI use cases and real-world examples that also cover peer support and crisis counseling programs.

Chapter 3: Case studies for mental health task sharing programs

This chapter explores how AI could enhance certain task-sharing programs by identifying opportunities to apply these tools across the program and by sharing real-world examples of emerging use cases.

- Section 1 outlines the implementation cycle for scaling a previously tested mental health task-sharing program and highlights where AI could unlock opportunities and strengthen task sharing models.
- Section 2 discusses real-world examples that illustrate how different organizations are piloting or integrating AI into their workflows.

While this chapter was created with evidence-based task-sharing programs in mind, the insights and examples throughout could also apply to the broader spectrum of mental health skill-building programs (such as programs that train volunteers or peer support specialists and programs that upskill mental health professionals).

We also acknowledge that there are a variety of interpretations and uses of task-sharing models and that not all settings will reflect the approaches included in this chapter. Common definitions for stages in implementation cycles are suggested to provide visibility on the kinds of activities that are often included in each specific stage. Real-world examples are included in Section 2, though we recognize these do not reflect the realities of all countries or contexts, where task-sharing approaches may differ and where cultural relevance may limit the use of current AI tools. Organizations can explore minimum viable product versions that are adapted to available technology resources and data availability.

Section 1: Implementing a task-sharing program

Applying a task-sharing program to a mental health context requires a multiphase approach — the implementation cycle. This approach allows programs to be as effective as possible in the context they are being used and allows them to be scaled into more and bigger contexts. As explained in chapter 2, task sharing programs can be delivered through different types of nonspecialists in a range of settings, each of which will have their own unique approaches and challenges.

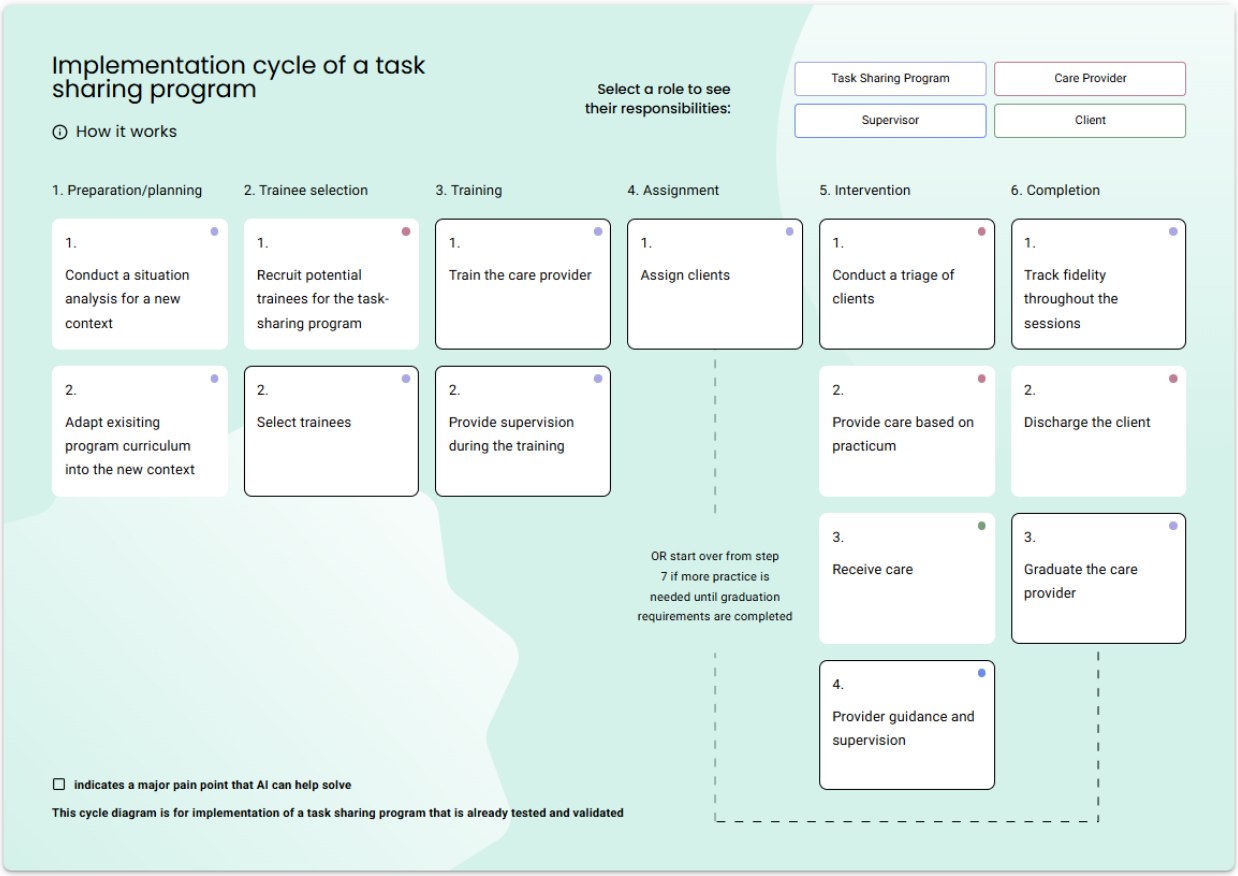
Methodology

The approach in this section was developed through a comprehensive process involving interviews with task sharing programs, workshops, and expert reviews. Interviews with representatives from organizations—such as CETA, Friendship Bench, GMH Lab at Columbia, Mental Health for All Lab at Harvard Medical School, Kids Help Phone, National Council for Mental Wellbeing, SAMHSA, StrongMinds, The Trevor Project, UNICEF, Wellcome, WHO, and 7Cups provided insights into their program cycles, challenges in scaling the programs, and potential AI solutions that could address these challenges. Workshops exchanged knowledge among practitioners and AI experts, and helped to identify best practices and areas where AI could scale program efforts.

This framework illustrates the key phases of implementing task sharing programs that have been tested and validated in one context and may be adapted for implementation in a new setting. Every program may not follow this exact sequence, and phases or steps may differ depending on contextual needs and program design. Additionally, while some steps involve clear areas where AI solutions can provide substantial support—such as with trainee selection and supervision—other steps have limited or uncertain

opportunities for improvement through AI. The cycle is intended as a flexible guide, encouraging programs to adapt it to their unique operational structures and strategic goals, particularly in integrating AI to improve various stages of the program.

The implementation cycle follows a generic approach and may include elements that are not applicable to all contexts. The visual of the example implementation cycle shared below follows a gradual progression and includes six phases which could be used by non-task sharing programs: program development, trainee selection, training, assignment, intervention, and completion. Each phase is critical to maintaining quality, ensuring provider competency, and optimizing client outcomes. At the end of the completion phase, implementation requires ongoing support and supervision, where AI might also play a role.



For the purposes of this field guide, a **phase** is a major stage in the program's implementation cycle (program adaptation or trainee selection, for example) that groups together activities with a common objective. Phases organize the implementation process into manageable segments, providing a clear road map from start to finish. Within each phase, **steps** are specific tasks that contribute to achieving the phase's objective (for example, in phase one, program adaptation, a step would be to conduct a situation analysis for a new context). These steps provide operational guidance, ensuring systematic execution of each phase (learn more).

For the purposes of this field guide, a **phase** is a major stage in the program's implementation cycle. Each phase groups together activities that share a common objective or purpose. For example, the **program adaptation** phase involves activities aimed at customizing the program for a new context, while the **training** phase focuses on preparing care providers to deliver mental health interventions. Phases help organize the

overall process into manageable and logical segments, providing a clear road map to start and finish implementing a program.

A **step** is a specific action or task within a phase that contributes to achieving the phase's objective. Steps are detailed actions that describe the practical activities required to move the program forward. For instance, within the program adaptation phase, the steps include "Conduct a situation analysis for a new context" and "Adapt existing program curriculum into a new context." Steps provide operational guidance on how to implement each phase, ensuring that all necessary actions are accounted for and executed systematically.

Instructions for interacting with the diagram

1. **Select your role**. Click on your user type (task-sharing program, supervisor, care provider, or client) to highlight the steps most relevant to your experience in a task-sharing program.

   Role definitions
   - Task-sharing program: The program responsible for training and supporting nonspecialist providers to deliver evidence-based interventions.
   - Supervisor: An experienced professional or a trained nonspecialist who provides guidance, support, and quality assurance to care providers.
   - Care provider: A nonspecialist in training (that is, trainee) or a nonspecialist who completed their training and who delivers mental health interventions to clients under supervision or without supervision, depending on the phase and type of the program.
   - Client: An individual receiving mental healthcare or support through the task sharing program.

2. **Explore the implementation cycle**. Hover over each phase to see a brief description of each stage.
3. **View challenges and AI opportunities**. Click on a step to reveal the main challenges and the ways that AI can help improve the step, if applicable.
4. **Access detailed AI use cases**. Within relevant steps, click on a specific AI use case to explore detailed insights, including users, applications, and design considerations.

Phase 1: Program adaptation

The program adaptation phase begins by evaluating how possible it is to implement the program in a new context, such as in a new country or demographic group. Next, local implementers trained in the program adapt the existing program curriculum to fit local cultural, linguistic, and regulatory needs, ensuring it is relevant in the next context and will be used. Task-sharing programs can also set up partnerships with local healthcare institutions and government bodies to align with broader mental health systems.

Step A: Conduct a situational analysis for a new context

What happens in this step?

AI could initially support local experts by providing preliminary insights about the context the program is being applied to. It could analyze existing literature or data on mental health needs and support needed, identify persisting needs, or flag potential mismatches in content or language. However, this phase requires

robust and specific knowledge of the context and situation, so AI's potential for impact may be limited in this phase, and it would require additional human oversight.

Implementers review existing data on local mental health needs, human resources, infrastructure, health system capacity, and relevant cultural and linguistic factors to determine whether the task-sharing program is feasible and how it should be tailored for the specific context. Typically, this step involves reviewing published literature and information produced outside of traditional publishing and distribution channels, examining local epidemiologic data (for example, the prevalence of mental health conditions), meeting and conducting interviews with local stakeholders (such as government officials, nongovernmental organizations (NGOs), or community leaders), and mapping out available mental health services and referral pathways. By gathering both quantitative and qualitative information, implementers can identify meaningful gaps and opportunities that will shape the design of the task-sharing program.

Step B: Adapt existing program curriculum into a new context

Programs work with local experts and partners to prepare evidence-based intervention materials—such as manuals, handouts, and training guides—to match local cultures, languages, and health system contexts. This step typically involves translating materials into relevant or dominant languages, incorporating culturally relevant examples into modules, and collaborating with local mental health professionals to ensure the program will be accepted. A structured adaptation process can include reviews by local experts, pilot testing, and final validation to balance core evidence-based content with cultural relevance.

Phase 2: Trainee selection

Recruiting the right nonspecialist providers is important for a program's success. Potential trainees may volunteer or be identified through local partners based on experience and interest. They are then selected with established criteria to ensure they have the skills, motivation, and time to participate in the training required and the delivery of the program.

Step A: Recruit trainees for the task sharing program

Potential trainees can learn about the program via targeted e-mail distribution. They also could learn about the program through word of mouth, social media channels, customized text messages, or paper flyers or postcards with a QR code distributed or posted through community centers, schools, workplaces or businesses. Recruitment is also often done within the organization taking on the program—for example, a healthcare system moving some tasks to existing personnel or upskilling existing staff. Recruitment often involves collecting basic information in line with privacy best practices and relevant regulations about each applicant's demographics, education, language skills, and interest and experience in mental health.

While AI can aid in the recruitment process of trainees for a task-sharing program (such as summarizing information about prospective trainees), there may be limited opportunities to substantially enhance it because creating an applicant pool and attracting talent relies on human outreach and judgment.

Step B: Select trainees

What happens in this step?

In this step, many different individuals evaluate applicants based on their past experiences, connection to the target population (including physical proximity), and qualities such as available capacity to engage, ability to connect meaningfully with people through compassion and empathy, and more-technical skills

(such as problem-solving ability). Coordinators ensure that trainees can be effectively trained to deliver evidence-based mental health interventions.

Challenges

Three challenges are most prevalent during this step:

- **Limited applicant pool**. If incentives (such as stipends or career development pathways) are unclear to prospective trainees, it can be challenging to attract individuals who can reliably commit to training and ongoing service delivery.
- **Time-intensive screening process.** Conducting thorough interviews, role-plays, or background checks can strain program resources, especially if local staff such as program coordinators and trainers are limited.
- **High dropout risk.** Participants may leave mid-training due to personal, financial, or family obligations, resulting in understaffing. Volunteers may also dropout if the reality of task sharing differs from their initial expectations.

Opportunities for AI

- **AI-powered screening tool.** While not applicable in all contexts (for example, where candidates do not have resumes or cannot attend virtual interviews), AI tools can be used to review candidates' resumes assessing skills suited for delivering mental health interventions after, having collected basic information. AI tools can also assess candidates' responses to AI-based simulations and rate soft skills (such as communication clarity) in line with local language, customs and contexts, or tech skills. If a candidate is selected for the training program, these simulations could be the baseline skill assessment to recommend additional training to close the gaps in competencies.

Relevant use case(s)

- **Applicant screening tool**

Phase 3: Training

In this phase, selected care providers undergo structured training sessions. These sessions are designed to provide volunteers with the knowledge and skills required for them to deliver evidence-based psychological support. Sessions can include instructor-led training and experiential learning sessions (such as roleplaying) to teach trainees how to apply best practices to real-life scenarios.

There are substantial opportunities for AI to enhance the training phase of task sharing programs by supporting both the training of care providers and the supervision process.

Step A: Train the care provider

In this step, care providers attend training sessions that can blend instructor-led content, role-plays, group work, and interactive exercises. The training usually starts and ends with assessments that test whether participants can apply important skills effectively and measure the effectiveness of the training to build skill sets.

Challenges

Three challenges are most prevalent during this step:

- **Resource-intensive training process.** Interactive sessions, role-plays, and practical exercises may require substantial human and financial resources.
- **Limited number of experienced trainers.** Having a small pool of qualified supervisors or master trainers may create bottlenecks, especially in remote or under-resourced areas.
- **Hard to ensure adherence to protocol.** Without a robust oversight system and systematic recordkeeping, it could be difficult to ensure providers are covering essential modules in the program.

Opportunities for AI

We acknowledge there may be limitations in data from some countries or contexts needed to ensure the use of AI is culturally appropriate.

- **Adaptive e-learning**. AI-driven instructor-led training modules can adjust the pace and difficulty of training sessions based on individuals' needs. They can also recommend additional training modules based on the gaps they spotted during the baseline assessment and throughout the training.
- **Role-play simulations**. AI-powered simulations allow trainees to practice unlimited client scenarios, including more complex cases in a risk-free environment.

Relevant use case(s)

- **Adaptive training interface**
- **AI training room**

Step B: Provide supervision during training

In this step, supervisors oversee practice sessions and offer real-time or near-real-time feedback, to ensure trainees are properly implementing the intervention. Regular check-ins throughout the training program help reinforce therapeutic protocols and address challenges volunteers may encounter during their work, such as complex cases or emotional strain.

Challenges

Three challenges are most prevalent during this step:

- **Supervisor shortage**. Many programs may not have enough qualified supervisors to meet high demand, leading to minimal or delayed feedback, which can negatively impact the quality of care offered and clinical outcomes.
- **Inconsistent feedback.** Even when supervisors are available, their focus can vary widely—some emphasizing technique-specific skills over foundational helping skills, for example— leading to inconsistencies in trainees' knowledge of and ability to implement interventions.
- **Trainee burnout**. Handling emotionally heavy sessions can lead to high stress, and trainees may not get adequate emotional support during training.

Opportunities for AI

- **Augmented feedback**. AI can evaluate session audio or transcripts and offer preliminary skills assessments for supervisors to validate. We acknowledge this may require additional consideration when it comes to how it compares to the specifics of other countries or contexts.

- **Supervisor dashboards**. Aggregated data can highlight the trainees who need additional mentoring or emotional support.

Relevant use case(s)

- **Post-session feedback report**
- **Supervisor dashboard**

Phase 4: Assignment

Once care providers complete initial training, in some contexts, they are assigned clients by trainers or supervisors, based on their readiness, the case complexity, and client needs. Allocating assignments in this way balances caseloads and ensures that complex cases are handled by more experienced providers.

In contexts where providers are assigned individual clients, there are substantial opportunities for AI to enhance the assignment phase of task-sharing programs, by assisting supervisors in matching trainees with the right cases. However, we acknowledge that in many contexts, nonspecialist care providers work with the population within their area, and there is rarely room for assignment within these programs. Also, given the importance of building trusted relationships through these existing mechanisms, AI's potential for impact is limited here.

Step A: Assign clients

In this step, programs or supervisors typically match care providers to clients based on availability and, in some cases, other factors such as skill set, language fluency, gender preferences, or the severity of the client's condition. This process ensures that care providers are handling appropriately leveled cases while maximizing coverage.

Challenges

- **Inefficient or ad hoc matching.** If cases are assigned on an ad hoc basis, in some instances, matching may lead to suboptimal patient assignments (for example, a new care provider may be assigned to a complex case, while an experienced provider goes underutilized). Additionally, one-to-one manual matching may also be labor-intensive for trainers or supervisors.
- **Scheduling bottlenecks.** Geographic distance, lack of technology (such as poor phone or internet coverage), and limited staff capacity may lead to missed appointments and hamper overall client engagement. Some teams also may note difficulty in coordinating volunteers' availability, especially if they have competing responsibilities or limited connectivity options.

Opportunities for AI

- **Smarter provider–client matching.** AI algorithms can consider provider experience, language, and client complexity and needs to match the client with the right care provider and reduce wait times — ultimately improving adherence and coverage.
- **Automated scheduling tools.** AI-driven platforms can send reminders, reassign appointments if care providers become unavailable, and help reduce administrative burdens.

Relevant use case(s)

- **Provider-client matching**

Phase 5: Intervention

Care providers begin client interactions by conducting triage, assessing mental health needs, and then delivering interventions under ongoing supervision. Ideally, if clients require higher-level care, the program facilitates a referral to the appropriate care pathway. Throughout this phase, organizations continue providing real-time guidance and supervision to support providers to ensure care delivered is within acceptable quality parameters.

There are substantial opportunities to test and validate how AI can enhance the intervention phase of task-sharing programs by supporting triage, assisting care providers in delivering evidence-based interventions, and providing real-time guidance.

Step A: Conduct triage of clients

In this step, care providers assess clients' mental health statuses, risk factors, and immediate needs using brief screening tools. If a client's needs exceed task sharing's scope, ideally they are referred for specialized care. This might also happen before getting assigned to a provider.

Challenges

- **Resource limitations.** Even if a client is identified as high-risk, local referral networks might be nonexistent or overburdened, making it difficult to provide timely, advanced care.
- **Stigma.** Clients may minimize symptoms to avoid labeling, complicating accurate triage.
- **High caseload.** Limited staff and an excess of new referrals may result in rushed or superficial triage, which could create a risk of care providers inaccurately assessing the severity of a case.

Opportunities for AI

- **Risk assessment**. Algorithms could augment detection of red-flag symptoms or behaviors of clients during screening responses to indicate which clients are higher risk and provide these insights to the care provider. This might not be applicable in certain contexts where risk assessments may not be conducted digitally.
- **Summary of relevant information.** AI might facilitate the process to summarize intake forms and highlight information that is most relevant for a given context.

Relevant use case(s)

- **Case severity triage facilitator**

Step B: Provide care based on the program's curriculum

In this step, trainees deliver evidence-based interventions under supervision. They follow a manualized protocol, focusing on practicing the foundational helping skills and therapy techniques they learned during training.

Challenges

- **Protocol drift**. Over time, providers may omit or replace components of the intervention, especially in high-volume or chaotic settings (such as group sessions with multiple distractions).
- **Complex cases**. Some clients may face overlapping challenges, such as interpersonal violence, homelessness, or chronic illness, which can overwhelm inexperienced care providers who lack advanced support.

- **Limited session records**. Paper-based or minimal documentation and variability in documentation quality may hinder continuity and real-time feedback. Some providers skip detailed notes because of time constraints or technology barriers.

Opportunities for AI

- **Session recordings for data intake.** AI tools can automatically transcribe session audio, highlight intervention steps, and feed structured data into program databases, enabling more consistent record keeping and timely feedback for care providers. The AI tools could also use fidelity rating sheets that already exist in the program to rate providers based on the same criteria that clinical supervisors are using.

Relevant use case(s)

- **Documentation assistant**

Step C: Receive care (clients)

In this step, clients attend sessions to learn coping strategies, receive psychoeducation, and collaborate with care providers. Some interventions require "homework" or ongoing practice between sessions.

Step D: Provide guidance and supervision

In this step, supervisors review care providers' care session notes or recordings and then share their observations, recognizing successes and pinpointing improvement areas. During these review sessions, providers can also voice any struggles, ensuring a cycle of ongoing skill development and building a feedback loop for the program.

In cases where nonspecialist care providers face urgent clinical, ethical, or emotional challenges, supervisors may listen in during sessions with clients and provide immediate guidance afterward. By offering timely support, supervisors can help maintain care providers' well-being and reinforce quality standards of interventions.

Challenges

- **Supervisor overload**. A single supervisor may support numerous care providers across multiple sites, limiting their capacity to offer timely and personalized feedback.
- **Limited in-person access**. Remote or rural programs face geographic constraints, limiting supervisors' ability to observe sessions directly and reducing the possibility of them providing timely feedback to volunteers.
- **Variable oversight**. Some supervisors may focus on administrative responsibilities such as scheduling, documentation, or program reporting over clinical fidelity and skill development.

Opportunities for AI

Acknowledging the limited data and research in some countries or contexts, these opportunities to use AI may have limited applicability in some circumstances.

- **Real-time guidance**. AI tools could provide on-the-spot recommendations for care providers and show them the next steps they can take, including potential responses they could provide to clients based on intervention protocols.
- **Structured feedback**. AI could highlight recurring competency gaps so that supervisors can focus on the critical development areas of the care provider.

Relevant use case(s)

- **Real-time guidance AI companion**
- **Post-session feedback report**
- **Supervisor dashboard**

Phase 6: Completion

Throughout sessions, supervisors will track how well care providers are implementing interventions as intended (clinical fidelity) to maintain quality standards. Once clients have completed their care, they are discharged, and care providers in training who have met program requirements graduate, marking their readiness to practice the learned interventions in their communities with appropriate supervision.

The completion phase focuses on ensuring care providers are ready to complete the task sharing training program, providing support in graduation, and facilitating the transition into community-based practice.

There are substantial opportunities for AI to enhance the completion phase of task sharing programs by tracking fidelity across sessions and supporting the post-graduation process for care providers.

Step A: Track fidelity throughout sessions

Supervisors will continue supporting care providers through regular supervision at the appropriate cadence to prevent provider drift and ensure high-quality care is being delivered throughout the sessions. Supervisors focus on assessing care providers based on their technique-based competencies (problem-solving steps, for example) and foundational helping competencies (such as empathy) to ensure providers are upholding clinical fidelity.

Challenges

- **Observer bias**. Supervisors may be inconsistent or more lenient with some care providers due to pre-existing biases.
- **Resource constraints**. Frequent fidelity checks may be infeasible, especially in large-scale or remote programs.
- **Sparse documentation**. Without standard forms or thorough logs, supervisors may lack the context or crucial information on how care providers did in the sessions to monitor and evaluate fidelity.

Opportunities for AI

Acknowledging the limited data and research in some countries or contexts, these opportunities to use AI may have limited applicability in some circumstances.

- **Automated fidelity checks**. AI can scan session transcripts for meaningful phrases or steps to confirm adherence to intervention protocols, with the results of the scan available to be accessed after every session.
- **Trend alerts**. Dashboards might highlight a drop in a care provider's fidelity score, prompting the supervisor to intervene.

Relevant use case(s)

- **Post-session feedback report**
- **Supervisor dashboard**

- **Documentation assistant**

Step B: Discharge the client

Once goals are met and clients exhibit clinically substantial improvement, clients are formally discharged from the program. Providers typically discuss warning signs of relapse, self-help strategies, and community resources with clients before they are discharged.

There are limited opportunities for AI to enhance the discharge process in a task sharing program because the decision to discharge a client is typically grounded in nuanced clinical judgment and interpersonal dynamics that are difficult and potentially risky to automate.

Step C: Graduate the care provider

Providers demonstrating consistent fidelity and competence receive formal recognition and continue providing treatment with the appropriate supervision.

Challenges

- **Drift**. Without refresher sessions, newly graduated providers may lose meaningful competencies over time and drift from the program's protocols.
- **Post-program adherence.** Provider drift may increase if providers are not supported by ongoing supervision.

Opportunities for AI

Acknowledging the limited data and research in some countries or contexts, these opportunities to use AI may have limited applicability in some circumstances.

- **Refresher microlearning**. On-demand AI modules could help graduates sustain skill levels post-graduation
- **Role-play simulations**. AI-powered simulations and real-time guidance allow graduates to continue practicing what they learned from the program.

Relevant use case(s)

- **AI training room**
- **Adaptive training interface**

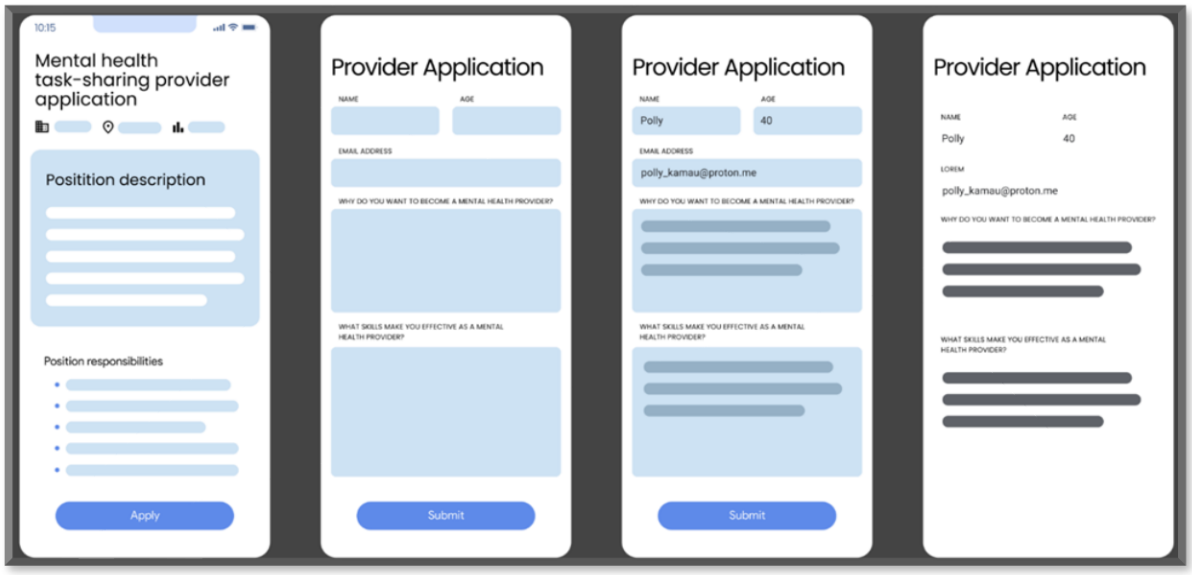
Section 2: Real-world examples using AI

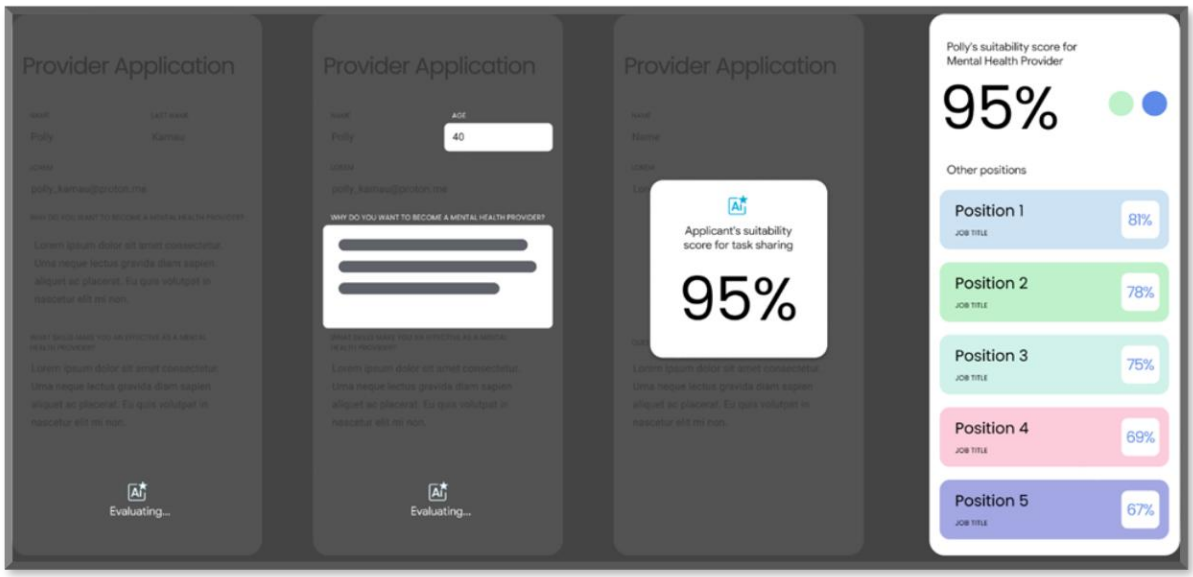
This section explores potential AI use cases that could help address challenges present throughout the implementation cycle and expand the reach of mental health task sharing programs. Each use case includes detailed scenarios that show how AI can be applied in practice, identifies the opportunities unlocked with these tools, describes direct end users, and provides real-world examples of AI in mental health task-sharing in certain contexts.

Through practical examples and real-world applications, this chapter demonstrates how AI has the potential to enhance the capacity and impact of mental health task sharing programs, paving the way for scalable and sustainable mental health services. Throughout, though not explicitly listed, clients are always among the end users benefiting from each of these solutions, given that the overall goal of improving efficiency, training, and other aspects is to have these lead to better client care as a result.

We acknowledge that some of the use cases and examples presented may require additional consideration when it comes to how they compare to the specifics of other countries or contexts.

Use case 1: Applicant screening tool





End user(s):

- Task-sharing program

Description:

Applicants who applied to become a care provider undergo an AI-enabled screening process to determine their willingness to participate in the task-sharing program and evaluate if they have the skills needed to be a mental health task-sharing provider.

Example scenario:

An applicant interested in becoming a care provider completes an online prescreening assessment through the **applicant screening tool**. The AI tool asks a series of adaptive questions to evaluate the applicant's alignment with the program's selection criteria (such as relevant work experience or physical proximity to the target population). The applicant screening tool assesses the applicant's responses for critical thinking, communication skills, and motivation. Once the applicant finishes the assessment, the AI tool generates a preliminary suitability score and identifies areas for follow-up during the interview process. The application automatically notifies the program coordinator about qualified candidates, streamlining the selection process.

Opportunities unlocked:

The **applicant screening tool** could address several challenges in mental health task-sharing programs, including the following:

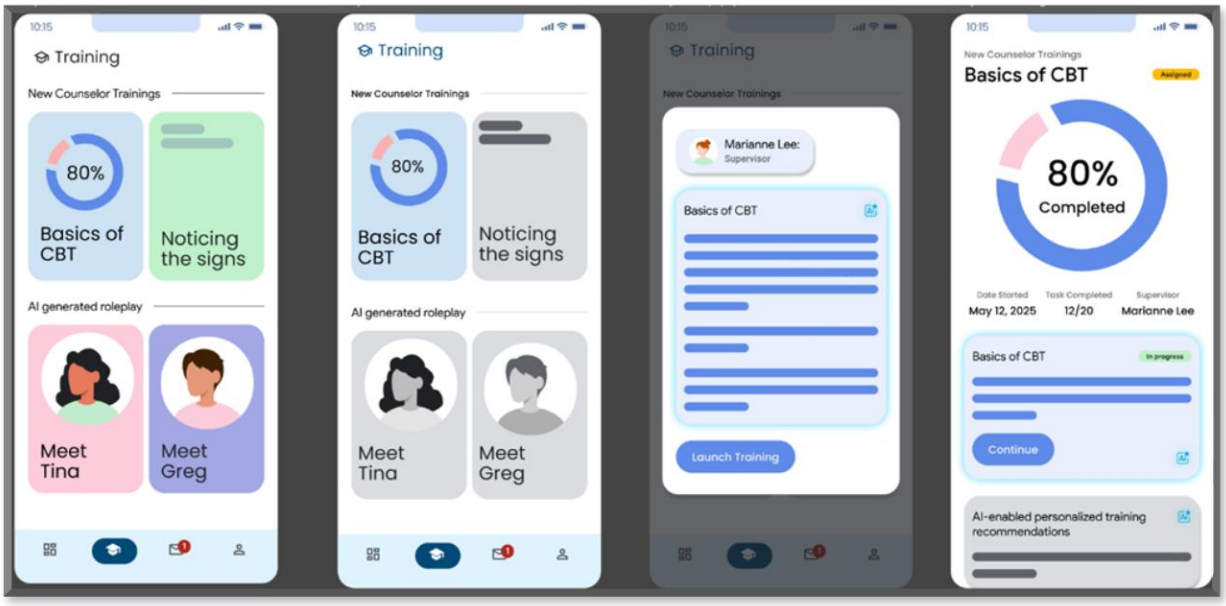
- **Time-intensive, variable screening process.** The AI tool could automate initial assessments using a set of criteria and shortlist qualified candidates, which could help standardize the care provider evaluations and substantially reduce the workload on program coordinators and trainers who would otherwise need to conduct time-consuming interviews, role-plays, or background checks.
- **High dropout rates.** The AI tool could evaluate applicants' skills and engagement levels, identifying potential risks for dropout early in the process. This feature could help minimize mid-training attrition and ensure consistent staffing levels.

How could the end user(s) benefit from this solution?

The primary end users who could directly benefit from this solution are the following:

- **Task sharing organization.** Responsible for managing the recruitment and selection process, the task sharing organization could benefit from automated shortlisting and suitability scoring. They could also benefit from a generated provider competence matrix that can help adapt the training to better suit the specific cohort of trainees, thereby reducing their administrative burden and allowing them to focus on final interviews and onboarding. Additionally, reducing dropout rates could save an organization time and resources, and standardizing processes helps to ensure that all care providers have a similar quality.
- **Client.** Reducing dropout risks and ensuring that selected volunteers are thoroughly screened, high-quality candidates would help provide clients with exceptional and consistent care.
- **Care provider in training.** Providing a clear description of the job experience and requirements helps applicants decide if they can and want to commit to the program, saving time for those who aren't interested in going through the screening and training process.

Use case 2: Adaptive training interface



End user(s):

- Care provider

Description:

Trainees (care providers in training) who complete the core training modules can use the training curriculum interface with an adaptive e-learning algorithm to receive additional, personalized training that reinforces specific skills. Care providers can continue using this interface after graduating from the task-sharing programs to refresh their training and avoid losing knowledge and skills.

Example scenario:

After completing core training modules, a trainee logs into the **training curriculum interface** and sees a personalized dashboard featuring evidence-based training modules tailored to the strengths and weaknesses identified during the roleplaying exercise they participated in during the interview (such as cognitive restructuring, empathetic listening, problem solving, or flexible thinking). The AI tool tailors the training curriculum by adjusting the pace and difficulty based on the provider's baseline assessment and ongoing performance.

For example, a trainee struggles with accurately identifying thought patterns in cognitive behavioral therapy (CBT). The system detects this challenge based on practice rounds with the AI-generated scenarios, based on the feedback supervisors provide manually, or based on the feedback that the provider companion (see use case 9) gather during the practice sessions. The adaptive e-learning algorithm then suggests additional learning materials, practice scenarios, and simpler modules before advancing to more-complex techniques.

Opportunities unlocked:

The **training curriculum interface** could address several challenges in mental health task-sharing programs, including the following:

- **Resource-intensive training processes.** Traditional interactive sessions, role-plays, and practical exercises require substantial human and financial resources. The adaptive e-learning could reduce

dependency on human facilitators by offering personalized, self-paced training, allowing trainees to practice repeatedly without logistical constraints.

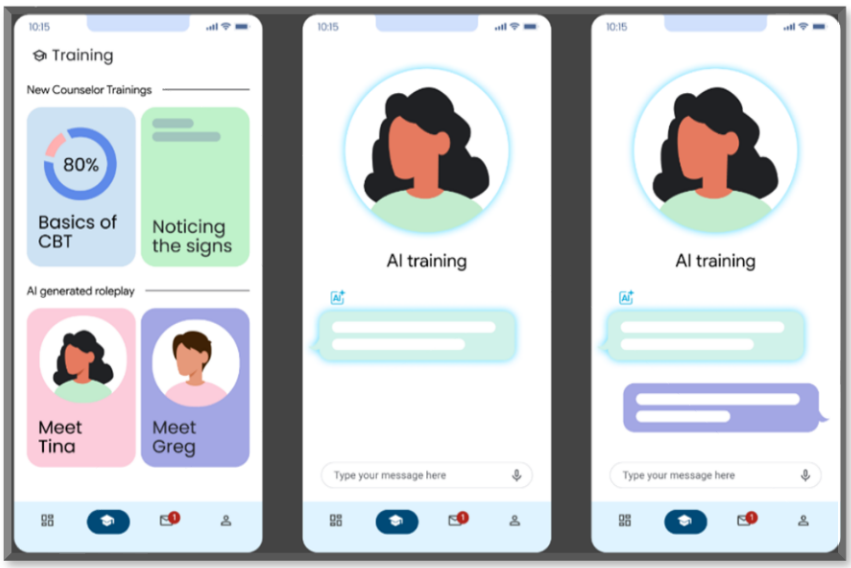
- **Ensuring adherence to protocol.** Maintaining consistent training quality is challenging without robust monitoring. The **training curriculum interface** could track each trainee's progress and ensure all essential modules are completed at a pace that is comfortable to the trainee. By standardizing training, feedback, and performance tracking, the AI tool could ensure that best practices are followed, regardless of the trainee's location. This approach allows all trainees to complete the program equipped with the same skills through standardization.
- **Drift and post-program adherence.** Some care providers may struggle to maintain protocols and may experience a decline in knowledge or skills without ongoing supervision or support after the program. This **training curriculum interface** with an adaptive e-learning algorithm could serve as a continuous resource with microlearning modules to reinforce best practices and keep skills sharp even after program completion.

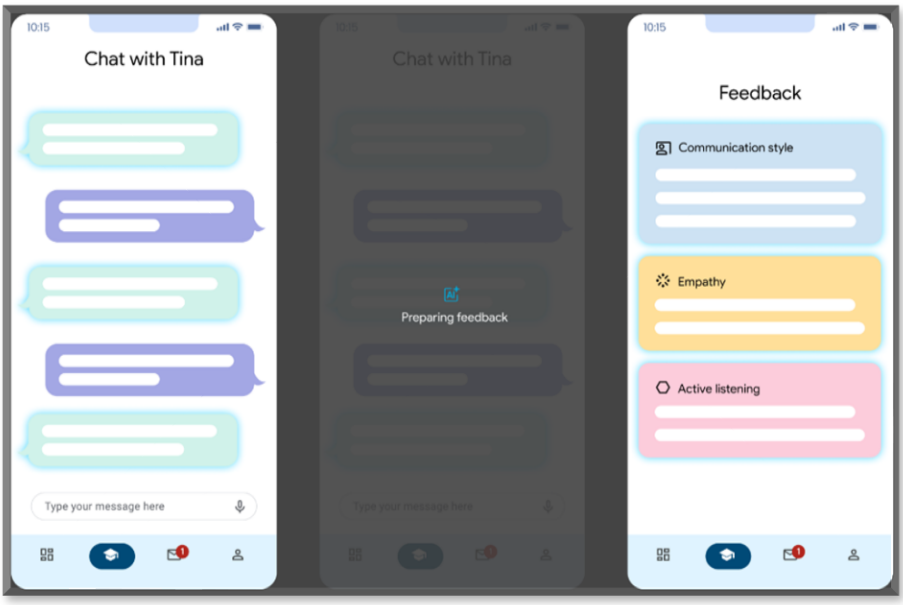
How could the end user(s) benefit from this solution?

The primary end user who could directly benefit from this solution is:

- **Care provider.** Trainees or care providers could engage with the platform for personalized, adaptive learning experiences that help them address gaps in their skills and knowledge and get to the same level of competency as their peers.

Use case 3: AI training room





End user(s):

- Care provider

Description:

A trainee or care provider can engage in role-play with AI-generated clients that are designed to simulate diverse scenarios and test the care provider's ability to navigate different challenges. Trainees and care providers receive real-time feedback and tailored suggestions as the conversation unfolds, delivered through video, voice, or text.

Example scenario:

A trainee interacts with an AI-generated client who struggles with social anxiety and avoids social situations because they are afraid of being judged. The **AI training room** simulates a realistic conversation, allowing the trainee to practice identifying negative thoughts, developing coping strategies, and setting realistic goals. The AI system evaluates the trainee's communication style, empathy, and adherence to technical skills (such as CBT techniques), then provides real-time feedback to enhance learning and application of skills.

Opportunities unlocked:

The **AI training room** could address several challenges in mental health task sharing programs, including the following:

- **Lack of realistic scenarios.** Traditional role-plays are done with peers who may not accurately reflect real-life patients or situations, limiting opportunities for realistic practice. AI roleplays can present trainees with a wider variety of realistic scenarios.
- **Resource-intensive training process.** Traditional role-plays require human facilitators, physical spaces, and logistical coordination, making them costly and challenging to scale. By simulating realistic interactions virtually, the **AI training room** could decrease the reliance on human role-players, allowing trainees to practice a variety of realistic scenarios at their own pace.

- **Limited number of experienced trainers.** In many regions, a shortage of qualified trainers limits the number of mental health care providers who can be trained. The **AI training room** could bridge this gap by providing real-time feedback on therapeutic techniques, communication skills, and other skills required by the program. Trainees could learn independently while maintaining high training standards, ensuring consistent education even in remote or underserved areas.
- **Ensuring adherence to protocol.** Maintaining consistent training quality is challenging without robust monitoring. The **AI training room** could track each trainee's progress, ensuring that all essential modules are completed, and that the skills required by the program are consistently applied.
- **Drift and post-program adherence.** The **AI training room** could serve as a continuous resource with role-play simulations to reinforce best practices and ensure skills remain sharp long after the completion of a task-sharing program.

How could the end user(s) benefit from this solution?

The primary end user who could directly benefit from this solution is:

- **Care provider.** Trainees or care providers could directly engage with the **AI training room** and benefit from realistic practice scenarios that build confidence and competence in their skills.

Real-world example 1:

The Trevor Project is a suicide prevention and crisis intervention nonprofit organization that provides information and support to LGBTQ+ youth. Through text, chat, and phone calls, the organization connects young people in crisis with trained volunteer counselors who offer immediate and confidential support.

To expand the capacity of its services and meet increasing demand, The Trevor Project developed an AI-powered crisis counselor training bot. This AI tool simulates realistic conversations, enabling trainees to practice crisis scenarios in a safe, supportive, and scalable environment before live interactions with youth.

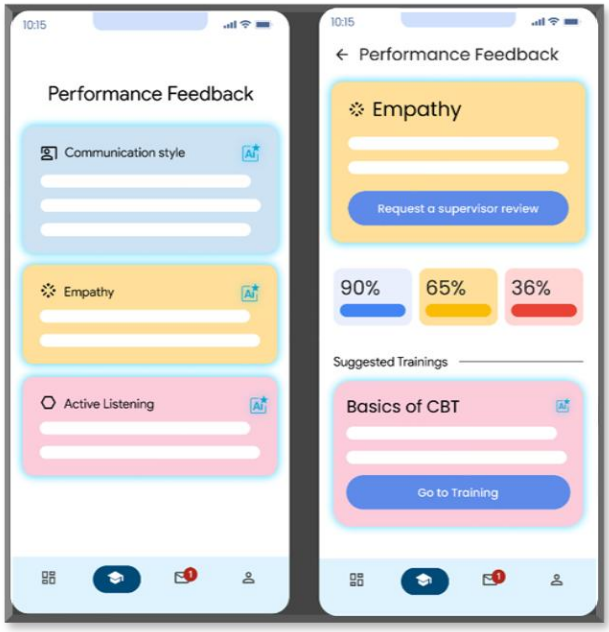
By using the crisis counselor training bot, The Trevor Project trained more than 1,000 crisis counselors in about one year, enabling the organization to scale their counselor base more quickly and serve a higher volume of LGTBQ+ youth in crisis.

Real-world example 2:

Partnership to End Addiction is a national nonprofit focused on preventing and addressing substance use disorders by providing families, communities, and professionals with practical knowledge they can trust to support young people at risk or struggling with addiction. Families and caregivers can get evidence-based resources and support services to take confident action in preventing and treating substance use.

To enhance its peer coach programming, Partnership to End Addiction has embedded AI simulations into their peer coach training curriculum. Simulations offer a cost-effective way to train coaches in having compassionate interactions with the parents they will be working with. AI is also allowing the Partnership to assess the quality of active coach conversations and provide feedback to the coaches to support ongoing skill development. AI is helping Partnership to End Addiction serve more families and improve the efficacy of this support.

Use case 4: Post-session feedback report



End user(s):

- Care provider

Description:

After a simulated training session in the AI training room or a client session, a trainee or care provider receives a report with meaningful insights, suggestions for additional modules, and the option to request an expert review from their task-sharing supervisor.

Example scenario:
After completing a simulated training session, a trainee or a care provider receives an AI-generated **post-session feedback report** detailing their performance. The report includes scores on meaningful competencies, such as empathy; adherence to therapeutic models, and listening skills. Additionally, it provides qualitative insights that will point out exactly where adherence veered from the expectation and exactly which part of delivery showed a deviation and did not meet the standard. For example, “Your empathy was high today, making your client feel at ease. However, adherence to the model was below expectations. Consider trying a different approach next time.” The report also recommends specific training modules that trainees should review again to strengthen underdeveloped skills. The trainee or care provider can then choose to request to have a supervisor review their feedback and provide further guidance. This structured feedback empowers trainees to continuously improve their skills through actionable insights and targeted recommendations.

Opportunities unlocked:

The **post-session feedback report** could address several challenges in mental health task sharing programs, including the following:

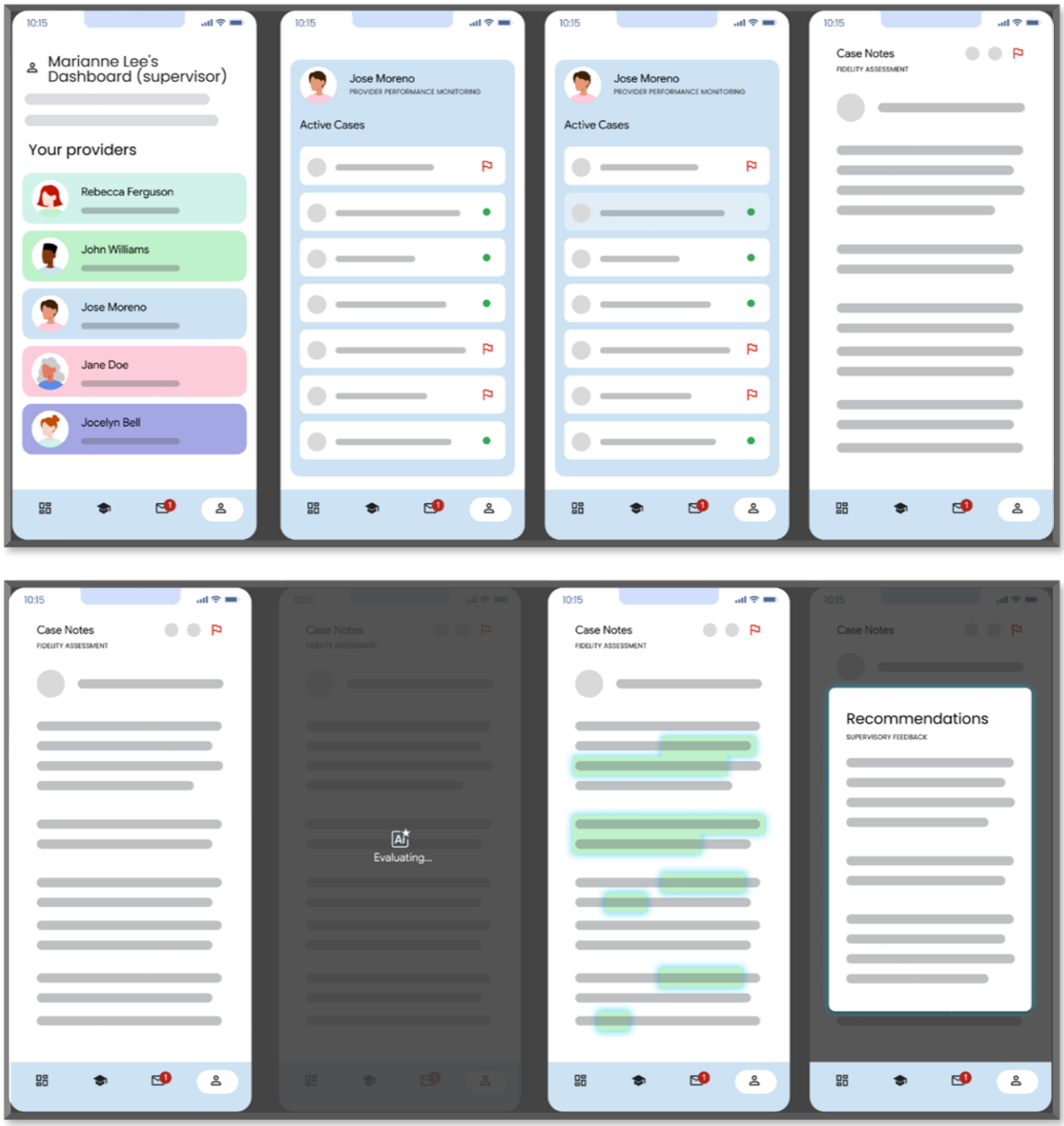
- **Supervisor shortage, limited time, and resource constraints.** Many training programs face a shortage of qualified supervisors with limited time and resources, which may result in delayed or infrequent feedback for trainees. The **post-session feedback report** could fill this gap by automatically generating performance evaluations immediately after each session, enabling timely feedback and maintaining training momentum without having to wait for human supervisors. Additionally, by streamlining the feedback process, the **post-session feedback report** allows supervisors to dedicate more time to trainees who require deeper guidance and support.
- **Protocol drift.** The **post-session feedback report** could also help maintain fidelity during the task-sharing program or after the trainee has graduated from the program. It could highlight meaningful phrases or steps to confirm that trainees or care providers are adhering to intervention protocols and indicate if there is a drop in a provider's fidelity score. Supervisors can then intervene, if necessary.
- **Inconsistent feedback and observer bias.** When supervisors are available, their feedback can vary substantially depending on their focus areas. For example, some supervisors may emphasize technique-specific competencies, while others may focus on foundational helping skills. The **post-session feedback report** could standardize evaluations by consistently assessing meaningful competencies across both of these focuses, such as empathy, technique adherence, and listening skills. It could offer balanced feedback across all critical areas, ensuring comprehensive skill development and reducing subjective biases.

How could the end user(s) benefit from this use case?

The primary end user who could directly benefit from this solution is:

- **Care provider.** Trainees or care providers could benefit from the **post-session feedback report** by using the actionable insights they receive on their strengths and areas of development to refine their foundational helping and technical skills.

Use case 5: Supervisor dashboard



End user(s):

- Supervisor

Description:

Through the supervisor dashboard, supervisors can easily access the client cases and feedback reports of the care providers assigned to them to monitor their work, evaluate their performance, and provide feedback as needed—for example, if they notice a score for a technical skill required by the program is declining. Pertinent notifications will be flagged for supervisors by an automatic alert.

Example scenario:

A supervisor logs into the **supervisor dashboard** to review the performance of several care providers who are under their supervision. The dashboard provides an overview of each provider's active cases, risk assessments, and session progress, and it highlights cases that need immediate attention.

For example, the dashboard alerts the supervisor if a provider's sessions need to be reviewed because it is showing inconsistencies in therapeutic techniques delivered. By accessing the detailed session analysis, the supervisor sees that the provider was not closely adhering to the techniques required by the treatment and received feedback on their communication style. The system recommends targeted training in those specific techniques to improve their skills.

The dashboard also offers performance metrics across important skills, such as empathy, technique adherence, and directive communication, to help the supervisor identify patterns across supervisees. By analyzing these metrics, the supervisor can tailor feedback, prioritize one-on-one coaching sessions, and allocate additional training resources where needed.

Opportunities unlocked:

The **supervisor dashboard** could address several challenges in mental health task sharing programs, including the following:

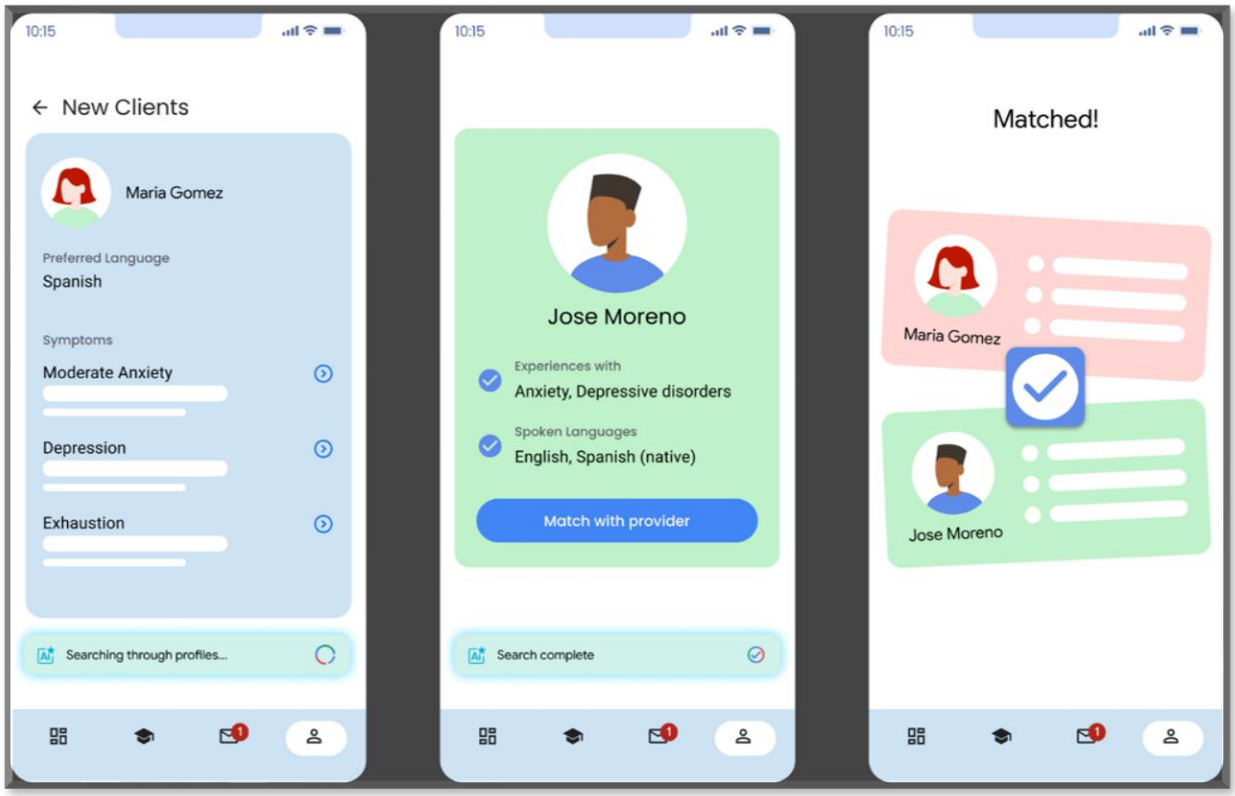
- **Supervisor shortage.** Many programs face a shortage of qualified supervisors, leading to minimal or delayed feedback. The **supervisor dashboard** could alleviate this shortage by automating preliminary assessments, highlighting high-priority cases, and streamlining feedback. Supervisors could then more efficiently manage larger caseloads without compromising quality.
- **Inconsistent feedback.** Without standardized evaluation criteria, feedback quality can vary substantially across supervisors. The dashboard could provide consistent, data-driven performance metrics and ensure that feedback is balanced across both technique-specific competencies and foundational helping skills, such as empathy and active listening.

How could the end user(s) benefit from this solution?

The primary end users who could directly benefit from this solution is:

- **Supervisor.** Experienced mental health professionals or trained nonspecialists who oversee care providers could use the dashboard to monitor the performance metrics of care providers during training, provide data-driven feedback, and deliver targeted coaching.

Use case 6: Provider–client matching



End user(s):

- Care provider

Description:

An AI-powered tool analyzes provider skills (such as foundational helping skills, technical skills, language fluency, and caseload capacity) and client needs (such as language preference, case complexity, and cultural background) and matches care providers to clients using a standardized matching algorithm.

Example scenario:

A newly trained care provider fluent in Spanish, is ready for their first client. Meanwhile, a new client who is experiencing moderate anxiety and who prefers to speak Spanish, awaits their initial counseling session.

The **provider–client matching** tool analyzes the care provider's skills, language fluency, and caseload capacity while also assessing the client's anxiety level, language preference, and cultural background. The AI tool matches care providers with clients based on similarities and compatibilities in each person's background.

The supervisor receives a notification that one of their supervisees has matched with a new client. The supervisor reviews the care provider's schedule to assess capacity and approves the match with one click. The care provider begins sessions with the client that they matched with. During the session, the client feels understood, building a strong therapeutic alliance from the start.

Opportunities unlocked:

The **provider–client matching** tool could address the following challenge in mental health task-sharing programs:

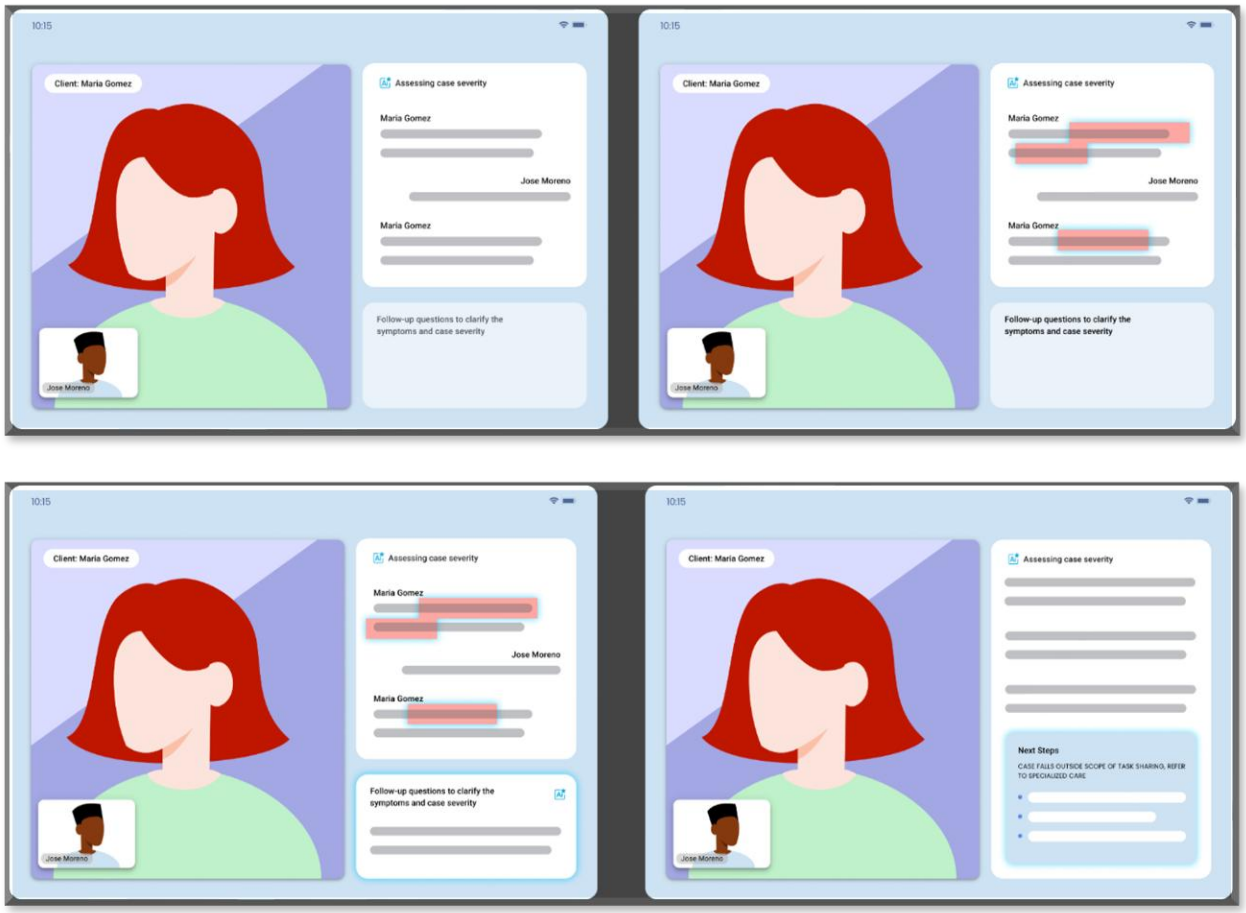
- **Inefficient or ad hoc matching.** Without structured criteria or data-driven tools, manual matching is labor-intensive and prone to human bias. The AI tool could apply standardized matching algorithms, to ensure that clients are matched with the most suitable care provider based on skill set, language fluency, case complexity, and client needs. This feature reduces mismatches in personality or capacity and optimizes the use of the provider's time and skills.

How could the end user(s) benefit from this solution?

The primary end users who could directly benefit from this solution are:

- **Care provider.** Care providers could benefit from being matched with clients that align with their skills, experience, and language fluency. More appropriately matches could lead to a higher confidence in and commitment to the program.
- **Client.** Clients could get better care from and feel more understood by care providers who align with their preferences, cultural background, and primary concerns (such as anxiety, depression).

Use case 7: Case severity triage facilitator



End user(s):

- Care provider

Description:

An AI-powered triage facilitator supports care providers during client intake by determining whether a case falls within the scope of task sharing. It can analyze conversations in real time, ask the care providers questions about the client's mental health, and then determine if the case is appropriate for task sharing or if it should be escalated to more specialized care. The tool helps standardize triage by identifying cases that require specialized care or providing tailored intervention guidance for cases that can be managed within task-sharing programs.

Example scenario:

During an intake session, the care provider meets with a new client who reports feeling stressed and tired all the time. The **triage facilitator** analyzes the conversation in real time, detecting answers that suggest severe depression. To guide the care provider, the tool suggests follow-up questions to clarify symptoms and assess the severity of the case.

If the client's needs exceed the scope of task sharing, the **triage facilitator** classifies the case as high-risk and recommends referring the client to specialized care. It generates a structured summary with flagged keywords and a risk severity score based on clinical tools, such as a patient health questionnaire (PHQ-9) that assesses the severity of a client's depression. If the case falls within the task-sharing program's scope, the **triage facilitator** provides guidance on appropriate interventions and next steps based on the program's structure.

Opportunities unlocked:

The **triage facilitator** could address several challenges in mental health task sharing programs, including the following:

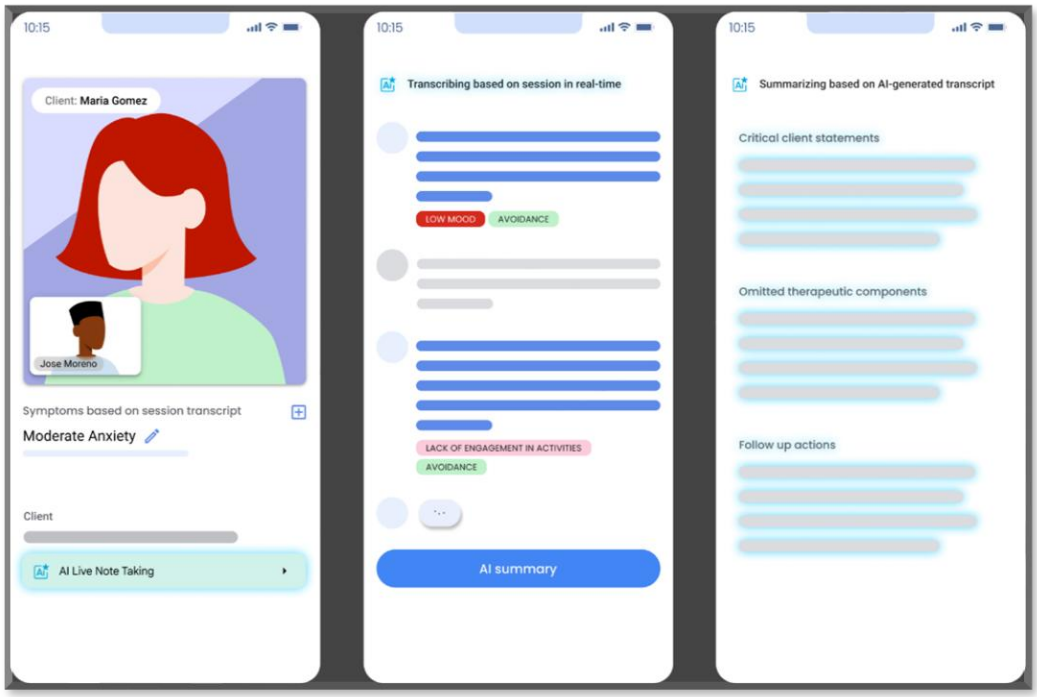
- **High caseload.** With limited staff and high referral volumes, triage can become rushed or superficial and vary in quality. The **triage facilitator** could streamline and standardize the process by automatically flagging high-risk cases and generating structured handover reports. The tool can ultimately help reduce administrative tasks and provide more accurate, standardized assessments of the severity of clients' mental health.
- **Appropriateness of care.** Appropriate triaging could help ensure that clients get the right level of care. Clients who benefit from task-sharing programs will be assigned to those, while clients needing higher levels of support will be referred to appropriate professionals.

How could the end user(s) benefit from this solution?

The primary end user who could directly benefit from this solution is:

- **Care provider.** Care providers who conduct client intake and triage assessments could benefit from real-time risk assessments and structured summaries of clients' mental health. Care providers could then ensure accurate severity evaluations and timely connections to specialized care.

Use case 8: Documentation assistant



End user(s):

- Care provider
- Supervisor

Description:

An AI-powered documentation assistant transcribes client sessions in real time. It generates structured summaries that highlight regularly occurring themes from the client and flags any missed therapeutic components from the care provider (for example, if they missed any questions they should have asked based on the structured intervention protocol).

Example scenario:

During a session, a care provider listens as a client describes ongoing struggles with motivation and daily routines. As the conversation progresses, the **documentation assistant** transcribes the discussion in real-time and highlights recurring themes such as “low mood,” “lack of engagement in activities,” and “avoidance.”

At the end of the session, the **documentation assistant** cross-checks the notes against the intervention manual and identifies that the care provider missed a goal-setting discussion, which is an important component of behavioral activation in CBT. The AI tool then generates structured session notes for the care provider that summarize critical client statements, outline intervention steps, highlight any omitted therapeutic components (in this case, the goal-setting discussion), and provide follow-up actions suggested for the next session.

After reviewing the notes, the care provider notices that there has been an increased sense of hopelessness in the client’s statements. To ensure timely support, they flag the session to be reviewed by a supervisor immediately and then save the session notes.

Opportunities unlocked:

The **documentation assistant** could address several challenges in mental health task sharing programs, including the following:

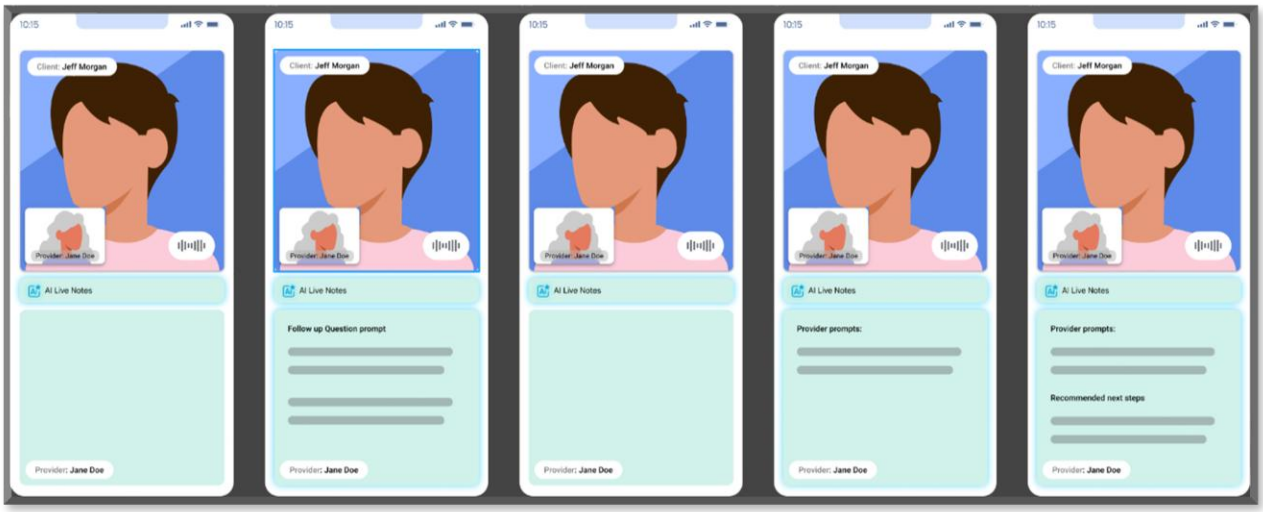
- **Protocol drift.** Care providers may unintentionally skip or alter processes during an intervention, especially in high-stress or high-volume settings. The AI tool could cross-reference session notes with evidence-based protocols to ensure that all essential components are covered and prompt follow-up actions are taken if needed.
- **Limited, irregular session records or sparse documentation.** Traditional paper-based or minimal documentation can hinder continuity and limit real-time feedback opportunities. Additionally, session notes can vary substantially between care providers, leading to inconsistent documentation and gaps in client history. The AI tool could transcribe sessions, organize structured data, and standardize notetaking formats to ensure accurate, detailed, and consistent documentation. This feature could reduce variability in notes, enhance continuity of care, lower the risk of provider burnout from administrative demands, and help supervisors to provide timely and targeted feedback based on comprehensive session records.

How could the end user(s) benefit from this solution?

The primary end users who would directly benefit from this solution are:

- **Care provider.** Care providers could benefit from automated notetaking, structured data organization, and protocol adherence checks. As a result, they could focus more on ensuring clients are engaged and maintaining high documentation standards.
- **Supervisor.** Accurate and consistent session records could help supervisors provide timely and targeted feedback to care providers. The AI tool's structured data could allow supervisors to track progress and provide tailored coaching for the areas where providers need additional support.

Use case 9: Real-time guidance AI companion



End user(s):

- Care provider

Description:

An AI-powered real-time guidance companion analyzes client conversations and provides real-time guidance, then suggests evidence-based next steps (such as follow-up questions) and prompts that will help care providers enhance therapeutic engagement (such as focusing on active listening).

Example scenario:

During a routine counseling session, a care provider listens as the client talks about feeling anxious and overwhelmed in daily life. Unsure how to proceed, the care provider hesitates.

The **real-time guidance AI companion** analyzes the conversation in real time, recognizing meaningful phrases related to anxiety and low mood. It suggests evidence-based next steps to the care provider, such as exploring the client's coping strategies and supporting the client in approaching rather than avoiding anxiety-inducing situations. The companion also tracks the care provider's communication style, noting that they are speaking most of the time. It prompts the care provider to shift to a more listening-focused approach to enhance engagement.

After the session, the **AI companion** generates feedback on how the care provider responded to its suggestions and provides any recommended follow-up actions. The care provider feels more confident, knowing they maintained therapeutic fidelity while receiving real-time guidance.

Opportunities unlocked:

The **AI companion** tool could address several challenges in mental health task sharing programs, including the following:

- **Supervisor overload.** A single supervisor often supports multiple providers, so their response times to urgent queries can be delayed. The **AI companion** could provide real-time guidance to care providers while staying aligned with intervention protocols, which will help reduce the need for immediate supervisor intervention and ease workload pressures.
- **Protocol drift.** Supervisors may not be able to observe sessions directly and ensure that care providers are following protocol during and after the program. The **AI companion** could bridge this gap by offering on-the-spot recommendations and structured guidance so that providers can receive timely support regardless of their location.
- **Variable oversight.** Supervisors may vary in their focus day-to-day, sometimes prioritizing administrative tasks rather than overseeing clinical fidelity. The **AI companion** could consistently reinforce evidence-based protocols, ensuring consistent oversight and high-quality care.

How could the end user(s) benefit from this solution?

The primary end user who could directly benefit from this solution is:

- **Care provider.** Care providers could receive real-time guidance, suggested next steps, and structured feedback, allowing them to navigate complex cases with confidence and competence.

Chapter 4: Success factors

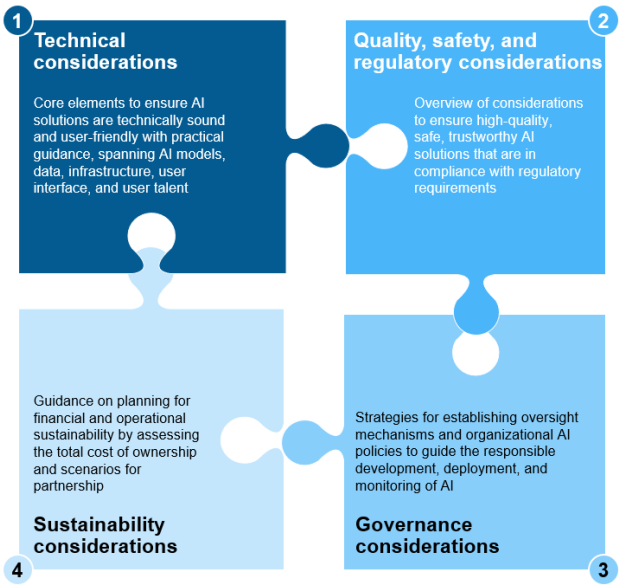
In the previous chapter, we explored potential ways AI could support and strengthen mental health task sharing programs, along with some real-world examples. However, identifying potential use cases is only the first step. Successful implementation of AI in task-sharing programs depends on whether AI solutions can be developed and integrated in a way that is technically sound, clinically safe, ethically governed, and operationally sustainable.

In this chapter, we outline four considerations task-sharing programs should keep in mind to responsibly and effectively implement AI in their programs:

1. ***Technical considerations.*** These core elements use practical guidance to ensure AI solutions are technically sound and user-friendly. Technical considerations include AI models, data, infrastructure, user interface, and technical talent.
2. ***Quality, safety, trust, and regulatory considerations.*** These considerations ensure high-quality, safe, and trustworthy AI solutions that are in compliance with regulatory requirements and meet best-in-class standards for responsible AI use.
3. ***Governance considerations.*** These strategies establish oversight mechanisms and organizational AI policies to guide the responsible development, deployment, and monitoring of AI.
4. ***Sustainability considerations.*** These considerations provide guidance on planning for financial and operational sustainability by assessing the total cost of ownership and scenarios for partnership.

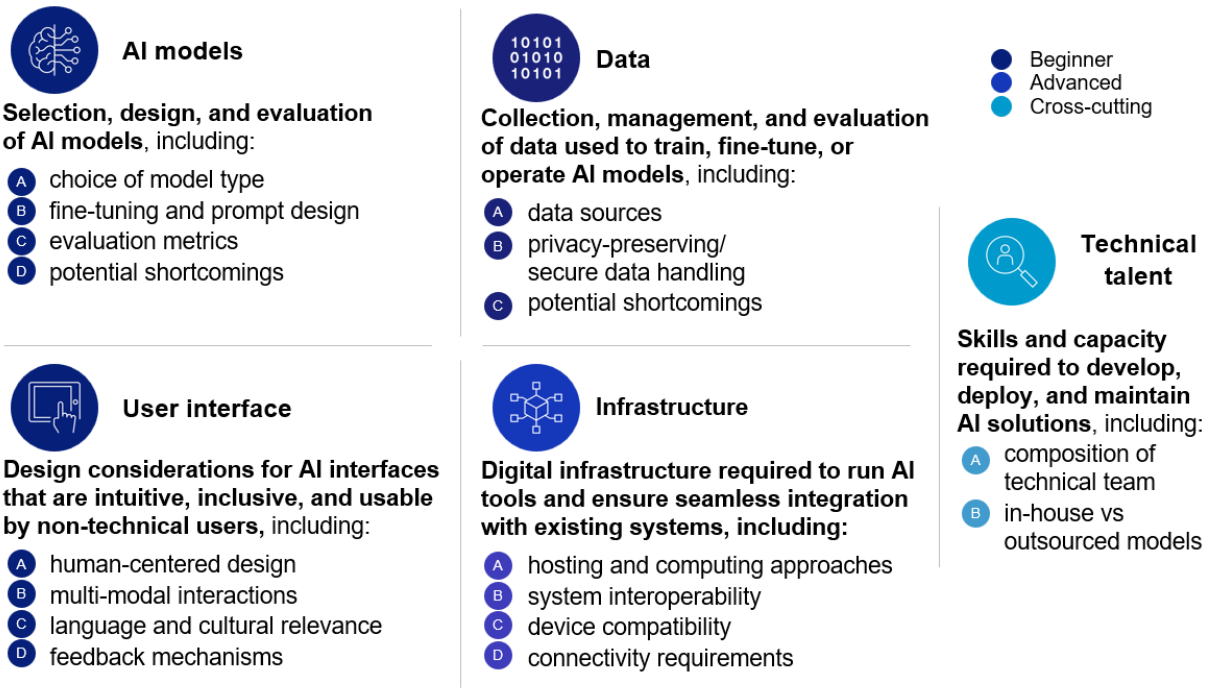
All four of these considerations are fundamental to integrate and scale AI in a way that enhances the delivery of evidence-based mental health services in task-sharing programs (Exhibit 3).

Exhibit 3



Section 1: Technical considerations

Successfully implementing AI in mental health task-sharing programs requires more than just the right AI model—it requires a robust technical ecosystem foundation and system. This section outlines five dimensions that should be considered to develop and adapt technically sound, user-friendly AI solutions:



AI models

Choosing a model

When considering AI for mental health task sharing, programs should understand which AI models are available to them, how to design prompts to receive the best output, and how to evaluate the responses received. Generally, two AI models are used most frequently proprietary models and open-source models.:

Proprietary models. Proprietary models are commercially developed and maintained by private companies, such as Anthropic (Claude), Google (Gemini), OpenAI (GPT), and Meta (Llama). These models are typically accessed through user interfaces or APIs. They offer high performance, reliable time during which the solution is up and running, and dedicated customer support for developers, making them appealing to programs seeking to adapt AI with minimal setup.

Programs should be aware that these models often come with usage-based pricing models, so the cost could rise quickly with ongoing use. Additionally, their closed-source, private nature limits the ability to customize the models for their specific needs. For programs prioritizing ease of use and reliability over customization, proprietary models may be a suitable choice, but they may require programs to budget for recurring operational costs.

Open-source models. Open-source models are freely available AI models developed by the community or technology companies, and they are hosted on various platforms. These models offer flexibility to customize the model for your specific needs, such as adapting it to local languages or workflows.

Programs should be aware that these models may require more technical expertise and infrastructure to deploy and maintain, which may involve additional budgeting for the initial setup. Moreover, for programs

without in-house AI expertise, using open-source models may require partnering with technical collaborators or external vendors.

Practical guidance:

- If your program has limited technical resources but needs high-performance, consider proprietary models.
- Work with a technology partner to evaluate and select the most appropriate model for your needs.
- Begin with smaller, simpler implementations before scaling to more-complex applications.
- Consider security and privacy requirements specific to mental health data.

Implementers are encouraged to examine differences between example AI models to help programs choose the approach that best fits their technical capacity, customization needs, and budget. Bear in mind that these model capabilities and offerings may evolve over time.

Feature options for programs to consider when selecting an AI model		
Type	Proprietary	Open-source
Parameters	Very large	Variable
API access	Yes	Community or direct
Cost	Usage-based	Free
Multimodal capabilities	Yes	Limited text only
Customization	Fine-tuning via API	Full customization
Support	Dedicated developer support	Community support
Performance	Very high	Variable
Safety	Advanced safety features	Variable depending on implementation
Size	Very large	Variable
Adaptability	High	Very high

Key definitions

Parameters: Variables that a model learns during training and that determine how the model processes input to generate output.

API (application programming interface): A set of rules and protocols that enables software applications to communicate with each other to exchange data, features, and functionality. In the context of AI models, an API enables developers or other software to send input (for example, text) to a model (for example,

GPT-4) and receive output (for example, written responses) in real-time, without needing to host or manage the model themselves.

Multimodal capabilities: A model's ability to understand and generate multiple types of data, such as text, images, or video.

Support: Availability of assistance (via resources provided or others) to developers using the model and who provides support for implementation

Fine-tuning and prompt design

Once the AI model is selected, programs can optimize their performance for mental health task-sharing scenarios through fine-tuning or prompt design.

Prompt engineering. Prompt engineering involves structuring the inputs given to the AI model to elicit the most useful and context-appropriate responses. Prompts could include explicit instructions (for example, "respond in an empathetic tone") to guide the model.

Fine-tuning. Fine-tuning is when a selected model is trained using your own data. For example, in the case of task-sharing programs, models can be trained using anonymized transcripts of counseling sessions. This process allows the model to learn domain-specific terminology, program-specific protocols, and possible cultural nuances to improve its performance in different contexts. Programs should be aware that fine-tuning typically requires access to a large volume of high-quality and well-labeled data, computational resources, and technical expertise (learn more).

Example fine-tuning strategies to improve and customize AI models:

Parameter-efficient fine-tuning (PEFT). In this approach to customizing AI models, only a small subset of the model's parameters are updated, rather than the entire model. Techniques such as low-rank adaptation (LoRA) allow the model to adapt to new tasks or domains with lower computational cost in terms of resources required, making this a practical option for programs with limited technical infrastructure or data.

Full fine-tuning. Full fine-tuning is a more comprehensive approach to model customization that involves updating all of the model's parameters using task-specific data. This method can yield the highest levels of performance and alignment from AI models—especially for complex or sensitive use cases—but it requires substantial computational resources, access to large annotated data sets, and advanced technical expertise. Full fine-tuning is typically suited to programs with dedicated AI development teams and robust infrastructure.

Practical guidance:

- Use prompt engineering to test and refine base AI models before committing to fine-tuning.
- Consider fine-tuning if your use case requires a lot of customization.
- Be aware of any harmful biases in your training data before you start on any fine-tuning process to prevent the model from reinforcing these harmful biases.

Evaluation metrics

AI models evolve as underlying data shifts and usage patterns change. Continuously evaluating the performance of these models is essential to ensure that models are not only technically accurate but also clinically appropriate.

Programs could assess the performance of AI models using a combination of conventional benchmarking metrics and domain-specific metrics that reflect the models' intended use (for example, their adherence to therapeutic protocols).

Common benchmarking metrics include the following. These constitute information that the person in the team with technical competency in the topic would review:

- **Accuracy:** the proportion of correct predictions made by the model
- **Precision:** the proportion of accurate results among all positive predictions to understand how often the model is correct when it makes a positive prediction
- **Recall:** the proportion of true positives correctly identified by the model
- **F1 score:** the harmonic mean of precision and recall, balancing both metrics
- **Latency and throughput:** respectively, measures of how quickly a model returns results and how many tasks the model can handle at a given time, both of which are critical for real-time interventions such as in-session feedback

Practical guidance:

- Define domain-specific success metrics that align with your use case goals (for example, protocol adherence).
- Use standard benchmarks alongside human evaluations (for example, providers or supervisors) to ensure technical and contextual quality.
- Ensure that your AI governance committee periodically reviews evaluation metrics and compares them to benchmarks to maintain safe and reliable model performance.
- Continuously refine benchmarks using feedback from diverse users and deployment data.

Key definitions

AI governance committee: A multidisciplinary committee that is responsible for setting the governance policies and guidelines, ensuring regulatory compliance, supervising the creation and implementation of AI tools, managing and mitigating the risks, and establishing transparency and accountability with the stakeholders involved. This committee should be established as part of the AI governance structure in your program. More details are covered in "Section 3: Governance considerations

Potential shortcomings

Despite best efforts, AI systems could produce flawed outputs. Programs could proactively manage three common shortcomings of AI models—bias, hallucinations, and susceptibility to adversarial attacks—to maintain user trust and solution integrity.

- **Bias** occurs when AI models are trained on data sets that fail to adequately represent the communities that the task-sharing program serves. Bias could cause outputs to be less accurate or culturally inappropriate.

- **Hallucinations** are responses that are generated by a model and sound plausible but are factually incorrect or misleading. For example, a model could fabricate statistics that could confuse users and decrease their confidence in the solutions using AI.
- **Susceptibility to adversarial attacks** refers to how a model can be manipulated through crafted inputs to generate unsafe, biased, or harmful outputs (adversarial attacks). These attacks could bypass safety filters and pose risks when AI tools are deployed in an unsupervised manner.

Practical guidance:

- Involve diverse stakeholders in data sourcing, annotation, and testing to reduce bias.
- Fine-tune and evaluate models with local, real-world examples to improve relevance and reduce hallucinations.
- Incorporate monitoring systems and automated filters to detect and block adversarial prompts.
- Clearly communicate AI limitations to users, emphasizing AI's role as a copilot in their work, not a decision-maker.
- Ensure that your AI governance committee routinely audits models and updates safeguards as threats evolve.

Data

AI models for task-sharing programs are only as good as the data they have. Programs should vet their data sources, safeguard their data, and understand any shortcomings their data may have.

Data sources

The most appropriate data sources depend on the intended AI use case. For mental health use cases, data sources may include anonymized counseling transcripts, structured session notes, supervision notes, cases or notes generated by experts, or screening tool results.

Task-sharing programs could aim to use data that is representative of the populations they serve—across age, gender, geography, language, and culture. Data collected internally during client sessions is often the most relevant, but it must be collected with appropriate consent and ethical safeguards. Publicly available or third-party data sets could also be used for AI models; however, it is important to account for potential gaps in population representation, limitations in quality, or potential biases introduced in the original collection process.

Practical guidance:

- Where possible, use in-house or contextually relevant data collected via informed content to maximize cultural and contextual alignment.
- Implement quality checks for data accuracy and completeness.
- Evaluate public data sets for representativeness and population relevance before use.
- Clearly document data sources, limitations, and intended use cases.
- Create data governance policies and put them into practice as part of your program's AI governance system *(see "Section 3: Governance considerations" for more details)*.

Privacy-preserving, secure data handling

Due to the sensitive nature of mental health data, programs should implement strong privacy and security measures that go beyond the regulatory minimum. Programs should, for example, anonymize and de-identify data whenever possible, restrict access to data using role-based permissions, and establish clear data retention and deletion policies. Programs are also advised to check policies of private LLMs or open source technologies used, to clarify whether data that is gathered might be utilized for other purposes, such as training of larger models, given this may be in conflict with consent rules for this type of information.

When working with a technology partner, programs could establish binding agreements that mandate strong data protection measures and compliance with local regulations (such as GDPR or HIPAA). Additionally, all staff interacting with mental health data could receive training not only on technical protocols but also on broader ethical considerations around confidentiality and harm prevention.

Practical guidance:

- Anonymize or de-identify personal data wherever possible.
- Encrypt data during storage and transmission.
- Restrict data access using role-based permissions.
- Define, regularly review, and operationalize clear data retention and deletion policies.
- Train all staff on data privacy, security protocols, and ethical handling of sensitive data.
- Conduct routine privacy risk assessments and third-party compliance audits.

Potential shortcomings

Even with strong sourcing and processing practices in place, limitations in data could compromise AI's reliability, safety, and fairness. For example, data that underrepresents certain subgroups, such as adolescents and non-dominant language speakers, may result in biased or irrelevant outputs. Incomplete, outdated, or inconsistently labeled data could further reduce model performance in real-world applications.

Practical guidance:

- Audit data sets for quality, completeness, and subgroup representation.
- Monitor model outputs across populations to identify bias or performance gaps.
- Refresh training data regularly to reflect evolving language and care practices.
- Use deployment feedback to refine data pipelines and address issues early.
- Embed ethical review in AI governance processes to reduce harm and maintain trust.

User interface

The user interface (UI) is the part of an AI solution that users see and interact with on their screens. A good interface helps users understand what AI is doing and how to use it while making them comfortable working with these tools.

Users accessing mental health task-sharing programs may not be tech-savvy or may be using older devices. Therefore, UI should be simple, intuitive, and culturally appropriate. If the solution is confusing or hard to navigate, it could lead to frustration and low adoption even if the AI model behind it works well.

When selecting or designing the interface of AI solutions, programs should think carefully about human-centered design, multimodal interactions, language and cultural relevance, and feedback mechanisms.

Human-centered design

Human-centered design is an approach that focuses on users, their needs, and their real-world contexts to make solutions as useful as possible. For mental health programs adopting AI solutions, interfaces of these solutions should be designed based on how care providers, trainers, and supervisors would think and work while using these tools.

To be as effective as possible, programs could involve frontline providers, supervisors, and other intended users in the design process and ask for feedback early on. Their input could help ensure that the solutions support actual workflows, aligns with users' way of thinking, and reduces their mental load.

Ideal interfaces would have simple layouts, clear step-by-step flows, easy-to-understand buttons, and helpful visual clues, such as icons and color indicators.

Multimodal interactions

In mental health programs, users may be using mobile devices in their second language, have slow internet connections, or have limited experience using AI solutions.

To make the AI solutions more accessible and usable for this audience, interfaces could support multiple ways of interacting. For example, some providers may prefer typing, while others might find voice input more intuitive. For users with low literacy, visual aids such as icons, color-coded buttons, or images could be helpful so they do not have to rely on comprehending complex text.

Language and cultural relevance

In mental health programs, how information is presented by the AI solutions could be as important as what the information is saying. Providers in these programs often deliver care in local dialects and are trained to use empathetic, culturally attuned language when supporting clients. If an AI solution uses overly formal, clinical language, users may feel disconnected from the tool, which could lead to lower adoption rates.

For the best results, interfaces of AI solutions should be designed with plain, context-aware language that aligns with local providers' communication style and the communities they serve. This may include the following:

- avoiding medical or technical jargon on the interface.
- matching the tone and phrasing used in local training materials.
- ensuring prompts or outputs feel natural and easy to deliver to clients.

The AI solutions could allow for switching between multiple languages or dialects in relevant regions to reflect how care providers adapt in real time, depending on their clients' needs. In these scenarios where multiple languages or dialects are to be made available, it is critical to have those from the community or relevant population involved to obtain their input on language, prompts, and so on.

Feedback mechanisms

In all stages of the AI life cycle—development, deployment, and maintenance—continuous feedback from frontline users is essential to improve the performance of AI solutions and increase trust in these technologies.

Mental health programs could build feedback mechanisms into these AI solutions to evaluate users' experience, such as "thumbs up" and "thumbs down" buttons, a flag for "this recommendation was not right," or an emoji-based rating. They could also have more-formal, regular check-ins with users to understand how well the tool is working in the field.

These feedback mechanisms could also create a sense of ownership and collaboration among care providers to make them feel like partners in building these tools, not just end-users.

Practical guidance:

- Co-design with real users (providers and supervisors) to reflect local realities.
- Use visual guides, voice support, and simple prompts to make tools usable across literacy levels.
- Adapt language and cultural tone to match the way care providers speak with clients.
- Include quick, intuitive feedback options in the solutions along with structured, feedback check-ins with users.

Infrastructure

Infrastructure refers to the digital backbone that enables AI systems to operate reliably and securely at scale and integrate into existing software used in healthcare systems. Infrastructure is especially relevant for programs that have already implemented at least one AI solution.

For AI solutions relevant to mental health, infrastructure considerations include choices in hosting and computing environments, system interoperability, device compatibility, and connectivity requirements.

Hosting and computing approaches

Hosting and computing refers to where and how the AI system processes information and generates responses. To ensure that AI solutions are fast, secure, and reliable in real-world settings, programs should consider where the AI model is stored and accessed, and where the data is processed:

- *Where the AI model is stored and accessed (hosting).* An AI model could be located on an external server (for example, hosted on a cloud platform), could be embedded directly within the application on the user's phone or tablet (known as "local," "on-device," or "edge" deployment), or could be a combination of both (known as "hybrid hosting").
- *Where the data is processed (computing).* An input from the user (via text, image, or voice) could be sent to an external server (the cloud), could be handled directly on the user's phone or tablet (known as "local," "on-device," or "edge" processing), or could be a combination of both (a hybrid computing model).

There are three approaches to hosting and computing. Each approach—cloud, edge, hybrid—has trade-offs in terms of scalability, speed, and privacy:

- *Cloud-based approach.* A could-based approach allows AI models to be hosted and run on powerful cloud-based, external servers that are accessible over the internet. These solutions are scalable and easy to update, and they reduce the need for programs to invest in and maintain local infrastructure. However, cloud-based solutions may raise concerns in settings with strict data residency laws, which regulate the location of where data is stored (for example, some servers may

require data to stay within the country of deployment) or in places where the internet connections are unreliable or slow.

- *Edge approach.* An edge approach allows AI models to be run directly on the user's device (such as a smartphone or tablet). This approach could enable faster response times and greater privacy because data doesn't have to leave the device. However, the AI models need to be smaller in size and efficient enough to run on local devices, and devices need to have enough storage and processing power.
- *Hybrid approach.* A hybrid approach combines the strengths of both a cloud-based and an edge approach. For simpler tasks, the model is run on devices and data can remain on the device; for more complex tasks, the cloud servers are used when needed. This approach offers more flexibility and may be the ideal choice for task sharing programs looking to implement AI for a variety of tasks in environments with low connectivity and strict data residency requirements.

Practical guidance:

- Review data privacy regulations and the storage and processing power of user devices when choosing where your models will operate.
- Consider using the cloud-based approach during pilot phases to easily develop and update your models.
- Consider transitioning to the edge or hybrid approach if your users have poor internet connectivity or if local data privacy laws restrict data sharing.

System interoperability

System interoperability refers to how easily an AI solution can connect and work with other digital tools that are already in use in health systems, such as appointment scheduling systems and electronic health records. Seamless interoperability is achieved when AI is integrated into existing tools and workflows without being disruptive to frontline workers.

This integration is typically made possible through application programming interfaces (APIs), which are sets of rules that allow software applications to communicate and share information securely. For example, a training curriculum interface with an e-learning algorithm might use an API to retrieve a care provider's previous feedback reports to identify gaps in their knowledge and skill set. Choosing AI models with open, well-documented APIs could help reduce the setup time and technical burden of AI solutions.

Additionally, AI solutions that follow widely adopted health data standards, such as Fast Healthcare Interoperability Resources (HL7 FHIR), could be easier to integrate with other digital health tools within health systems.

Practical guidance:

- Select AI models that offer open and well-documented APIs as much as possible.
- Plan integration with your existing digital systems early in the process with in-house IT teams or technology partners.
- If you are planning to adopt ready-to-use solutions and looking to integrate with the broader health system later, prioritize tools that follow international data standards such as HL7 FHIR.

Device compatibility

Device compatibility refers to how well an AI solution works across different types of devices that users are likely to use. Devices used for mental health programs could include entry-level smartphones, tablets, or laptops.

In low-resource settings that mental health programs may operate in, these devices may have limited processing power, memory, or storage to run AI models. Therefore, it is essential to test AI solutions on a variety of devices to ensure that they work smoothly regardless of operating system, processing power, memory, or storage.

Developers could also design solutions that are device-agnostic and could provide clear information about minimum hardware requirements (namely the operating system, processing power, memory and storage necessary to use the solutions) so that programs can update the AI solutions or their devices accordingly.

Practical guidance:

- Avoid as much as possible AI solutions that require a lot of battery power or memory if operating in a low-resource setting.
- Provide clear guidance to the people implementing the AI solutions on the minimum and recommended device specifications.
- Test AI solutions on devices that reflect what care providers and supervisors use in the field.
- If possible, include an offline installation package for areas with limited internet access (more information on offline and low-connectivity modes below).

Connectivity requirements

Communities in low-resource settings where task sharing programs may operate could face intermittent or low-bandwidth internet connectivity. As a result, AI solutions that require a constant internet connection may become unavailable and ineffective in these settings.

To maximize usability, AI solutions should be designed with offline or low-connectivity modes. For example, data could be temporarily stored on a device and sync with the cloud when the internet becomes available. For time-sensitive use cases, such as an AI provider companion, offline access could be crucial.

Practical guidance:

- Assess internet availability and speed in the environments where AI solutions will be deployed.
- Prioritize solutions that could function offline or with low bandwidth.
- Provide guidance for users on what features require the internet and what can be used offline.

Technical talent

In addition to setting up the technical aspects of an AI model, implementing AI for mental health programs requires assembling a team with the right skills to design, develop, deploy, and maintain these AI solutions in settings that often have limited resources. Regardless of where your program is in its AI journey, the talent aspect is crucial.

Composing the technical team

While large-scale AI deployment may involve many specialists, a lean, well-coordinated development team could also be efficient and effective for mental health programs. One critical consideration is that the

technical team must work in close collaboration with a team of qualified clinicians—including doctors of psychology (PsdDs), doctors of philosophy (PhDs), medical doctors (MDs), licensed clinical social workers (LCSWs), licensed mental health counselors (LMHCs), or licensed professional counselors (LPCs)—who can help to ensure validity, safety, clinical best practices, and ethical application of the AI within the sensitive context of mental healthcare. Clinicians should be involved in defining appropriate use cases; simulating real-world cyber-attacks and evaluations to ensure quality, safety, and clinical appropriateness, and determining success criteria to launch. In this context, a lean AI team could include the following five roles:

- ***Product manager.*** A product manager connects the leadership of task-sharing programs, users of AI features, and the development team. They define success metrics, oversee the entire development process end-to-end, coordinate across technical and operational teams, and ensure that the solutions align with both user needs and program objectives.
- ***Machine learning (ML) engineer.*** An ML engineer uses raw data to design, fine-tune, and integrate AI models into user-facing applications. While the ML engineer's primary focus is to build AI models, they may need to design, build, and maintain data pipelines in circumstances where there is no built-in data interface or resources to hire a data engineer. (In large teams, these responsibilities will fall on data engineers.) The ML engineer ensures that data quality, accuracy, consistency, and scalability are upheld before AI models are built.
- ***UX/UI designer.*** The user experience (UX)/user interface (UI) designer creates simple, intuitive interfaces that reflect how users think and interact with digital solutions. They also focus on accessibility, multilingual support, and cultural relevance where appropriate.
- ***Software engineer.*** The software engineer builds the actual user-facing AI solution—whether it's a mobile app, web dashboard, or offline-compatible tool. They are responsible for integrating AI models into the product, developing secure APIs for the solution's ability to connect and communicate with other entities in a coordinated manner, and ensuring that the solution works on end-users' devices. In the mental health task-sharing context, the engineer could benefit from being familiar with optimizing apps for low-bandwidth and older devices.
- ***User support specialist.*** The user support specialist trains end-users (care providers and supervisors) on the AI model, responds to user requests, and shares user feedback with the development team so it can improve the AI solutions.

These roles could be adapted based on project scope and at times may be owned by fewer than five people. In early phases, one person may take on multiple roles if the task doesn't require specialized knowledge—for example, a product manager could also serve as a user support specialist. However, more-advanced programs may need to expand their team or outsource certain tasks.

In-house versus outsourced models

Once teams select a use case to implement, programs can decide whether to build and manage the AI solution internally, outsource certain roles or tasks, or combine both approaches. Each model comes with trade-offs in cost, control, speed, and sustainability.

- *In-house model.* Hiring an internal team offers the program greater control over design choices, data privacy, and long-term adaptability. This ownership would be especially important for programs working with sensitive mental health data and aiming to build capabilities that generate actionable insights from data. However, building an in-house model requires more up-front investment, especially in recruiting, onboarding, and retaining talent, so this may not be feasible for small programs. Further details are covered in "Section 4: Sustainability considerations."

- *Outsourced model.* Outsourcing the end-to-end development of AI solutions to a technology partner could help task-sharing programs move faster, especially if highly specialized expertise (such as advanced ML engineering or secure backend development) is needed. However, working with a technology partner could introduce challenges around contextual relevance (for example, they may not be as in tune with cultural sensitivities of the user population), communication gaps, financial and operational sustainability, and additional considerations to adequately address privacy.
- *Hybrid model.* Many task-sharing programs could benefit from a hybrid approach to building AI models. For example, a technology partner could build the AI solution including the model within it, while in-house staff could lead product management, implementation in the field, and user support. This approach balances technical depth with contextual alignment and allows programs to gradually build internal capacity over time, which is ideal.

Practical guidance:

- Decide whether to partner with a technology company or build your AI model in-house.
- If the hybrid model is chosen, define roles and responsibilities for both parties.
- Choose technology partners who have experience in mental health and/or digital health in low-resource settings.
- Maintain clear documentation throughout development to support future scaling and reduce vendor dependence.
- Designate internal owners early to ensure the program can adapt and maintain the solution over time.
- Partner with local universities or technology hubs to build technical capacity and reduce long-term costs.

Section 2: Quality, safety, trust, and regulatory considerations

Mental health skill-building programs looking to integrate AI solutions into their programs should think about how to ensure the effectiveness of interventions (quality), the prevention of harm (safety), the confidence of users and communities (trust), and regulatory compliance. These considerations become even more critical in settings where users may not be familiar with AI solutions and the potential risks that may come with them.

In this section, we explore how to ensure high-quality AI solutions, uphold safety through risk mitigation, build trust via ethical and transparent practices, and maintain regulatory compliance.

Ensuring quality

In healthcare, quality refers to the caliber of services that help in improving or reaching desired outcomes. In the context of mental health task sharing, quality could refer to the degree to which AI solutions enhance care providers' ability to serve clients, support effective training and supervision, and improve overall program delivery in line with evidence-based standards. Given these AI solutions are designed to support care providers rather than directly interact with patients, ensuring quality involves work in six areas:

Provider-facing utility and effectiveness. AI solutions should be validated in real-world settings through pilots before they are scaled widely. During this validation, solutions should demonstrate improvements in care provider training, adhere to protocols, or support decision-making in a way that ultimately contributes to better program outcomes. For AI solutions that offer assessments or recommendations (such as the triage facilitator), the performance of the solutions should also be benchmarked against clinical standards when they are being validated, following implementation.

Contextual and workflow alignment. AI solutions should be adapted to the care delivery model followed in the program, care providers' workflows, and local languages and norms. For example, a real-time guidance AI companion should use culturally appropriate language and reflect relevant intervention protocols used in the program. To enable a culturally adapted AI solution and avoid misinterpretations, programs should work with community stakeholders to adapt content and ensure that AI's training data includes local context if possible.

Continuous refinement. AI solutions should be designed to allow for ongoing refinement. Developers should establish mechanisms for user feedback, performance monitoring, and iterative updates over time. Programs should treat AI solutions as living tools that evolve and should plan to audit the outputs and outcomes of the AI solutions periodically.

AI model evaluation. AI solutions should be evaluated for safety and how well they are fulfilling their intended use, which requires programs to continuously evaluate an AI model's performance using relevant metrics (such as accuracy or reliability). Establishing feedback loops can also ensure ongoing quality and identify areas for improvement.

Benchmark model performance. AI solutions need to be benchmarked against relevant clinical standards and existing clinical practices to validate its effectiveness and demonstrate its added value in supporting care providers.

Define model success metrics. AI solutions should have clear measurable metrics to define the success of the AI models in achieving their intended purpose. Such metrics might include, an acceptable percentage of correct or appropriate recommendations made, target average satisfaction rating provided by users

regarding their experience with the AI solution, or evidence of all AI outputs meeting ethical guidelines and complying with relevant mental health regulations and data privacy laws.

Upholding safety

As mental health programs begin to adopt AI, ensuring safety becomes a shared responsibility across developers, implementers, and care providers. Upholding the safety of an AI solution means proactively identifying risks, minimizing risks, and putting safeguards in place to avoid preventable harm. To meet this standard, programs should take three steps:

Conduct risk assessments and maintain ongoing vigilance. Before implementing AI solutions, programs should conduct structured risk assessments to identify how and where AI could fail or cause harm. After deployment, this vigilance should continue. Programs should track moments where the AI has had unintended consequences (for example, if the client drops out of the program or if they are distressed by the tool) through regular safety audits and feedback systems. Some potential risks to be aware of are covered in "Section 1: Technical considerations."

Set up mitigation strategies and fail-safes. Programs should have a mitigation plan for each risk identified in the risk assessment. Some fail-safe mechanisms that programs could establish to mitigate risks are escalation pathways, human-in-the-loop oversight, and clear boundaries for AI's role. An example escalation pathway is if an AI solution detects high risk (a user indicating potential suicidal thoughts, for example), it could immediately alert a supervisor. Moreover, in a human-in-the-loop oversight process in task-sharing, a supervisor overseeing care providers could review AI's suggestions in difficult cases and ensure that suggestions are appropriate. This helps if AI misses nuance—care providers can apply their own judgment rather than automatically following AI's prompts in these situations. Finally, programs should inform care providers with disclaimers and onboarding materials that AI is not a licensed counselor to reduce overreliance and avoid "therapeutic misconception."

Foster a culture of safety. Programs should cultivate an environment with care providers where safety concerns are openly discussed. Programs could train providers on limitations to AI and how to responsibly use AI, encourage community feedback, and create transparent mechanisms for logging and addressing any issues. Safety should be an integral, evolving part of programs' operations and culture.

Building trust

Trust is the foundation that allows users and communities to feel confident using AI solutions in their workflows. If people do not trust the AI solution or the practices followed by the organization behind it, they could feel resistant toward the initiative, and the potential impact of the solution would go unrealized. Therefore, programs should actively build and maintain trust by making AI solutions more transparent, respecting users' rights and privacy, ensuring the technology is available and accessible to all users, and installing human oversight and accountability throughout the process. Substantial elements of establishing and upholding trust include the following:

Explainability and transparency. People are more likely to trust AI if they understand what it does and why it does it. Wherever possible, AI solutions should provide clear explanations for their recommendations. For example, if an AI triage solution suggests that a client may have depression, it should indicate the factors (the symptoms the client exhibited, for example) that led to that conclusion in understandable terms. Additionally, care providers using the solutions should be trained in how the AI works, its reliability, and its limitations. This way, care providers can confidently interpret AI outputs. This openness could reduce misinformation and concerns about AI.

Data privacy and consent. Mental health data is highly sensitive, so strict privacy protections are nonnegotiable. Programs should follow global data protection standards—for example, they should collect

the minimum amount of data necessary, store data securely in an encrypted way, and ensure users know what data is collected and how it's used. Since client data is being used to train AI models or being analyzed in many of the AI use cases outlined in this field guide, programs should obtain consent for data usage from patients and disclose to clients that AI is being used in any provider–client interactions. Communities could be more receptive to AI solutions when they see that their confidential information is safeguarded and their dignity is upheld.

Technology availability and access. AI solutions should be developed and implemented to serve all segments of the population without bias or discrimination. Therefore, programs should ensure that AI solutions work effectively for different demographic groups across gender, ethnicity, language, and socioeconomic status, and that the benefits of these solutions are broadly and fairly distributed. For example, if internet connectivity is an issue in remote communities, an offline or low-bandwidth version of the AI solution could be offered. Ensuring availability and access of the technology promotes trust because people see the initiative as just and inclusive, rather than favoring certain groups.

Human oversight and accountability. An essential part of building trust in AI solutions is assuring users that humans remain in charge of decisions and accountable for mistakes that may be made by AI solutions. Programs should ensure that there is always a qualified human overseeing the AI's operations and available to take care of issues that are escalated, if needed. Programs could set up governance structures to oversee all AI solutions from development to implementation. Governance structures are covered in "Section 3: Governance considerations."

Global guidelines for AI uniformly emphasize resolutions such as: "do no harm", ensure effectiveness, be transparent, ensure fairness, protect privacy, and keep humans in control. These principles should be the backbone of any AI solution implemented in high-risk settings including healthcare and mental health. How to integrate these principles into a task-sharing program's AI journey through a robust governance structure is covered in detail in "Section 3: Governance considerations."

Maintaining regulatory compliance

Integrating AI solutions into mental health task-sharing programs presents both immense opportunities and complex regulatory challenges. As outlined in "Chapter 3: Use cases," AI could enhance various stages of mental health task-sharing programs, from trainee selection and training to intervention and supervision. When implementing an AI use case from this field guide, it is critical to understand its classification to maintain regulatory compliance.

The use cases described in this field guide illustrate how AI could move beyond simple administrative support and into areas that could directly influence client care and clinical decision-making. As a result, some AI solutions mentioned might be classified as ***software as a medical device (SaMD)*** *or* ***medical device software (MDSW)***. These classifications could substantially increase the likelihood of regulatory oversight.

Software as a medical device or medical device software is software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device. "Medical device" in this context means any instrument, apparatus, implement, machine, appliance, implant, reagent for in vitro use, software, material or other similar or related article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purpose(s):

- diagnosis, prevention, monitoring, treatment, or alleviation of disease
- diagnosis, monitoring, treatment, or alleviation of (or compensation for) an injury
- investigation, replacement, modification, or support of the anatomy or of a physiological process

- supporting or sustaining life
- control of conception
- disinfection of medical devices
- provision of information by means of in vitro examination of specimens derived from the human body

Regulatory bodies such as the **U.S. Food and Drug Administration (FDA)** and authorities overseeing the **EU Medical Device Regulation (EU MDR 2017/745)** have specific requirements and guidance for SaMD (or MDSW, as it is known in the European Union). Their definitions and risk classifications, which dictate the level of required oversight, hinge on the AI tool's function and the potential risk to patients if the tool fails.

Let's look at some starting thoughts for how the use cases from chapter 3 might potentially affect regulatory considerations. Note that this is just an initial set of considerations. The legal team or relevant lawyer should examine and define the full list of items to analyze from a regulatory perspective. This list of starting thoughts is not meant to be used as counsel. It is worth noting that regulation for novel technology, especially as it intersects with impact on health or when it is aimed at lowering distress or improving wellbeing, is in flux, requiring expert counsel for navigation:

- **Applicant screening tool.** This AI tool could assess applicants' suitability for delivering mental health interventions by evaluating their communication skills, and predicting their likelihood of success. If this tool is used alone without human oversight, it could be challenging from a regulatory point of view. However, if a human is making the formal assessment and final decision, this tool would then stay as an administrative training tool.
- **Adaptive training interface and AI training room.** If these tools are used to provide personalized training, assess trainee competency, and simulate client interactions, they could be considered to be providing or influencing clinical training and practice. If the AI tool is providing feedback based on clinical protocols, this may be seen as diagnostic and, therefore, could be subject to medical device regulation. If human oversight is involved to assess the competency of the training and the AI tool is used as a support to their knowledge, the AI tool is likely to be considered a medical device. It is also worth noting that many countries have training program regulation and certification requirements, which can extend to community health workers, as well.
- **Post-session feedback report, supervisor dashboard, and documentation assistant.** These tools could analyze session data, provide feedback on provider performance, and flag any deviations from care protocols. Since these features may be considered as being directly related to the quality of care being delivered, these tools may raise regulatory questions. Regulators may consider these tools medical devices if they influence the quality of care. However, if it is clear that a trained professional is using the AI's output only as a support tool alongside their expertise and that a human is making the final decisions, medical device oversight may not be applicable.
- **Provider–client matching, triage facilitator, and provider companion.** These tools could play a direct role in assigning clients, guiding assessments, and informing the interventions being delivered. Because these tools could influence substantial aspects of client care and clinical decision-making, regulators are likely to consider these tools as medical devices.

It's important to note that regulatory landscapes differ across geographies. A tool classified in one way in the European Union might be treated differently in Australia, Canada, Japan, or the United States. Before starting the development of AI solutions, mental health programs should conduct thorough regulatory diligence in the country or region in which they intend to deploy their AI solutions. This research could help

inform safeguards and oversight that need to be implemented during the development and implementation of the AI solutions.

Of course, having these considerations in place will require you to think about your governance processes, which are procedures that ensure the development and implementation of technology within your organizational structure is safe and effective.

Section 3: Governance considerations

The previous section discussed potential ethical, legal, and practical considerations that may be associated with AI solutions, considering the sensitive nature of mental health data, research, and interventions. We encourage all organizations to establish clear principles for the responsible use of AI. Governance of AI solutions (AI governance) mitigates these risks and enhances the trustworthiness of these technologies.

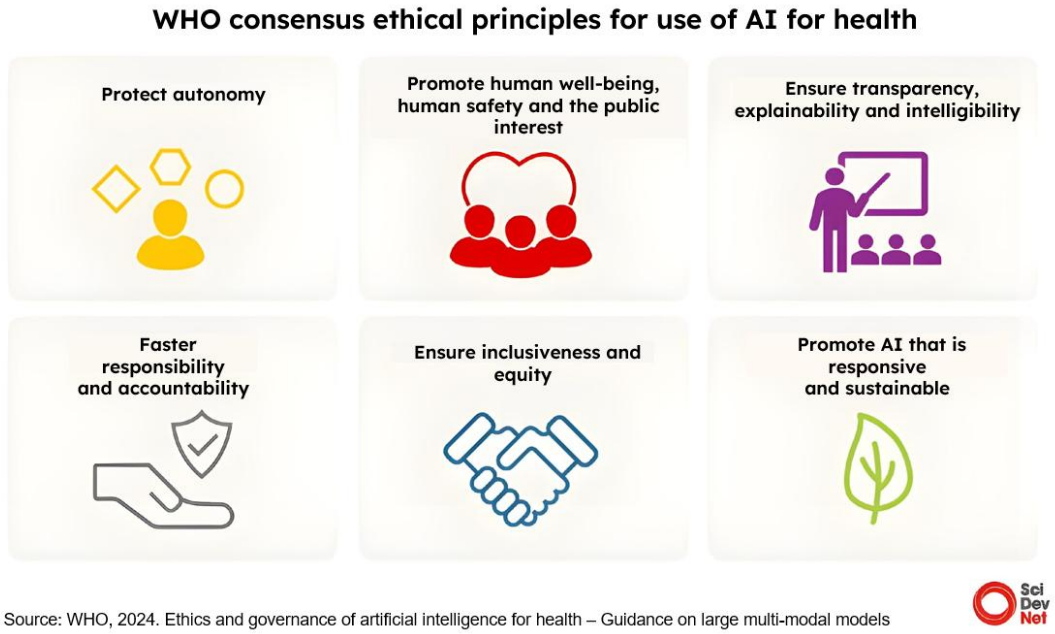
AI governance is **a system of checks and balances that oversees an AI solution throughout its life cycle**—from initial development to deployment, ongoing use, and continuous monitoring. It includes the policies and processes a program uses to review, assess, and manage AI solutions, ensuring their safe, responsible, and effective development while complying with local regulations. A comprehensive AI governance structure in mental healthcare can preserve patient safety, uphold ethical standards, ensure regulatory compliance, foster trust through transparency and accountability, and manage privacy concerns and other legal issues.

Although there is no one-size-fits-all governance framework, there are principles defined by international and national organizations to guide programs in the ethical development and application of AI solutions.

The World Health Organization (WHO) identified six core principles for **"responsible AI"** to guide the governance of AI in healthcare, including mental healthcare (Exhibit 4), though the final decisions to tailor the final framework used would lie on the implementing organization:

- **Protect autonomy.** Safeguard individuals' rights to make informed decisions, ensuring AI does not undermine personal freedom or autonomy.
- **Promote human well-being, human safety, and the public interest.** Prioritize the health, safety, and overall welfare of individuals and society when deploying AI solutions in healthcare.
- **Ensure transparency, explainability, and intelligibility.** Operate in ways that are understandable, with clear explanations for their decisions and operations.
- **Foster responsibility and accountability.** Establish clear lines of responsibility for AI's outcomes, ensuring that developers, users, and organizations are accountable for the technology's impact.
- **Ensure inclusiveness.** Promote the fair distribution of AI benefits, ensuring that AI applications in healthcare are accessible to and optimized for all populations, especially underserved groups.
- **Promote AI that is responsive and sustainable.** Encourage the development of AI that can adapt to changing needs, remain effective over time, and be sustainable for long-term use.

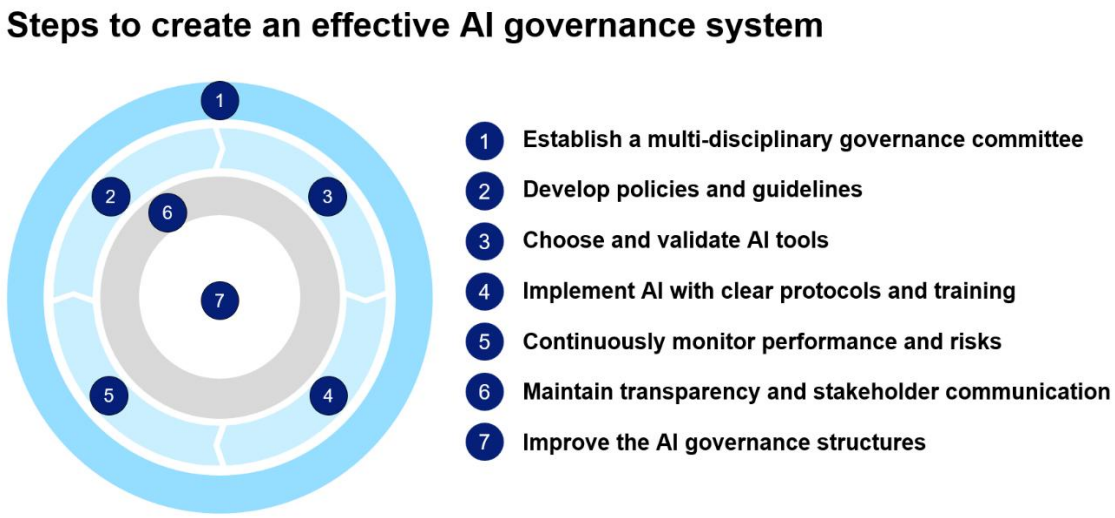
Exhibit 4



Citation: Abdallah Taha, "*AI risks in global health 'must be accounted for' – WHO*," SciDev.Net, January 23, 2024; Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models, World Health Organization, 2024

Establishing and maintaining an effective AI governance system rooted in these core principles requires a structured and practical approach. The following steps outline how a task-sharing program, regardless of its size, could integrate AI governance into its workflows to enable the responsible use of AI (Exhibit 5). These steps gather recommendations from existing AI frameworks originally developed by global authorities such as WHO, the World Economic Forum (WEF), and Coalition for Health AI.

Exhibit 5



Establish a multi-disciplinary governance committee

- **Form the committee** that will create the governance strategy, structure, and processes.
- **Aim for a group that brings diverse perspectives together,** including clinical, technical, ethical, and community perspectives.
- **Define the roles and responsibilities of the committee clearly by** setting the governance policies and guidelines, maintaining regulatory compliance, supervising the creation and implementation of AI tools, managing and mitigating the risks, and establishing transparency and accountability with the stakeholders involved.

Develop organizational policies and guidelines

- **Review existing laws, regulations, guidelines, and frameworks**—such as the general data protection regulation (*GDPR*), *HIPAA*, *or the Assurance Standards Guide by Coalition for Health AI (CHAI)*—to tailor recommendations to the local context and the program's needs. Note that this list is not exhaustive.
- **Consult peer organizations that have implemented AI solutions** to understand their interpretations of the existing resources and regulations and learn from their experiences.
- **Draft AI governance policies and guidelines** to be clear, measurable, actionable, and realistic—for example—*"All AI-generated content must be reviewed by a human prior to inclusion in patient healthcare records."*
- **Make the policies and guidelines accessible to users, and communicate the new rules of engagement with AI** throughout the program.

Choose and validate AI tools

- **Create a checklist of approval criteria when selecting an AI solution to implement—**for example *"Does the solution operate in the local language and consider cultural nuances?"*
- **Conduct a pilot with the AI solution before implementing it widely** to see how the solution performs with real users and data.
- **Work with developers to refine the AI solution** until all requirements set by organizational policies are met

Implement AI with clear protocols and training

- **Create an implementation plan in line with governance policies and guidelines.**
- **Educate users about benefits, risks, and boundaries of the AI solution** (for instance, what it should be used for and what it is not meant to do) **and the protocols set by the governance committee.**
- **Maintain a feedback loop between the governance committee so users** can report back any issues they have with the experience and clarify policies as needed

Continuously monitor performance and risks

- **Define key performance indicators (KPIs) that align with the program's goals**—for example, *"What percentage of referrals suggested by AI did the supervisors agree with?"*
- **Monitor user engagement and any disparities** to see if AI performs equally across the communities the task-sharing program serves and is used by the intended providers.

- **Maintain a log of AI-related incidents to conduct risk assessments—**for example, *"The AI gave a potentially unsafe recommendation, but the care provider caught it."*
- **Update the AI solution and adjust how staff interacts with it if there are risks flagged during monitoring**

Maintain transparency and stakeholder communication

- **Regularly update the internal and external stakeholders involved on how the AI solutions are performing**—for example, *"This quarter, our AI-powered triage solution helped identify ten new patients who needed specialized mental healthcare."*
- **Communicate any changes to the AI solutions and governance policies in place.**
- **Engage with local regulators and policymakers, when possible,** to share on-the-ground experiences.

Improve the AI governance structures

- **Hold retrospective committee meetings to monitor how well the governance processes are working,** and update the processes accordingly**.**
- **Track the evolving regulatory landscape and update organizational governance policies accordingly** to comply with new laws or regulations.

Over time, as programs gain experience and resources, they can refine these structures and adapt to evolving ethical and regulatory standards as well as new technologies. By taking an iterative approach, even smaller programs can develop a governance framework that is both resilient and scalable.

Section 4: Sustainability considerations

Integrating AI into mental health task sharing programs requires a thoughtful approach to ensure long-term financial and operational sustainability. To maintain the sustainability of an AI initiative, programs should treat AI as an ongoing expense, rather than being reliant on one-time grants and budget for its development, deployment, and maintenance.

When adopting AI solutions as part of mental health task-sharing programs, programs should assess the **total cost of ownership** by distinguishing between capital expenditures (the development and deployment costs) and operational expenditures (the maintenance costs).

Total cost of ownership of AI solutions

Multiple factors, including solution complexity, data requirements, infrastructure and technological resources, and user support could influence the total cost of ownership of AI solutions.

To understand the total cost of ownership and effectively allocate resources to manage it, these factors should be examined across three phases of the AI life cycle:

Development phase

This phase includes designing, training, and testing AI models before integrating them into programs' workflows. Programs could choose to fine-tune a pretrained open-source model or adapt a pretrained proprietary AI model for their solutions. This section is meant to be used as an informational guide based on public information and not as spending recommendations by the authors of this resource.

Another option would be to build a mental health–specific custom AI model (an LLM) from scratch; however, this is a costly and time-consuming process. To provide a sense of scale, training a model similar to GPT-

3 was estimated to cost around $4.6 million with the lowest-priced cloud processor in 2020. While there are organizations that are trying to build LLMs for psychology, most digital health companies use vended LLMs.

Since building a custom LLM is not a feasible approach for many task-sharing or similar mental health programs and customization options are limited with proprietary models, we focus on AI solutions that embed pretrained open-source models in the following cost breakdown.

Key definitions

Train: Process of feeding data that is curated to help the model refine itself (that is, learn patterns based on the new data so that its reasoning is updated) for producing accurate responses to queries

Test: Phase following the training where the AI model is evaluated using new data to assess model's accuracy and effectiveness

The cost of developing an AI solution with a pretrained open-source model has the following components:

- **Cost for data acquisition and preparation.** Building a custom AI model, especially an LLM, requires vast amounts of high-quality data. For domain-specific applications, this involves collecting, annotating, storing, and managing large sets of data. For example, in 2025, at time of print, the cost of creating a high-quality training data set can range from $10,000 to $90,000, depending on the nature of data and the complexity of the annotation process. These costs will depend on the wage levels at the project site and can vary substantially.
- **Cost for model training and testing:** Substantial computing resources (such as graphics processing units [GPUs] and tensor processing units [TPUs] may be required to train the LLM on your data set and test it using a different data set (such as anonymized transcripts of counseling sessions). Due to their scalability and flexibility, cloud-based solutions can provide these GPUs and TPUs more conveniently and help manage computational needs. Training costs can vary substantially based on model size and complexity.
- **Other costs.** Beyond data and model development, programs should account for the cost of the talent required to build the AI solutions. At this stage, a lean development team includes a product manager, a machine learning engineer, an UX/UI designer, and a software engineer that works closely with clinical, legal, and regulatory experts. Visit "Chapter 4, Section 1*:* Technical considerations" has role descriptions of these team members.

Considering the factors above, the cost for the development phase of an AI solution using a pretrained open-source model can range from $35,000 to more than $150,000, depending on the project's complexity and specific requirements. This ballpark cost assumes the software development cost for the user-facing application in which AI is embedded would range from $15,000 to more than $100,000.

While this section emphasizes open-source models, task-sharing programs could determine which model type—proprietary or open-source—is better suited to their goals and resources in collaboration with their in-house IT team or an external technology partner.

Deployment phase

Once an AI model has been adapted for use, the next step is to deploy it within the program's technical infrastructure. Once AI models are deployed in existing technical infrastructure, they should be integrated into operational workflows and be accessible to end users through select solutions, including mobile apps or web-based platforms.

The cost of deploying and using an AI model has the following components:

- **Cost of integration.** Integrating AI solutions into operational workflows requires developing APIs, modifying existing software, and ensuring a seamless data exchange between platforms. These costs vary depending on the need for customization and the complexity of integration.

- **Cost of quality assurance.** After integration is completed, the development team ensures the AI solutions function as intended within the programs' workflows. This step includes validating the system's performance for speed, responsiveness, and accuracy and conducting quality assurance to identify and address any technical issues such as bugs prior to full-scale deployment.
- **Cost of training end users.** Adopting AI solutions requires programs to train different groups of end users on how to use AI solutions effectively and responsibly. This training may include onboarding sessions, interactive tutorials, and user manuals. Programs should budget to hire staff or technology partners to deliver this training.

Considering the factors above, the cost for the development phase for an AI solution using a pretrained open-source model can range from $15,000 to more than $40,000, depending on the project's complexity and specific requirements. This ballpark cost assumes the cost—from testing and from quality assurance that confirms the AI is compliant with healthcare regulations—would range from $10,000 to $25,000.

Maintenance phase

After deployment, AI solutions require ongoing monitoring and maintenance to ensure that they continue to perform as intended and adapt to the evolving needs of end users and mental health task-sharing programs.

The cost of monitoring and maintaining a pretrained open-source AI model has the following components:

- **Cost of cloud hosting.** Programs using cloud-based, open-source AI models must budget to host their models on the cloud. Fees may vary based on complexity of model—the larger the size of the model, the higher the cost will be. Many cloud platforms offer scalable options to manage costs.
- **Inference compute costs.** Running an AI model in real-world settings requires computational resources to handle user interactions and generate responses to user queries. The cost of processing these interactions depends on factors such as the number of users, the complexity of the query, and how efficiently the model processes the information.
- **Cost of performance monitoring.** Using monitoring tools to continuously monitor AI solutions is essential to detect and address performance issues such as inaccuracies and model drifts (in other words, when the model's performance gradually declines because of changes in data). Often, cloud-based platforms provide monitoring tools and alerts to identify performance issues early, with an additional cost.
- **Cost of regular updates and improvements.** AI models need periodic updates to incorporate new data, refine algorithms, and improve their accuracy. Associated costs would include fees for cloud computing resources that retrain the tool and fees for human resources (such as a machine learning engineer) to refine models based on changing needs or identified performance issues.
- **Cost of user support.** Ongoing support to users, including training sessions and help desks, should be provided to ensure that AI solutions are used effectively and any critical issues are addressed promptly. Programs should budget for at least one person dedicated to user support and escalate issues to the developer team (a user support specialist, for example).

Considering the factors above, the cost of the maintenance phase for an AI solution using a pretrained open-source model can range from $5,000 to $25,000 annually, depending on the project's complexity and specific requirements. This ballpark cost assumes that the program would spend 10 to 30 percent of their initial development cost annually on maintenance.

Mental health skill building programs could choose to fully own tasks in these phases or fully outsource these tasks to a technology partner. Next, we will cover different scenarios to develop and implement AI solutions in mental health task sharing programs.

Scenarios to develop and implement AI solutions

Deciding whether to build AI solutions in-house or partner with external technology companies is an important part of the process to adopt AI. This decision will likely vary for each program depending on its goals, use cases, technical and operational capacity, and financial resources.

Mental health task-sharing programs, which often operate with limited technical capacity and financial resources, may benefit from partnering with technology companies across the AI journey. These partnerships could help offset costs that may occur with building in-house capacity, such as the cost of building a development team.

The following section outlines common decision points across three phases of the AI life cycle and practical considerations for mental health task-sharing programs.

Development phase

- **Build.** Programs with access to domain-specific data, a development team with technical expertise, and substantial financial resources may choose to build a custom LLM from scratch. This option offers maximum flexibility and control but is associated with the highest cost for the development phase, especially for data preparation and model training and testing. This path is generally not feasible for mental health task sharing programs without a dedicated development team with AI talent.
- **Partner.** Since most mental health task sharing programs do not have in-house development teams and have limited financial resources, they could benefit from partnering with a technology company to build their AI use cases. This partnership will likely include developing an AI solution that embeds a pretrained open-source model. However, task sharing programs could determine which model type—proprietary or open-source—is better suited to their goals and resources in collaboration with their in-house development team or an external technology partner.

Deployment phase

- **Build.** Programs with strong in-house technical experts, including software engineers and product managers, may opt to deploy and integrate AI solutions internally. This path allows for more customization and control over how AI solutions fit within existing workflows, but this approach requires careful planning for compliance, cybersecurity, and quality assurance, especially in health-related settings.
- **Partner.** For programs without in-house technical experts, partnering with a technology company can simplify deployment. External partners can support AI integration into existing platforms, manage technical infrastructure, and ensure usability across devices. This option allows for faster implementation and may reduce the risk of deployment delays.

Maintenance phase

- **Build.** Programs with in-house development teams may opt to manage the maintenance of deployed AI solutions internally. While this approach offers greater control and flexibility to tailor improvements based on user feedback, it requires sustained investment in technical talent, monitoring systems, and user support.
- **Partner.** For programs without in-house development, the team could partner with a technology company to outsource technical maintenance, including performance monitoring, improvements, and user support. These services are often delivered through a service contract and could offer predictable costs and timely upgrades.

For mental health task sharing programs with limited financial resources and technical expertise, a hybrid approach may be feasible. A technology partner could build the AI solution that embeds the AI model, while in-house staff could lead product management, implementation in the field, and user support. This approach could be ideal because it balances technical expertise with the contextual knowledge of the program and allows programs to gradually build internal capacity over time.

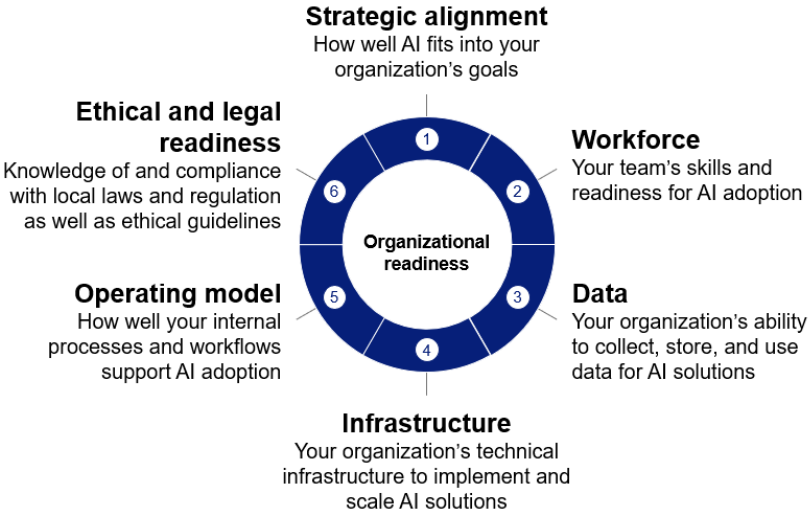
In this chapter, we explored the building blocks of using AI in a responsible and sustainable way—from making sure that solutions are technically sound and safe to setting up responsible governance and planning for long-term financial and operational sustainability. Up next, we share a simple readiness assessment to help you take stock of where your program is today and what steps might help you move forward in your AI journey.

Chapter 5: Assessment

Building on earlier chapters that discuss how AI can help mental health programs scale and the essentials of responsible AI adoption, this chapter focuses on the next step: assessing programs' readiness to adopt and implement AI solutions effectively.

To fully benefit from AI, task-sharing programs should first understand their current capabilities and the potential challenges they may face implementing AI, though some might be starting from scratch. This guide is meant to provide a blueprint for how to begin thinking about this. This chapter introduces a practical assessment framework to evaluate six dimensions that are critical to successful AI adoption (Exhibit 6). For each area, we provide a description, an example of what "good" looks like, and guiding questions to score your program's readiness.

Exhibit 6



The insights you gain from this assessment will help you plan the way forward in your AI journey. The next chapter contains additional resources that could support your journey.

Capability and readiness assessment: Are you ready for AI?

This assessment includes six dimensions with five questions per dimension, totaling 30 questions. Each question is scored on a scale from 0 to 4.

How to use this tool:

1. For each dimension, start by reading the description and the example that is considered "good" to understand what strong readiness looks like
2. Reflect on your program's current practices and capabilities
3. Answer each of the five questions using the rating scale below

4. At the end of the assessment, use both your overall score and the scores of each dimension to pinpoint specific areas of strength and areas requiring further improvement

   Rating scale:

   - 0: Not at all
   - 1: Partially, with substantial gaps
   - 2: Partially, with some gaps
   - 3: Mostly, with minor gaps
   - 4: Fully

Assessment:

1. Strategic alignment

Description: Strategic alignment refers to how clearly your program's AI ambitions are tied to its mission, priorities, and long-term goals for improving mental healthcare delivery and support. It also measures the degree of support and commitment from organizational leadership toward these AI initiatives—both of which are essential to ensuring that AI efforts are well-resourced and purposeful.

Example of "good": The program's strategic plan explicitly calls out the exploration or implementation of innovative technologies (including AI) as effective ways to expand the program's reach, efficiency, or quality of mental health programs. There is a written long-term vision for innovative technologies (including AI), it covers the next one to three years, and it is reviewed regularly to stay aligned with evolving strategic goals. Leadership believes in the importance of exploring and potentially adopting AI and communicates this vision clearly across teams. Financial, human, and technological resources are clearly allocated or planned for AI initiatives, with oversight by a designated leader or committee.

Questions:

For each question below, rate your program's current state from 0 to 4. Click on the icon in the upper right corner to see the rating scale.

1. Do you have an initial vision for how AI could evolve and be integrated into your services over the next one to three years?
2. Does your program have clearly defined goals for exploring or implementing AI that align with your mission and strategic priorities?
3. Does leadership explicitly communicate support for AI initiatives?
4. Is there a shared understanding among leadership and relevant departments of the potential opportunities and risks of using AI in your context?
5. Do your strategic plans include specific initiatives or resources (such as funding, personnel or time) that are required to evaluate or implement AI into your programs?

2. Workforce

Description: Workforce readiness refers to the extent to which your program's workforce possesses the necessary skills, knowledge, and adaptability to effectively engage with AI solutions while upholding ethical and client-centered practices.

Main elements of workforce readiness include the ones listed below. This assumes that mental health expertise is a crucial element that is already taken into account to assess readiness:

- **technical expertise** to support AI implementation and maintenance.
- **digital literacy** to use AI solutions and interpret and apply AI-driven insights.
- **change management experience** to adopt and integrate new tools effectively.

Example of "good": The program has a dedicated IT team with documented expertise in AI-related technologies (such as machine learning and natural language processing) or has a partnership with an external technology company. There are ongoing, structured training programs to enhance digital literacy among the staff, covering topics such as "interpreting AI outputs" and "recognizing potential bias". The program also has experience in managing technological transitions, supported by clearly defined change management strategies such as communication plans, role clarity, and feedback mechanisms.

Questions:

For each question below, rate your program's current state from 0 to 4. Click on the icon in the upper right corner to see the rating scale.

1. Do you currently have staff with expertise in areas relevant to AI, such as data analysis, software development, or machine learning (either in-house or through reliable partnerships)?
2. Are there designated individuals or teams responsible for exploring, implementing, and managing potential AI solutions?
3. Have you assessed the current digital literacy levels of staff who will interact with or interpret AI-driven insights?
4. Are there plans or resources allocated for training and upskilling staff on AI-related concepts and tools?
5. Does your program have experience managing technological change and ensuring user adoption of new systems?

3. Data

Description: Data readiness refers to your program's ability to collect, store, manage, and use data effectively for AI solutions. Having good data readiness means the program has a robust data infrastructure, has practices to ensure data quality, and adheres to security and privacy regulations. As covered in chapter 4, high-quality, well-governed data is critical for training, deploying, and monitoring AI solutions.

Example of "good": The program has robust, well-documented systems for securely collecting and storing relevant program data in compliance with ethical standards, regulations and applicable local privacy laws (such as HIPAA). Comprehensive data governance policies and operational processes are in place to routinely audit and ensure the accuracy, completeness, and secure handling of mental health data as well as privacy-preserving practices. The program can readily access and analyze data relevant to potential AI applications, supporting continuous improvements.

Questions:

For each question below, rate your program's current state from 0 to 4. Click on the icon in the upper right corner to see the rating scale.

1. Do you have established systems and processes for collecting relevant data related to your mental health programs?
2. Is your data stored securely and in compliance with relevant privacy regulations (such as HIPAA or GDPR)?
3. Do you have data governance policies that guide quality assurance, data security, and the ethical use of mental health data?
4. Do you have operational processes in place to regularly check the accuracy, completeness, and secure handling of your data?
5. Can you readily access and analyze the data that would be relevant for training or using AI models?

4. Infrastructure

Description: Infrastructure readiness refers to whether your program has the digital foundation needed to effectively deploy, integrate, and maintain AI solutions. Programs can assess their infrastructure readiness by evaluating network connectivity, hardware and software capabilities, and hosting and computing capabilities.

Example of "good": The program has high-speed, reliable internet connectivity across all operational areas. Current hardware and software systems are generally up-to-date and capable of handling modern applications. The IT infrastructure can support the integration of cloud-based AI services, and there are internal IT staff or external partners available to support the technical implementation and ongoing maintenance of AI systems. The existing systems are designed with scalability and flexibility in mind, using well-documented APIs to facilitate integration.

Questions:

For each question below, rate your program's current state from 0 to 4. Click on the icon in the upper right corner to see the rating scale.

1. Do you have reliable, high-speed internet connectivity across your operational areas?
2. Do your current hardware and software systems generally meet the requirements for running modern applications and handling data processing?
3. Do you have IT infrastructure that could potentially support the integration of cloud-based, on-premises (on-device), or hybrid AI solutions?
4. Do you have IT staff or external partners who could support the technical implementation and maintenance of AI systems?
5. Are your existing systems designed with scalability and flexibility in mind to accommodate new technologies?

5. Operating model

Description: Operating model readiness evaluates whether the program's internal processes, workflows, and culture are prepared to support the adoption of AI solutions. Programs with good operating model readiness have processes for evaluating new technologies, a culture of innovation, effective communication, and project management capabilities.

Example of "good": The program has documented and regularly reviewed processes for evaluating, piloting, and adopting new technologies. There is a culture that encourages innovation, learning, and responsible experimentation. Effective communication and collaboration exist across departments. Project

management practices are consistently applied to the implementation of new initiatives. Teams demonstrate flexibility in adapting new workflows to incorporate AI solutions.

Questions:

1. Does your program have established processes for evaluating and adopting new technologies?
2. Does your program promote a culture of innovation, learning, and responsible experimentation within your teams?
3. Do you have effective communication and collaboration channels across different departments that would be involved in AI initiatives?
4. Are there established project management systems in place to oversee the implementation of AI initiatives?
5. Is your program flexible in adapting to new workflows and processes that might be introduced by AI solutions?
6. Are end users or the community ready to accept the use of AI in the program?

6. Ethical and legal readiness

Description: Ethical and legal readiness evaluates your program's awareness of and ability to comply with relevant ethical standards and regulatory requirements related to the use of AI in mental health or healthcare. Programs with good ethical and legal readiness understand potential biases, ensure data privacy and security, and establish clear processes for risk mitigation and stakeholder feedback.

Example of "good": The program understands and monitors applicable international and local legal and regulatory requirements for AI use in mental health or healthcare. Staff across the program are aware of the ethical considerations related to AI, including—but not limited to—bias, privacy, transparency, and consent. Clear, up-to-date governance policies are in place to guide data privacy, security, and the responsible use of AI solutions. The program has structured processes for identifying, assessing, and mitigating potential ethical and legal risks associated with AI. Feedback mechanisms are also in place to gather and respond to stakeholder concerns and feedback on the use of AI in a timely, transparent manner.

Questions:

For each question below, rate your program's current state from 0 to 4. Click on the icon in the upper right corner to see the rating scale.

1. Does your program have a clear understanding of applicable international and local legal and regulatory requirements that may apply to the use of AI in your programs?
2. Are your staff aware of the ethical considerations related to AI, including—but not limited to—bias, privacy, transparency, and consent?
3. Do you have governance policies in place to guide data privacy, security, and the responsible use of AI?
4. Do you have processes in place to identify, assess, and mitigate potential ethical and legal risks associated with AI?
5. Do you have formal mechanisms to collect and address concerns or feedback from stakeholders regarding the use of AI in your programs?

Interpreting your readiness score

Below, you will find more information about this stage, highlighting common strengths and areas of development for programs in this stage to guide your next steps in your AI journey.

Scoring and interpretation:

Total score range	**Readiness stage**	**Description** (details below)
0–40	Stage 1—Awareness	These programs have limited AI readiness. They should focus on foundational learning and leadership engagement.
41–80	Stage 2—Exploration	These programs have initially explored AI and plan to integrate it. They should begin building capabilities and identifying opportunities.
81–100	Stage 3—Implementation	These programs have started to pilot AI initiatives. They should build infrastructure and integrate AI into workflows.
101–120	Stage 4—Optimization	These programs have mature AI adoption and have successfully implemented at least one or two use cases. They should focus on scaling, refining, and continuously improving solutions.

Stage 1—Awareness (0–40)

Programs in this stage are in the early phases of considering AI and most likely have the following characteristics:

- There may be limited awareness of AI capabilities and their potential applications within mental health programs.
- Strategic alignment is low, with little to no mention of AI in organizational plans.
- Data infrastructure, technical capacity, and talent readiness are likely to have substantial gaps.
- Ethical and legal considerations related to AI may not be well-understood.

Programs in this stage could focus on the following areas to build capabilities and move forward:

- **Initial education:** Build foundational knowledge of AI and potential use cases relevant to your mental health programs.
- **Use case prioritization:** Identify one or two potential areas where AI could support your program while considering technical; quality, safety, and trust; governance; and sustainability aspects needed to integrate AI solutions into your program.
- **High-level assessment:** Perform a basic review of the program's current data availability and technology infrastructure while considering the potential challenges related to working with mental health data.

- **Leadership alignment:** Initiate discussions with leadership to secure buy-in for further exploration, emphasizing the importance of governance considerations to ensure responsible AI use.

Consider revisiting the following chapters:

- Chapter 2: Background for foundational knowledge
- Chapter 3: Use cases to review case studies
- Chapter 4: Success factors for main dimensions and challenges to consider when integrating AI solutions

Stage 2—Exploration (41–80)

Programs in this stage are actively exploring the potential of AI and beginning to plan for its adoption. These programs most likely have the following characteristics:

- There is some level of strategic alignment, and there are some initial discussions about how AI could support organizational goals.
- Efforts to assess data infrastructure, technical capacity, and talent needs are underway.
- There may be some initial ethical and legal considerations.

Programs in this stage could focus on the following areas to build capabilities and move forward:

- **Detailed use case analysis:** Conduct an in-depth analysis of prioritized use cases considering model selection, data, and integration needs.
- **Technology assessment:** Perform a more thorough assessment of current technology infrastructure—such as hardware and software, hosting and computing capabilities, and network connectivity—evaluating its capacity to support AI implementation.
- **Talent gap analysis:** Assess the capabilities of the workforce and decide whether to build AI solutions in-house or to partner with a technology company. Results from the technology assessment could be used to decide which tasks to outsource in this step.
- **Preliminary budgeting:** Estimate the total cost of developing AI solutions in-house, or request a quote from your technology partner if you are outsourcing development.

Consider revisiting the following chapters:

- Chapter 3: Use cases to review case studies
- Chapter 4: Success factors for technology and sustainability considerations

Stage 3—Implementation (81–100)

Programs in this stage have moved beyond planning and are starting to implement pilot AI projects or early-stage applications. These programs most likely have the following characteristics:

- AI initiatives are more aligned with strategic objectives, and there is increasing leadership support.
- Investments are being made in data infrastructure, technical capacity, and talent development.
- There is a growing awareness of ethical and legal considerations, and some initial policies may be in place.

Programs in this stage could focus on ensuring they have the following areas built out:

- **Governance:** Set up an AI governance structure and establish policies around the responsible use of AI solutions.
- **Risk assessment:** Conduct a preliminary risk assessment to identify and mitigate potential ethical and operational risks that could occur with the prioritized use cases.
- **Infrastructure setup:** Set up necessary hardware, software, and cloud resources to support pilot projects.
- **Performance monitoring:** Implement systems to monitor AI performance, accuracy, and impact.
- **Pilot:** Develop and pilot one or two AI use cases with clear evaluation plans and risk mitigation strategies.
- **Training and onboarding:** Provide training and onboarding to users to teach them how to use and interact with the AI solutions.
- **Data preparation:** Clean, transform, and prepare data to improve AI models, adhering to the data governance policies of your program.

Consider revisiting the following chapter:

- Chapter 4: Success factors for technology and sustainability considerations

Stage 4—Optimization (101–120)

Programs in this stage have a mature approach to AI adoption and most likely have the following characteristics:

- The AI strategy is well-defined and closely aligned with organizational goals.
- Some AI solutions have been piloted and were found beneficial in their mental health programs.
- There is a strong focus on continuously improving, optimizing, and scaling AI solutions.
- Data infrastructure, technical capacity, and talent are well-developed and actively managed.
- Ethical and legal considerations are proactively addressed with a robust governance structure.

Programs in this stage could focus on the following areas to build capabilities and move forward:

- **Scale-up:** Expand successfully piloted solutions to new operational areas or start piloting different AI use cases.
- **Continuous improvement:** Regularly evaluate and refine AI models, operational processes, and governance structures to stay aligned with evolving organizational priorities and the technological landscape.
- **Thought leadership:** Publish your experience and learnings to guide other mental health programs that are in the early stages of AI adoption.

Completing a readiness assessment is a critical first step in your program's AI journey. By identifying the stage of AI development your program is currently at and the areas it has to develop further, you lay the foundation for more strategic, informed, and responsible AI adoption. The following section has a curated list of resources to support you as you continue to build capabilities and successfully and sustainably adopt AI into your mental health programs.

Chapter 6: Additional resources

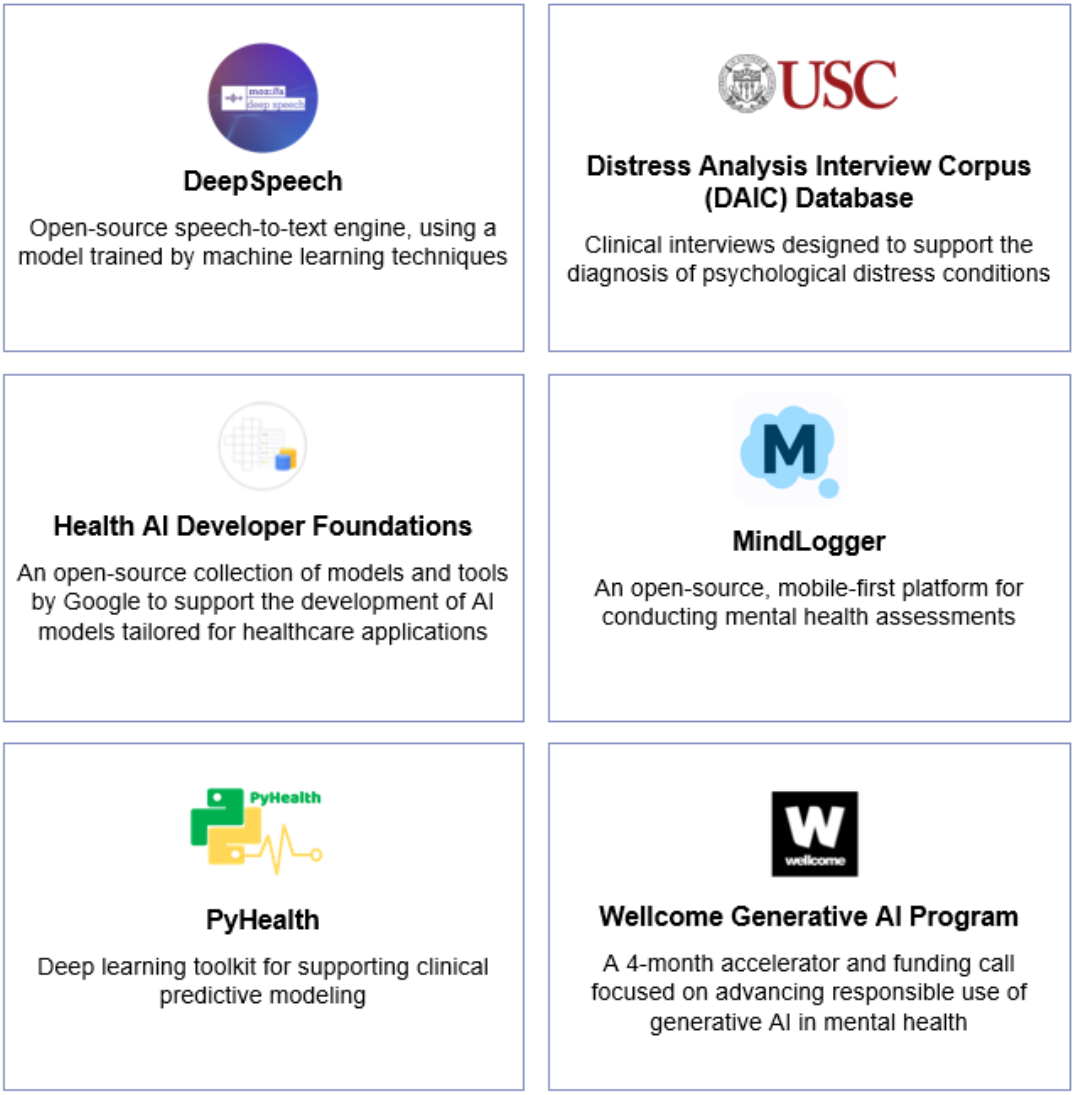
This is a preliminary list of practical tools and further readings that can help support your program's AI journey. These resources supplement this field guide and can deepen your understanding of responsible AI practices across technical and ethical domains.

The practical tools section features open-source models, toolkits, and funding opportunities that could help kickstart or accelerate the development and deployment of the AI use cases mentioned in the field guide.

These were selected based on each being open-access, clearly linked to health or medical applications, proved to be used in such applications, and shown to directly support or have the possibility to be linked to AI capabilities. This is a nonexhaustive sample based on publicly available information. The list is not an endorsement of these organizations or solutions and is for reference only. Organizations should make their own decisions on what tools meet their needs.

The further readings section offers timely guidance and thought leadership from global organizations on how to design, govern, and scale AI in a way that is human-centered, safe, and trustworthy.

Practical tools



Links to resources below

- DeepSpeech: https://github.com/mozilla/DeepSpeech
- DAIC Database: https://dcapswoz.ict.usc.edu/
- Health AI Developer Foundations: https://developers.google.com/health-ai-developer-foundations
- MindLogger: https://mindlogger.org/
- PyHealth: https://pyhealth.readthedocs.io/
- Wellcome Generative AI Program: https://wellcome.org/grant-funding/schemes/generative-ai-accelerator

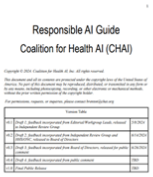
Further readings

Ethics and Governance of Artificial Intelligence for Health

The WHO guidance on ethical principles and governance structures for AI in health

Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Language Models

The WHO report outlining ethical considerations and governance guidance for large language models in health


Responsible AI Guide

Coalition for Health AI (CHAI)

Responsible AI Guide

A practical guide developed by Coalition for Health AI (CHAI) to promote responsible AI development and use in healthcare

Blueprint for Trustworthy AI

A blueprint developed by Coalition for Health AI (CHAI) to support trustworthy AI implementation and assurance in healthcare

The People + AI Guidebook

A practical guide by Google for designing human-centered AI products, with actionable insights across the AI life cycle

The Opportunities and Risks for Large Language Models in Mental Health

A peer-reviewed article by Google that summarizes how large language models are being used for mental health education, assessment, and intervention. It also highlights key opportunities and associated risks

Links for these readings are provided below:

- Ethics and Governance of Artificial Intelligence for Health: https://www.who.int/publications/i/item/9789240029200
- Ethics and Governance of Artificial Intelligence for Health - Guidance on Large Language Models: https://www.who.int/publications/i/item/9789240084759
- Responsible AI Guide: https://chai.org/wp-content/uploads/2025/03/CHAI_Responsible-AI-Guide_03202025.pdf
- Blueprint for Trustworthy AI: https://chai.org/wp-content/uploads/2024/05/blueprint-for-trustworthy-ai_V1.0-2.pdf
- The People + AI Guidebook: https://pair.withgoogle.com/guidebook/
- The Opportunities and Risks for Large Language Models in Mental Health: https://mental.jmir.org/2024/1/e59479/PDF

Closing thoughts

As mental healthcare continues to outpace the resources available, thoughtfully integrating AI into mental healthcare programs offers a powerful path forward—one that could expand access, enhance quality, and support the programs and people at the heart of support and care. We hope this field guide sparks new ideas, informs strategic decisions, and inspires collaboration across sectors.

Thank you for being part of the journey to increase access to mental health support and care for those in need—and to build more scalable, equitable, and human-centered systems of care.