Office of Artificial Intelligence Policy
UTAH DEPARTMENT OF COMMERCE

Division of Professional Licensing
UTAH DEPARTMENT OF COMMERCE

Guidance Letter

Best Practices for the Use of Artificial Intelligence by Mental Health Therapists

April 2025

Prepared for the Utah Office of Artificial Intelligence Policy and the Utah Division of Professional Licensing

Summary

Artificial intelligence (AI) technologies present significant opportunities to enhance mental health therapy while introducing novel ethical and practical challenges. This document establishes **guidelines for the responsible use of AI** by licensed mental health therapists within therapeutic, supervisory, and educational contexts.

Modern AI systems learn patterns from large training data sets to generate outputs (e.g., answers) for new inputs (e.g., questions). These systems can produce inaccurate or undesirable outputs. The likelihood of inaccurate or undesirable outputs can depend on the specific AI tool (e.g., some AI tools may learn wrong patterns) and the input (e.g., AI tools tend to be less reliable when either the input or the corresponding output is rare or unprecedented). We categorize AI technologies into predictive AI (i.e., systems that output specific labels or values) and generative AI (i.e., systems that produce flexible outputs like text, images, or conversational responses) and list themes to consider in weighing the advantages and disadvantages of their use.

The use of AI in mental health therapy offers several **potential benefits**. These include increased efficiency through automation of routine tasks, expanded capacity to process and analyze large amounts of patient data, and independent data-driven assessment to supplement therapist judgment. Patient-facing AI tools can provide personalized and engaging user experiences while improving the accessibility and availability of mental health support. However, **important risks** accompany these benefits. AI systems may produce inaccurate or biased outputs that lead to harmful decisions. They may respond inadequately to patients with diverse backgrounds and unique needs. Therapists may develop overconfidence in or overreliance on AI systems. Patient privacy and data security remain ongoing concerns. Technological barriers can create access inequities. Patients may experience over-immersion or overreliance on AI tools. Perhaps most fundamentally, AI interactions may lack the genuine emotional connection central to effective therapy.

The document examines various **applications of AI in mental health therapy**. These include AI-assisted risk assessment and diagnosis, monitoring of therapist performance, automated transcription and summarization of therapy sessions, treatment plan development, patient intake via conversational bots, bot-assisted therapy homework and learning experiences, and bot-administered talk therapy. Each use case presents unique benefits and challenges that require careful consideration.

Mental health therapists should adhere to several key guidelines for responsible AI implementation. Mental Health therapists must obtain **informed consent** from patients after clearly disclosing benefits, risks, and data practices. **Data privacy and safety** require verification of appropriate data handling standards and Business Associate Agreements with AI providers. **Therapist competence** demands ongoing education about AI capabilities, limitations, and proper use. **Patient safety** considerations include assessing digital literacy and risk factors for technology-related issues before implementing AI tools. Therapists should

carefully consider **unique patient needs** that may affect AI tool appropriateness and effectiveness. **Continuous monitoring** of AI outputs for accuracy and effectiveness should occur with frequency based on risk factors.

Therapists should maintain authorship transparency when interacting directly with AI systems by clearly disclosing when AI has contributed to documentation or communications. They should thoroughly **review AI-generated content** before including it in patient records or treatment decisions. Most importantly, therapists must evaluate **critically** and maintain professional judgment when considering AI-suggested diagnoses or treatments.

For patient-AI interactions, therapists should assess the **patient-technology relationship** before recommending AI tools. The **primacy of patient well-being** must be maintained, ensuring AI augments rather than diminishes the quality of care. **Contingency planning** prepares for potential service disruptions or malfunctions. Therapists must verify **ethical AI conduct** appropriate to system capabilities, understand **emergency response** limitations, and establish appropriate protocols. Regular **interaction monitoring** helps assess the effectiveness and safety of patient-AI exchanges.

Similar principles apply in educational contexts. **Student needs assessment** should precede AI incorporation in training. **Educational integrity** prioritizes effective learning outcomes over technological convenience. **A critical review** of AI-generated student assessments prevents overreliance. **Appropriate monitoring** of student-AI interactions ensures educational benefit.

These guidelines aim to help mental health therapists navigate the integration of AI technologies while upholding professional standards and prioritizing patient welfare. By implementing these best practices, therapists can harness AI's benefits while mitigating potential risks in therapeutic, supervisory, and educational settings.

Contents

1. Introduction

The rapid advance of artificial intelligence (AI) technologies, particularly generative AI technologies, has the potential to substantially transform the behavioral health field. These tools may offer promising opportunities to enhance behavioral healthcare (Thakkar et al., 2024; Ettman and Galea, 2023; Olewade et al., 2024). Examples include improved diagnostics, increased accessibility to mental health services, and enhanced patient-therapist interactions. These and other developments may improve the experiences of patients and therapists. Patients may enjoy easier access to high-quality care. Since AI technologies tend to excel specifically at automating menial and repetitive tasks, therapists may find it particularly beneficial to transfer such tasks to AI and focus their efforts on fostering the human connection with their patients and building robust therapeutic relationships.

However, new challenges accompany the new opportunities that emerge with AI technologies. Integrating AI technologies into behavioral health introduces ethical, legal, and practical challenges. Concerns about data privacy and security arise in using AI technologies, as these technologies often involve recording, storing, and sharing data, which, in the context of behavioral health services, can include sensitive patient information. Issues around informed consent, transparency, and accountability become increasingly complex in the context of AI-assisted behavioral health services. Additionally, when a mental health therapist or a patient becomes over-reliant on the use of these technologies, it can have detrimental effects on the provided service and the patient's well-being. Generative AI technologies, specifically, have the potential to generate unpredictable and, thus, potentially inaccurate responses in conversations with mental health therapists and potentially insensitive and/or harmful responses in discussions with patients. These and other risks of the use of AI technologies can create conflicts between the desire to enjoy the benefits of the new technology and a therapist's duty to align their practice with the ethics code of their profession (American Association for Marriage and Family Therapy, 2015; American Mental Health Counselors Association, 2020; American Psychological Association, 2016; American Psychiatric Association, 2013; Association of Social Work Boards, 2014; National Association for Alcoholism and Drug Abuse Counselors and National Certification Commission for Addiction Professionals, 2021; National Association of Social Workers, 2021; National Board for Certified Counselors, 2023).

On the one hand, the great potential for benefits motivates a thorough exploration of AI technologies and their various use cases in behavioral health. On the other hand, the substantial risks associated with using these technologies in behavioral health create a responsibility to prevent and mitigate harm that may arise from such use. Healthcare organizations, healthcare providers, and providers of AI tools all carry a part of this responsibility. How should a mental health therapist who wishes to use AI tools navigate the new technologies and the associated responsibilities? This document aims to establish clear guidelines on the ethical and responsible use of AI by licensed professionals in behavioral health services.

To this end, this document includes a primer on the current state of AI, a general risk-benefit comparison for different types of AI, a collection of use cases of AI technologies in mental health therapy, and a list of guidelines for using AI technologies by mental health therapists. These guidelines are meant to help mental health therapists of behavioral healthcare navigate the complexities of introducing AI technologies while safeguarding patient well-being and upholding professional standards.

2. Scope and Definitions

This guide is concerned with the use of generative AI technologies that can assist a mental health therapist in a licensed profession in offering services within the scope of their professional license to patients, supervision to colleagues in their profession, and education and training to existing licensees (i.e., continuing education), as well as students seeking licensure. It does not concern the use of AI tools outside a dedicated therapeutic relationship between a patient and their licensed mental health therapist. The use of conversational bots aimed at improving mental health and/or well-being falls only within the scope of this guide when a patient has an ongoing therapeutic relationship with a licensed mental health therapist who recommended or prescribed the use of the conversational bot to the patient.

For some concepts that appear frequently in this document, one can find different definitions in different sources. This report assumes the following definitions of the concepts, most of which are in accordance with the Utah State Code.

Artificial intelligence

“Artificial intelligence” means a machine-based system that makes predictions, recommendations, or decisions influencing real or virtual environments. (Utah State Code 13-72-101)

Artificial intelligence technologies

“Artificial intelligence technology” means a computer system, application, or other product that uses or incorporates one or more forms of artificial intelligence. (Utah State Code 13-72-101)

Generative artificial intelligence

“Generative artificial intelligence” means an artificial system that (i) is trained on data; (ii) interacts with a person using text, audio, or visual communication; and (iii) generates non-scripted outputs similar to outputs created by a human, with limited or no human oversight. (Utah State Code 13-2-12)

Mental health therapist

"Mental health therapist" means an individual who is practicing within the scope of practice defined in the individual's respective licensing act and is licensed as:

1. a physician and surgeon or osteopathic physician engaged in the practice of mental health therapy;
2. an advanced practice registered nurse specializing in psychiatric mental health nursing;
3. an advanced practice registered nurse intern specializing in psychiatric mental health nursing;
4. a psychologist qualified to engage in the practice of mental health therapy;
5. a certified psychology resident qualifying to engage in the practice of mental health therapy;

6. a physician assistant specializing in mental health care under Section 58-70a-501.1 of the Utah State Code;
7. a clinical social worker;
8. a certified social worker;
9. a marriage and family therapist;
10. an associate marriage and family therapist;
11. a clinical mental health counselor;
12. an associate clinical mental health counselor;
13. a master addiction counselor or
14. an associate master addiction counselor.

(Utah State Code 58-60-102)

Patient / Client

"Patient" means an individual who consults or, is examined, or is interviewed by an individual licensed under this chapter who is acting in the individual's professional capacity (Utah State Code 58-60-102).

Supervisor / Teacher

“Supervisor” or “teacher” means a licensed mental health therapist offering formal training and/or expert or clinical supervision to a student of a licensed behavioral health profession or to another licensed mental health therapist.

Student / Trainee

“Student” or “trainee” means an individual receiving formal training and/or expert or clinical supervision of their service in a licensed behavioral health profession by a licensed mental health therapist.

3. Background on Artificial Intelligence Technologies

Several times in human history, a new technology has transformed how we work, live, and participate in society. The invention of the combustion engine, for example, has led to the development of motorized vehicles and thus transformed how we travel and transport goods. The discovery of semiconductor technology has led to the development of digital calculators and, later, personal digital computers, and thus, has transformed how we calculate and store information. The rise of such transformative technologies tends to be followed by a period of uncertainty during which

- entrepreneurs test the scope and limitations of using the new technology to create new products and services and/or make the production and delivery of existing products and services more efficient via automation,

- small-business owners and service providers test how they can integrate these new products and services in their business operations and interactions with their clients in a beneficial way and

- educators and workers try to identify what knowledge one should acquire to participate effectively in the economy and society as it is changed by new technology.

The current period of uncertainty follows the development of new AI technologies, which were made possible by a collection of computing hardware and algorithm innovations. Debates on the level of technological knowledge needed to effectively and safely use AI technologies in one's professional and personal life are ongoing. The remainder of this section offers concise explanations of the foundations of modern AI technologies, types of AI technologies, and the data-privacy implications of using these technologies.

3.1 Foundations of Modern Artificial Intelligence

Most attempts to assess human intelligence in an educational setting or to use human intelligence in a professional setting are based on finding correct answers for some specific questions. For example, in an elementary school math exam, students are given sets of math problems and tasked to find a correct solution for each problem. A part of the intellectual work of a mental health therapist is to offer a considerate and contextually sensitive response to a patient's communications. Another part is to consider a patient record and formulate a correct or at least plausible diagnosis. In machine learning, the question, including the information on which the answer should be based (e.g., a patient record), is called an *input*. The answer (e.g., a diagnosis) is called an *output*. Modern machine-learning algorithms are based on showing a computer program many pairs of inputs and the corresponding desired outputs. The computer program can learn a *pattern* that connects the inputs shown with their corresponding outputs. A computer program that has learned a pattern in such a way is often called *artificial intelligence*

(AI). The set of input-output pairs shown to an AI to facilitate its learning of a pattern is called *training data*. Once the AI has learned the pattern, it can use the learned pattern to generate an output for an input that has not been shown before. This means that a user of the AI can prompt it with a question for which the AI has never seen the correct answer, and the AI will respond with a plausible answer based on its learned pattern.

3.2 Failure Modes for Modern Artificial Intelligence

This paradigm of training a machine or computer by showing it a training data set is the foundation of modern AI technologies. It is a powerful paradigm, but the trained AIs can generate wrong or undesired outputs in individual cases. These failures of AIs are typically associated with one of the following failure modes:

1. **The learned pattern includes errors.** The machine-learning paradigm described above makes it possible for correlational relationships among features of inputs and outputs to be learned as causal relationships. For example, a computer may be shown a set of patient records and corresponding diagnoses. The training data may contain many examples of women in the 20-40 years age range who have been diagnosed with postnatal depression. This may prompt a computer to learn that women in the 20-40 age range are likely to have postnatal depression, regardless of whether they have any children or have ever been pregnant. This error in the learned pattern may lead a computer to propose postnatal depression as a diagnosis for a childless 28-year-old woman who has no children and has never been pregnant.

2. **An input is outside the scope of an AI's intended use.** Many "AI-powered" products are designed for a specific and narrow use. For example, a developer may develop an AI assistant to summarize talk therapy sessions for mental health therapists. With this purpose in mind, the developer uses training data that includes transcripts of therapy sessions and corresponding high-quality summaries of these sessions. The developer then makes the AI assistant available to mental health therapists, one of whom discovers that it is possible to submit patient records instead of therapy transcripts as inputs to the AI assistant and receive output. While the output may look like a proper summary of the patient record, this summary may be missing critical information about the patient because the computer has never seen a high-quality summary of a patient record and thus does not know how to generate one.

3. **The correct output requires case-specific information that is not included in the input.** Sometimes, the input does not include case-specific information important for generating the correct output. This lack of information can lead AI tools to generate inaccurate outputs. Consider an AI tool designed to make a diagnostic recommendation based on transcripts of a patient's conversation with their therapist. The transcripts may fail to capture the patient's mannerisms, gestures, tone, and mood during the conversation. The therapist who has noticed these contextual features during the

conversation may thus develop a more accurate diagnosis than the AI tool that has no access to the contextual information.

4. **An input is rare, at least in the training data.** If some input scenarios are rare or missing in the training data, the computer may have seen few or no instances of this input during training. The learned pattern then does not apply to the rare input scenario, and consequently, the computer generates inaccurate outputs for these input scenarios. For example, an AI-powered chatbot designed to assist individuals with anxiety may have been primarily trained on conversations related to social anxiety and generalized anxiety disorder. Suppose a user presents with a rare anxiety subtype, such as health anxiety or obsessive-compulsive disorder (OCD). In that case, the chatbot might fail to provide appropriate support or misinterpret the user's concerns.

5. **An output is rare, at least in the training data.** A computer may have seen few or no examples of this output in the training data if a particular diagnosis, treatment plan, or intervention is uncommon. The learned pattern then does not connect inputs to these rare outputs, and it may be that the computer generates these rare outputs with a much lower frequency than desired or not at all. Patients with rare conditions or who would benefit from uncommon interventions may thus be disadvantaged when the output of an AI tool solely determines diagnoses or treatment plans.

6. **An input or output is under or overrepresented in the training data.** An AI tool may generate inaccurate outputs when some input-output pairs are under or overrepresented in its training data, even if all possible inputs and outputs are included in the training data. For example, suppose the training data for an AI-powered depression screening tool primarily consists of cases from urban populations. In that case, the AI's learned pattern may put too much focus on environmental stressors and not give enough consideration to the lack of access to resources as a risk factor for depression.

7. **Inputs include incorrect assertions.** Some conversational bots and other generative AI tools sometimes provide “hallucinated” outputs that include decidedly wrong information. Hallucinations tend to happen when generative AI tools are asked to provide information they cannot access or when the query includes wrong assertions. For example, when asking an AI to “summarize the patient's relationship with their sister” for a patient who does not have siblings, the AI may describe the patient's relationship with a made-up sibling.

3.3 Types of Artificial Intelligence

To understand the landscape of AI products, it can be helpful to distinguish between two types of AI based on the types of outputs that they generate: predictive AI and generative AI (Narayanan and Kapoor, 2024).

Predictive AI is an umbrella term for AI tools whose output is a prediction of a label for an input. Examples include

- diagnostic tools that recommend one of several possible diagnoses given a patient's data,
- sentiment analysis tools that label a patient's mood (e.g., a value out of the five words "happy," "sad," "angry," "neutral," or "anxious") based on a transcript of their conversation with their mental health therapist, and
- risk assessment tools that output a value (e.g., a value chosen out of the three labels "high," "medium," "low," or a number ranging from 0 to 1) to describe a person's risk of harming themselves or others.

Because these AI tools have substantial limitations on what outputs they generate, it is easy (or at least easier than in some other cases) to enumerate all possible failure scenarios and develop mitigation strategies for each scenario.

Generative AI is an umbrella term for AI tools that are less restricted in their outputs. Outputs can include free-form text, images, audio, or video. Examples include

- summarization tools that, when given a long text (e.g., a transcript of a talk therapy session) as input, generate a short summary of the input text,
- recommendation tools that generate recommended personalized treatment plans for a therapist based on a patient's data and
- chatbots or other conversational bots that engage in conversation with patients outside of therapy sessions with their therapist.

The flexibility of generative AI tools makes it impossible to account for every output they can generate and how that output may aid or harm therapists and patients. In this sense, the use of generative AI tools is more similar to using a car than predictive AI. There are simply too many ways in which accidents can happen to prepare for every single one of them individually. The skill of safely driving a car is obtained through many hours of driving experience and a driving instructor's guidance. Similarly, users of generative AI tools can learn to anticipate the likelihood of failure and quickly respond to failures of the AI tool appropriately through extensive experience with the AI tool in a controlled or assisted setting. Accordingly, our guidance on the safe use of AI technologies in mental health therapy includes sections specific to generative AI use.

Drawing a clear line between predictive AI and generative AI becomes difficult when predictive AI tools have many possible outputs or when generative AI tools are flexible enough to be used to predict labels. For example, it is in principle possible—although not recommended as it has the potential to violate HIPAA and related privacy laws—to share a patient record with a general-purpose language model (e.g., OpenAI's ChatGPT, Google's Gemini, Microsoft's

Copilot, or Anthropic's Claude) and ask it to generate a diagnosis, thereby effectively performing the task of a predictive AI. While the language model will likely comply, it is important to remember that these general-purpose AI tools are not designed for this use and that the quality and accuracy of diagnostic recommendations obtained in such a way can vary strongly. Some AI tools also combine a predictive AI for a specialized purpose with a generative AI (e.g., a language model), which lets users interact with the predictive AI via written or spoken language instead of requiring them to learn how to use a specific graphical user interface.

3.4 Implications for Data Privacy

The modern machine learning paradigm requires large amounts of data to continue improving the capabilities of AI tools. The demand for large data sets creates a strong incentive for providers of AI tools to collect data from their customers' interactions with their AI tools. Specifically, providers of AI tools for mental health services and mental-health-related services may be incentivized to collect and use patient data to improve their AI systems and sell these datasets to other entities for AI development purposes.

The practice of collecting and selling/sharing sensitive data is always associated with risks, such as

- data accidentally being made publicly available through human error,
- data being misused by one of the persons charged with data handling or
- data being stolen and misused or published by an external person.

In addition to these general risks, the collection and selling/sharing of sensitive data for training AIs bear further privacy risks because it may be possible for users to retrieve sensitive information about individual people whose data was included in the AI's training via *adversarial queries* (Nasr et al., 2023; Schwarzschild et al. 2025). Typically, these adversarial queries are inputs that users have meticulously and purposefully designed to trick AIs into revealing sensitive data. In rare cases, however, it is also possible that the benign use of an AI leads to the disclosure of sensitive data.

The likelihood of adversarial queries to an AI successfully leading to the disclosure of sensitive data tends to be low. Still, it changes frequently as AI developers improve AI security and hackers strengthen their capabilities. This complex landscape of privacy risks makes it more important than ever for mental health therapists to

- carefully review service agreements, business associate agreements, and other agreements that govern the collection and use of patient data and

- obtain a patient's informed consent before using AI tools that record or transcribe therapy sessions or involve interactions between a patient and a conversational bot beyond patient-intake procedures.

4. Potential Benefits and Risks of the Use of Artificial Intelligence Technologies in Mental Health Therapy

Emerging AI technologies developed and/or used in mental health therapy can be associated with a host of potential benefits and risks. While the concrete, complete risk-benefit profiles of each AI tool will depend on the specific tool, use case, and context, it is possible to identify general themes with which the potential benefits and risks of many AI tools in mental health therapy align.

4.1 Potential Benefits and Risks of the Use of Predictive Artificial Intelligence Technologies

Use cases for predictive AI in mental health therapy involve analyzing data to help therapists make decisions (Lee, 2021). Accordingly, **potential benefits** tend to fall into one of three categories:

- **Increased efficiency through automation.** The AI-assisted automation of decision-making tasks can help therapists and therapist supervisors to work more efficiently. For therapist and therapist supervisors working long hours, saving time on tasks that can be made more efficient through the use of AI technologies can improve their work-life balance and personal health, and it can improve the quality of the therapeutic, educational, and supervisory services that they offer to their patients, students, and trainees. The increased efficiency can also allow therapists to serve an increased number of patients and therapist supervisors to serve an increased number of students and trainees.

- **Increased scope or capacity through automation.** AIs can process large amounts of data quickly, thereby allowing therapists and therapist supervisors to base their assessments, recommendations, and decisions on holistic analyses that would be impractical to do otherwise. This increase in capacity has the potential to lead to more accurate diagnoses, a higher quality of care, and, in some cases, greater transparency in the therapist's decision-making process.

- **Independent data-driven assessment.** Therapists can compare their assessments of patient conditions, therapy sessions, etc., to those provided by an AI. This data-driven "second opinion" can provide a therapist with additional insights that can improve decision-making. Similarly, an independent data-driven perspective can potentially aid therapist supervisors in assessing their trainees' performance and progress more holistically and/or more accurately.

Risks tend to fall into one of six categories:

- **Risk of inaccurate or wrong AI outputs.** Many commercially available AI tools advertise their high accuracy and low failure rates. Nonetheless, it is almost always possible for AIs to generate inaccurate or wrong outputs, for example, via one of the failure modes listed under *Failure Modes for Modern Artificial Intelligence*.

- **Risk of inadequate response to patient backgrounds and unique needs.** Several examples of the use of predictive AI on large populations of people have highlighted that, in such cases, subpopulations can be negatively affected. Some AI tools provide accurate predictions and good recommendations on average but consistently lead to undesirable outcomes for patients with specific backgrounds. These systematic deviations occur when some patient backgrounds or unique needs are not adequately represented in the AI's training data (see *Failure Modes for Modern Artificial Intelligence*). The resulting harms can be hard or even impossible to detect by individual patients or therapists because they may only become evident when comparing several large patient groups with different backgrounds and/or needs. Undetected and unmitigated, these inadequate responses to patient backgrounds and unique needs can lead to large groups of patients receiving inadequate care.

- **Risk of human overconfidence in AI outputs.** Some of the potential harm from wrong or implausible AI outputs can be prevented or mitigated through appropriate human oversight. When therapists or therapist supervisors become overconfident in their AI tools, however, they may decide

  - not to review an AI's output or
  - to uncritically trust an AI's judgment over their own, even when they notice discrepancies.

  Therapists who neglect their responsibility to review AI outputs critically may fail to prevent harm to patients. In educational settings, inaccurate AI outputs may harm a trainee's education if their supervisor does not detect and correct them.

- **Risk of human overreliance on AI assistance.** A therapist who offloads many of their analysis and decision-making tasks on AI tools can become so reliant on these tools that they feel unequipped to perform these tasks without AI tools. Being able to analyze patient data and make therapy-related decisions independently ensures that therapists can

  - critically review AI outputs,
  - detect mistakes made by AIs, and

    - produce accurate diagnostic analyses, well-supported therapeutic decisions, and treatment plans when an AI tool fails or becomes unavailable.

  AI tools may become unavailable for various reasons, including but not limited to computer viruses, internet or power outages, or product discontinuation by or bankruptcy of the providing company.

- **Privacy risks.** The use of AI tools on patient or student data requires this data to be collected and motivates its digital storage. The collection and storage of such data are associated with the risk of breaches of confidentiality through accidental or purposeful data mishandling. If the collected data is used to train AIs, an additional privacy risk arises because AIs can leak training data in response to adversarial queries (see *Implications for Data Privacy*).

4.2 Additional Potential Benefits and Risks of the Use of Generative Artificial Intelligence Technologies

Generative AI has facilitated the development of capable and flexible conversational bots. These conversational bots include chatbots, audiobots, and “digital humans” (i.e., animated videobots). The various potential use cases of such conversational bots can involve therapists, patients, therapist supervisors, and therapist trainees as users. For these user groups, using such conversational bots tends to be associated with the potential benefits and risks specific to human-machine interactions in which a machine mimics human responses. **Potential benefits** of such interactions tend to fall into one of five categories:

- **Ease of use.** When querying or adding data to a patient record or other data collection would otherwise involve a user interface with various dropdown menus, checkboxes, and so on, sharing the data request or the new data with a conversational bot can make the process easier or more pleasant for the user.

- **Engaging user experiences.** Due to the immersive nature of interactions with conversational bots, users who feel motivated or energized by interactions with others may be more engaged in tasks that involve conversational bots than in tasks that do not involve any conversational element.

- **Personalized service.** Conversational bots may have access to detailed information about their users through memory of their conversations or other means. A conversational bot developed to customize its service based on such data may use this information to create personalized user experiences that match a user’s individual preferences or needs.

- **Increased sensitivity to a user’s background and unique needs.** If developed accordingly, a conversational bot can adjust its responses to consider a user’s unique needs that may arise from their economic, racial, ethnic, cultural, or other backgrounds. In a therapeutic setting, a conversational bot sensitive to its patient's background can help patients feel validated and understood, especially when it is difficult for their therapists to relate to the patient's personal experiences. In an educational setting, a conversational bot sensitive to a student's background can potentially provide more effective learning experiences.

- **Increased availability of support.** A conversational bot whose purpose is to support a patient in maintaining or achieving mental health and well-being can, at least in principle, be available anywhere and at all times via a smartphone application or similar device or service. This creates an opportunity for improved care for patients who would benefit from immediate support on demand and/or continuous 24-hour support. A conversational bot developed for educational purposes can support students in their learning whenever the students decide to devote time to their studies.

Risks tend to fall into one of six categories:

- **Technological barriers for users.** Access to technology, including reliable internet connectivity and suitable devices (e.g., smartphones or computers), is not evenly distributed across all populations. This can create a barrier for individuals from low-income communities, rural areas, and certain marginalized groups, potentially exacerbating existing inequalities in access to mental healthcare. Additionally, positive and successful user experiences may require users to have
    - a certain level of trust in a technology's capability and alignment and
    - a certain level of technical proficiency.

  Users with a low level of trust in the technology or technical proficiency may face challenges in effectively interacting with these tools. They may feel intimidated or frustrated by the experience. This tech-induced frustration can make therapy for these patients less effective. For therapists in training, tech-induced frustration can lead to poor learning outcomes.

- **User over-immersion**. Some conversational bots can mimic human behavior in very immersive ways. A user of such a bot may form an emotional connection with the bot and may prefer excessive interactions over healthy human social connections, negatively affecting their emotional well-being (Fang et al., 2025). In such situations, users may
    - neglect meaningful real-life relationships,
    - fail to develop healthy coping mechanisms for challenges that they experience in social interactions,
    - suffer a decline in essential social skills and a diminished capacity for genuine human connection,

  - develop a diminished sense of agency and autonomy through overreliance on the conversational bot's guidance and support,
  - experience anxiety and other withdrawal symptoms when access to the AI tool is limited.

- **Lack of emotional connection.** While AI tools can mimic human language and emotional expressions, they cannot fully replicate a human therapist's nuanced emotional intelligence and empathy. This can lead to a lack of genuine emotional connection and understanding, potentially hindering the therapeutic process. An AI tool may struggle to accurately interpret and respond to subtle emotional cues, such as tone of voice, body language, and micro-expressions, potentially leading to a user feeling isolated and misunderstood.

- **Lack of sensitivity to a patient's background and unique needs.** AI tools may not be adequately equipped to understand and respond appropriately to the unique cultural nuances and values of diverse patient populations. This can lead to culturally insensitive or inappropriate responses from conversational bots, potentially alienating and offending users from different cultural backgrounds. Specifically, if the datasets used to train AI models lack diversity (see *Failure Modes for Modern Artificial Intelligence*), the AI may not be able to effectively address the specific mental health needs of patient populations that are underrepresented or not represented in those data sets.

- **Undesired consequences from out-of-scope use.** AI tools can be misused or abused, potentially leading to harmful consequences. Over-reliance on generative AI tools could potentially erode professional standards in mental healthcare. This could lead to a decline in the quality of care therapists provide and a devaluing of the essential human element in the therapeutic process.

- **Declining engagement over time.** Digital mental health interventions, including AI-powered tools, often face challenges with user retention and sustained engagement. Initial enthusiasm may quickly wane as users experience intervention fatigue, leading to decreasing usage patterns and eventual abandonment. This dropout phenomenon can result in
  - incomplete therapeutic processes that fail to deliver meaningful clinical outcomes,
  - false impressions of treatment efficacy when only engaged users are measured in evaluations,
  - frustration and self-blame among users who interpret their disengagement as personal failure rather than a design limitation,
  - resource misallocation when healthcare systems invest in technologies that remain underutilized beyond initial adoption phases.

The challenge is particularly pronounced for users with certain mental health conditions that inherently impact motivation and follow-through, potentially creating a paradoxical situation where those who might benefit most struggle the most to maintain consistent engagement with digital interventions.

- **Hallucinations and harmful advice.** Some conversational bots and other generative AI tools sometimes provide “hallucinated” outputs that include decidedly wrong information (see *Failure Modes for Modern Artificial Intelligence*). If unchecked, these hallucinated outputs may lead therapists to adopt ineffective or even dangerous treatment plans or make decisions based on false or misleading information. Generative AI tools interacting with patients may directly provide harmful advice to a patient, instill or strengthen harmful beliefs and opinions that a patient holds, or encourage patients to act in ways that are harmful to themselves or others.

Together with the risk categories listed under *Potential Benefits and Risks of the Use of Predictive Artificial Intelligence Technologies*, the risk categories listed here outline the risk landscape for the use of generative AI tools. A comparison with the list of categories of potential benefits of generative AI tools suggests that many categories of potential benefits and risks can be paired as two sides of the same coin. For example:

- Conversational bots create **new ways to interact with information technology**. For some, these new ways may improve the ease of use. For others, these new ways to interact with technology can cause tech-induced frustration.

- **Immersive interactions with conversational bots** can have positive effects on users. However, users who become over-immersed may experience various adverse effects.

- Generative AI tools can create **personalized user experiences** using personal user data. These personalized experiences can make patients feel understood and validated. However, patients can feel misunderstood or offended when the user experience is not adequately personalized. For these patients, AI-augmented therapy may be less effective than a traditional therapeutic approach that focuses on building an emotional connection with a therapist.

- Generative AI tools may be equipped to converse appropriately (sometimes even more so than a therapist) with **patients of various economic, racial, ethnic, cultural, or other backgrounds** if the diverse challenges associated with these backgrounds were addressed during the AI’s development. If they were not addressed during the AI’s development, the AI tool may be particularly insensitive to a patient’s background and unique needs and thereby risk alienating or offending the patient.

Assessing whether potential benefits outweigh the risks when introducing a generative AI tool in therapy can be difficult. The outcome of such a risk-benefit analysis can depend on how a specific AI tool was developed, a patient’s background, relationship to technology, mental health conditions, and how the therapist uses (or instructs the patient to use) the AI tool to augment the patient’s therapy.

5. Use Cases for Artificial Intelligence Technologies in Mental Health Therapy

This section includes a non-exhaustive list of use cases for AI in mental health therapy and their risk and benefit profiles. Many AI products currently advertised for this use tend to address one or several of these use cases.

5.1 Use Cases for Predictive Artificial Intelligence Technologies

AI-assisted assessment of the risk of harm to self or others

AI tools can analyze patient data, such as past medical records, therapy notes, and - with appropriate consent - even social media activity to predict the likelihood of self-harm or harm to others.

Potential benefits

- **Increased efficiency through automation.** Automating parts of the risk assessment process can reduce the time it takes a therapist to conclude a risk assessment. Identifying patients with an elevated risk of self-harm or violence more quickly can create a greater window of opportunity for timely interventions that can potentially prevent tragic outcomes.
- **Increased scope or capacity through automation.** Time saved through automating part of the risk assessment process can free a therapist to focus more on other complex aspects of patient care.
- **Independent data-driven assessment.** By analyzing vast amounts of data, AI can potentially improve the accuracy of risk assessments compared to traditional methods, which may be subjective and prone to human bias.

Risks

- **Risk of inaccurate AI outputs.** AI tools may generate inaccurate predictions, leading to unnecessary interventions (i.e., false-positive predictions) or missed opportunities for intervention (i.e., false-negative predictions). Inaccurate assessment can occur, for example, when conversations include language nuances such as metaphors, idioms, irony, or sarcasm, which may be challenging for an AI to interpret correctly. Another risk factor is situations where the correct interpretation of a conversation requires important context (e.g., a patient’s cultural or religious background) that may affect how they express themselves and what cultural norms give the context for their relationships with others. These and other factors can lead to inaccurate risk assessments, and some of these inaccurate assessments can have detrimental and potentially even fatal consequences.

- **Risk of human overconfidence in AI assistance.** A therapist's expert knowledge and personal understanding of a patient's experiences can be necessary for an accurate risk assessment. When a therapist uncritically trusts an AI's risk assessment over their own, these essential factors are not included in the risk assessment. The resulting risk assessment may be less accurate and ultimately detrimental to the patient and the people around them.

- **Risk of human overreliance on AI assistance.** Like any other digital tool, an AI-driven risk-assessment tool can become unavailable anytime for various reasons (see *Potential Benefits and Risks of the Use of Predictive Artificial Intelligence Technologies*). **Individual patients may also** choose to opt out of AI tool usage in their therapy. In such cases, therapists must realize all aspects of their responsibilities, including assessing a patient's risk of self-harm and harm to others without the AI tool.

- **Risks to patient privacy.** These AI technologies can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on a patient's data may reveal that data in response to adversarial queries. Risk-assessment tools used by mental health therapists may collect past medical records, therapy notes, and other personal data for their assessment. Using an AI-driven risk-assessment tool thus bears the risk of a severe breach of confidentiality.

AI-assisted diagnosis

AI-assisted diagnostic tools can analyze patient data to aid therapists in a clinical setting in identifying mental health conditions.

Potential benefits

- **Increased efficiency through automation.** A therapist's diagnosis of a patient's mental health conditions may depend on the therapist's impressions gathered from interactions with their patients and various patient data, such as intake forms, medical records, and referral notes from other therapists. Reviewing these and other patient data can be partially or fully automated using AI tools. Automating the data review can increase efficiency in a therapist's diagnostic work.

- **Increased scope or capacity through automation.** Automating diagnostic tasks can allow therapists to diagnose a larger number of patients. Additionally, therapists may consider a larger amount of patient information in each diagnosis because automated data review allows therapists to analyze large amounts of data quickly.

- **Independent data-driven assessment.** A therapist or an AI (or both) can arrive at inaccurate diagnoses for various reasons. A discrepancy between the therapist's diagnosis and a diagnosis obtained from an AI's review of the patient's data indicates that at least one of these diagnoses is incomplete or inaccurate. The discrepancy thus motivates a critical review of both diagnoses. Using the output of a diagnostic AI tool to contrast a therapist's diagnoses of their patient's mental health can increase the accuracy of the therapist's diagnoses.

Risks

- **Risk of inaccurate AI outputs.** Diagnoses provided by AI tools can be inaccurate. The risk of an inaccurate diagnosis can be elevated when
    - the accurate diagnosis would include rare conditions or conditions that may have been underrepresented during an AI tool's training and development or
    - demographic, cultural, religious, or other aspects of a patient's background may affect how a condition presents itself and may have been underrepresented during an AI tool's training and development.
- **Risk of human overconfidence in AI assistance.** A therapist's expert knowledge and personal understanding of a patient's experiences can be necessary for arriving at an accurate and complete diagnosis of a patient's mental health conditions. When a therapist uncritically trusts an AI's diagnosis over their own, these essential factors are not included in the diagnosis, which may thus be less accurate and ultimately detrimental to the patient.
- **Risk of human overreliance on AI assistance.** Like any other digital tool, an AI-driven diagnostic tool can become unavailable anytime for various reasons (see *Potential Benefits and Risks of the Use of Predictive Artificial Intelligence Technologies*). **Individual patients may also** choose to opt against the use of such an AI tool in their therapy. In such cases, it is important that therapists can realize all aspects of their responsibilities, including their diagnostic work, without the AI tool.
- **Risks to patient privacy.** These AI technologies can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on a patient's data may reveal that data in response to adversarial queries. Diagnostic AI tools used by mental health therapists may collect past medical records, therapy notes, and other personal data for their assessment. Using an AI-driven diagnostic tool thus bears the risk of a severe breach of confidentiality.

AI-assisted monitoring of therapist performance

AI tools can evaluate therapist's performance based on transcriptions or recordings of their interactions with patients. Tools can calculate performance scores according to performance metrics set by a therapist for self-monitoring or a supervisor for training.

Potential benefits

- **Increased efficiency through automation.** AI-driven performance-monitoring tools may help supervisors provide performance feedback to trainees more efficiently than before by automating several aspects of the review and analysis of the trainees' therapist-patient interactions. In self-monitoring and training settings, AI-driven performance monitoring can improve treatment fidelity by evaluating how faithfully or accurately a therapist implements a treatment plan.

- **Increased scope or capacity through automation.** The efficiency gained through automating some aspects of therapist supervision may allow a supervisor to supervise a larger number of trainees simultaneously. It may also allow trainees to receive more detailed feedback on their performance than from a supervisor who has to divide their time across several trainees and other job responsibilities.

- **Independent data-driven assessment.** An AI-driven performance monitoring tool can provide a data-driven performance assessment that is not subject to the same biases that might affect assessments from human supervisors, which might, in some cases, rely too heavily on their judgment of a trainee's character rather than their most recent performance. In addition to efficiency gains, if a supervisor selects one or several performance measures to be used over an extended supervision period, AI-calculated performance measures can improve the consistency of their trainees' performance scores. If the supervisor also shares this decision and the selected performance measures with their trainees, trainees can benefit from a more transparent assessment process with clear avenues for self-improvement.

Risks

- **Risk of inaccurate AI outputs.** AI tools can produce inaccurate performance scores. Such inaccurate scores can result from the AI tool misinterpreting aspects of the therapist-patient interaction and/or the therapist's approach to treatment. These misinterpretations are particularly common when the therapist-patient interactions include language nuances such as metaphors, idioms, irony, or sarcasm, which may be challenging for an AI to interpret correctly. Another risk factor is situations where the correct interpretation of a conversation requires important context (e.g., a patient's cultural or religious background) that may affect how they express themselves and react to their therapist's suggestions.

- **Risk of human overconfidence in AI assistance.** A supervisor or trainee who is overconfident in the AI tool may decide to trust an AI's assessment of a trainee's performance over their assessment, even when these two assessments conflict. This may cause inaccurate performance assessments to detrimentally influence trainees' self-perception, learning strategy, and professional skill development.

- **Risk of human overreliance on AI assistance.** Like any other digital tool, an AI-driven performance assessment tool can become unavailable anytime for various reasons (see *Potential Benefits and Risks of the Use of Predictive Artificial Intelligence Technologies*). It is also possible that individual trainees or their patients may choose to opt against the use of such an AI tool in their therapy. In such cases, a supervisor needs to be prepared to fulfill their supervisory responsibilities in any such case.

- **Risks to patient privacy.** These AI technologies can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on trainees' and patients' data may reveal data in response to adversarial queries. Performance-assessment tools used by mental health therapists or their supervisors may collect
    - copies of text interactions between therapists and their patients,
    - transcripts, or audio or video recordings of person-to-person therapist-patient interactions,
    - and other sensitive data.

  Using an AI-driven performance-assessment tool thus bears the risk of exposing sensitive trainee or patient data.

5.2 Use Cases for Generative Artificial Intelligence Technologies

Automated transcription of therapist-patient interactions

Speech-to-text technology enables therapists to transcribe spoken words during therapist-patient interactions (e.g., talk therapy sessions) into written text automatically.

Potential benefits

- **Increased efficiency through automation.** Automated transcription alleviates the need to take manual notes during a therapy session. The transcripts or AI-generated summaries can help therapists efficiently write reports on patients' status and progress.

- **Increased scope or capacity through automation.** Reducing the labor associated with note-taking during therapy sessions frees therapists to fully focus on the interaction with their patients while the interactions occur.

- **Independent data-driven assessment.** Automated transcripts may lead to more comprehensive and accurate patient records. Therapists can use the computerized transcripts to review the exact dialogue of a therapy session when designing a personalized treatment plan.

Risks

- **Risk of inaccurate AI outputs.** AI-generated transcripts can include transcription errors that can misrepresent the conversation. Such inaccuracies can arise, for example, due to algorithms with better speech recognition for certain accents or forms of expression. The inaccurate information in transcripts can harm patients and therapists in various ways. For example, inaccurate transcripts can lead to inaccurate patient records and, in turn, to patients receiving inaccurate diagnoses and inefficient treatment plans from their therapist and/or other healthcare providers. Additionally, if transcripts are subpoenaed or divulged via discovery, courts may assume that the information in these transcripts is accurate and allow them as evidence in legal disputes. In extreme cases, AI hallucinations can lead to completely fabricated and unsubstantiated details being included in a patient's record, which may be used to convict a patient of a crime that they have not committed.

- **Risk of human overconfidence in AI assistance.** Overconfidence in AI-generated transcripts may cause therapists to include transcription errors in a patient's permanent records, thereby failing to mitigate the harms that can arise from inaccurate AI outputs (see above).

- **Risk of human overreliance on AI assistance.** A therapist who becomes over-reliant on AI-generated transcripts may refrain from taking personal notes and consulting their memory of therapy sessions. The practice of memorizing the contents of a therapy session and revisiting those memories after the session is necessary to critically review AI-generated transcripts and identify and correct inaccuracies in such transcripts.

- **Risks to patient privacy.** These AI technologies can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on a patient's data may reveal that data in response to adversarial queries. The existence of such transcripts also opens up the possibility for them to be subpoenaed or divulged via discovery. (For example, therapists are mandated reporters for certain things divulged in therapy, such as child abuse. A victim could subpoena the transcript as evidence that the therapist had information that they should have reported but did not.)

Automated summarization of therapist-patient interactions

Language models can assist therapists by generating summaries of therapist-patient interactions based on therapy session transcripts, therapist notes, and/or patients' electronic records.

Potential benefits

- **Increased efficiency through automation.** Automated summarization can streamline documentation practices for therapists. Improving the efficiency of writing reports on patients' status and progress can, for example, decrease therapists' workload, potentially improving a therapist's work-life balance. Patients may also benefit from this efficiency gain because an energized and well-rested therapist may be able to offer an improved quality of care.

- **Increased scope or capacity through automation.** The AI-generated summaries can offer a comprehensive overview of therapy sessions, potentially exceeding the level of detail that one would reasonably expect from a therapist's manually taken notes. They may thus improve a therapist's ability to identify key themes and issues in their patient's therapy.

- **Independent data-driven assessment.** The AI-generated summaries can offer a data-driven alternative to a therapist's evaluation of a patient's condition and progress. Therapists can use AI-generated summaries to review and potentially refine their assessments critically.

Risks

- **Risk of inaccurate AI outputs.** AI-generated summaries can include hallucinations, misinterpretations, and other errors that can misrepresent a patient's electronic records. Such errors can arise, for example, when the correct interpretation of a patient's condition would require important context like their facial expression, tone of voice, and gestures that they made during their interactions with their therapist. That information is not included in the AI's inputs. An AI-driven summarization tool may emphasize long and detailed record entries more than short and concise entries (or vice versa). Such biases can lead to a skewed AI-generated patient electronic record summary. They are particularly prone to arise if patient records

    - include entries from multiple therapists and/or other healthcare providers with different documentation styles
    - include some entries that a therapist generated with AI assistance and some entries that the therapist created with assistance from AI tools.

  The inaccurate information in AI-generated summaries can harm patients and therapists. For example, they can lead to patients receiving inaccurate diagnoses and inefficient treatment plans from their therapist and/or other healthcare providers. Additionally, suppose a patient’s electronic records are subpoenaed or divulged via discovery. In that case, courts may assume that the information in these records is accurate and allow them as evidence in legal disputes. In extreme cases, AI hallucinations can lead to completely fabricated and unsubstantiated details being included in a patient’s record, which may be used as evidence in a legal dispute.

- **Risk of human overconfidence in AI assistance.** Overconfidence in AI-generated summaries may cause therapists to include inaccurate information in a patient’s permanent records, thereby failing to mitigate the harms that can arise from inaccurate AI outputs (see above).

- **Risk of human overreliance on AI assistance.** Over-reliance on AI-generated summaries may cause therapists to refrain from documenting their observations in their patients’ records. Their observations can be different and sometimes more accurate than the information in an AI-generated summary. This can be the case, for example, when vital context on the patient’s background is missing from the information provided to the AI tool. In such a case, the therapist must review and amend the AI-generated summary as necessary.

- **Risks to patient privacy.** These AI technologies can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on a patient’s data may reveal that data in response to adversarial queries. Moreover, AI-generated summaries included in a therapist’s notes or a patient’s electronic records can be subpoenaed or divulged via discovery. (For example, therapists are mandated reporters for certain things divulged in therapy, such as child abuse. A victim could subpoena the therapist’s notes as evidence that the therapist had information that they should have reported but did not.)

Automated summarization of multiple patient data modalities

Language models can be used to consolidate a patient's history from various records, which may include different types of data.

Potential benefits

- **Increased efficiency through automation.** Automated data summarization of patient data can substantially improve a therapist’s workflow for generating comprehensive overviews of a patient’s condition and progress.

- **Increased scope or capacity through automation.** AI-generated data summaries can facilitate the summarization of data types that a therapist cannot process and summarize on their own. For example, AI-generated summaries could incorporate information from many transcripts of therapy sessions, therapy notes, user activity data from a patient's use of phone apps that offer mental health support, etc.

- **Independent data-driven assessment.** Summaries considering large and diverse data sources may provide a cohesive overview and holistic, data-driven understanding of a patient's condition. Such a summary may uncover patterns or essential details that a therapist may have missed.

Risks

Compare risks of *Automated summarization of therapist-patient interactions*.

AI-assisted design of treatment plans

AI tools can provide treatment recommendations and generate treatment plans based on patient data and established clinical guidelines.

Potential benefits

- **Increased efficiency through automation.** The AI-assisted design of treatment plans can substantially improve a therapist's workflow for developing treatment plans for their patients.

- **Increased scope or capacity through automation.** If AI tools are developed to consider an extensive catalog of professional sources in creating treatment recommendations and/or treatment plans, such tools can enable therapists to base their treatment plans on a wide range of evidence-based practices. If these tools are also frequently updated by their developers to incorporate state-of-the-art knowledge, the AI-generated treatment recommendations may include new therapeutic approaches that a therapist has not yet been able to include in their continued education.

- **Increased sensitivity to a user's background and unique needs.** If developed accordingly, an AI tool can adjust its treatment recommendations to consider a patient's unique needs that may arise from their economic, racial, ethnic, cultural, or other backgrounds. For example, the AI tool might recommend treatment approaches that are demonstrated to be most effective with elderly populations if the patient is a senior citizen. Such tailored interventions have the potential to improve the efficacy of therapy and thereby improve patient outcomes.

Risks

- **Risk of inaccurate AI outputs.** AI can generate unhelpful treatment recommendations based on inaccurate data in a patient's record or a misinterpretation of a patient's unique needs. AI hallucinations may cause an AI to assume a patient's record includes

- **Risk of human overconfidence in AI assistance.** Overconfidence in AI-generated treatment recommendations may cause a therapist to opt for an AI-generated treatment plan instead of the treatment plan the therapist would have developed independently. Information that should inform a treatment plan but is not included in a patient's electronic records can then not influence the development of the treatment plan.

- **Risk of human overreliance on AI assistance.** Depending on AI tools to develop treatment plans can diminish a therapist's own critical judgment and expertise. This judgment and expertise is important for critically assessing the accuracy of AI-generated treatment recommendations and to provide treatment recommendations when AI tools are not available or a patient has opted out of the use of such tools in their therapy.

- **Risks to patient privacy.** These AI technologies can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on a patient's data may reveal that data in response to adversarial queries.

Patient intake via conversational bots

A conversational bot can interact with a patient during intake and collect medical information and other relevant data about the patient through conversation with the patient. It can also assist in the procurement of a patient's informed consent.

Potential benefits

- **Increased efficiency through automation.** Automating some or most tasks related to patient intake via conversational AI can reduce administrative burdens for a therapist or their administrative co-workers or staff. Conversational bots can take on time-consuming tasks like gathering basic patient information (contact details, demographics, insurance), scheduling appointments, and obtaining consent.

- **Engaging user experiences.** By incorporating interactive elements, such as personalized questions and multimedia content, the AI can make the intake process interactive and enjoyable for patients. The augmented intake procedure can create a comprehensive and detailed patient history without causing the anxiety, stress, or frustration that some patients may feel when being asked to fill out a large number of long questionnaires and forms.

- **Increased sensitivity to a user's background and unique needs.** A conversational bot may communicate with a patient in a way that accommodates the patient's background and unique needs. For example, many AIs are multilingual and can converse with users in a language of the user's choice. They may also accommodate varying literacy levels, cultural backgrounds, and preferred communication methods (e.g., voice, text). Aspects of the communication format, such as volume of voice or font size of text, can be adjusted to accommodate a patient’s preferences and/or hearing or visual impairments.

- **Increased availability of support.** An AI tool for patient intake can operate 24/7. Its around-the-clock availability allows patients to complete their intake at any time before their first appointment at their preferred pace.

Risks

- **Risk of inaccurate AI outputs.** AI-generated intake documentation may include inaccuracies that arise from a conversational bot missing or misunderstanding information provided by a patient. Additionally, a generative AI may hallucinate new, inaccurate information and include it in the intake documentation without the patient’s knowledge.

- **Technological barriers for users.** Some patients may face technological barriers, such as limited access to technology, lack of digital literacy, or difficulty using technology due to age, disability, or other factors. These patients may find completing their intake with a conversational bot difficult or impossible. They may have a negative intake experience and may even decide not to seek mental health support because of this negative experience.

- **Lack of sensitivity to a patient’s background and unique needs.** A conversational bot may fail to detect or misinterpret signs that indicate a patient’s unique needs and may thus not appropriately accommodate such needs. Patients may have a negative intake experience and may even decide not to seek mental health support because of this negative experience.

- **Undesired consequences from out-of-scope use.** Patients may try to talk to an intake bot about a topic that is unrelated to their intake. A conversational bot without appropriate guardrails may engage in a conversation that is outside the scope of its intended use. These out-of-scope conversations can lead the intake bot to produce inappropriate comments or harmful advice. Patients may ask the bot for real-time therapeutic support in emergencies, which a bot designed for patient intake would not be equipped to offer. Especially when an AI intake bot operates outside of standard office hours, it may be hard for a healthcare organization's therapist or staff to detect such an emergency situation and intervene appropriately. Some patients may also try to deliberately deceive an intake bot to gain access to sensitive information about a therapist or their practice. Without appropriate guardrails, an intake bot can fall for such phishing attacks.

- **Risk of human overconfidence on AI assistance.** Hallucinations by generative AI, tech-induced frustration, and other factors may lead to intake documentation being inaccurate or incomplete. A therapist who does not confirm with patients that the information collected by an intake bot is correct, risks basing their analyses and their patients' treatment on wrong information.

- **Risks to patient privacy.** The use of conversational bots can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on a patient's data may reveal that data in response to adversarial queries.

Bot-assisted patient homework

Assigning homework that involves patients interacting with conversational bots can reinforce therapeutic techniques between sessions. For example, such bots can guide patients through CBT exercises, meditation, or day planning.

Potential benefits

- **Increased capacity through automation.** Using conversational bots for therapy homework allows therapists to include guided exercises in treatment plans without allocating a substantial amount of time within therapy sessions to these exercises. As a result, the therapist can use the time within therapy sessions to focus on other aspects of their patient's treatment.

- **Personalized service.** Conversational bots can adjust their responses to a user's personal preferences, which they deduce from user data collected over time. Customized user experiences can make the interactions between a patient and the conversational bot more engaging for the patient. Therapy homework with a conversational bot's assistance may thus be more effective than therapy homework without the bot.

- **Engaging user experiences.** In addition to offering a personalized user experience, some conversational bots can incorporate interactive elements, such as multimedia content and gamified progress updates. These elements can make the therapy homework fun for some patients.

- **Increased sensitivity to a user's background and unique needs.** A conversational bot may communicate with a patient in a way that accommodates the patient's background and unique needs. For example, many AIs are multilingual and can converse with users in a language of the user's choice. They may also accommodate varying literacy levels, cultural backgrounds, and preferred communication methods (e.g., voice, text). Aspects of the communication format, such as volume of voice or font size of text, can be adjusted to accommodate a patient's preferences and/or hearing or visual impairments.

- **Increased availability of support.** A conversational bot can operate 24/7, thereby allowing the patient to complete their guided homework at their preferred time. Specifically, therapy homework can involve practicing coping strategies with the bot's guidance in real-time scenarios.

Risks

- **Risk of inaccurate AI outputs.** A conversational bot's responses can be inaccurate if it misses or misunderstands information provided by a patient. Additionally, a conversational bot may sometimes provide incorrect, inappropriate, or non-evidence-based responses due to generative AI hallucinations or other limitations of the bot's development. These inaccuracies can lead to discrepancies between the conversational bot's guidance and the therapist's advice, potentially causing confusion and/or hindering a patient's progress.

- **Technological barriers for users.** Some patients may face technological barriers, such as limited access to technology, lack of digital literacy, or difficulty using technology due to age, disability, or other factors. These patients may find it difficult or impossible to complete their therapy homework with a conversational bot. Their negative experiences with their homework assignments can diminish the effectiveness of their homework assignments and their therapy.

- **Lack of sensitivity to a patient's background and unique needs.** A conversational bot may fail to detect or misinterpret signs that indicate a patient's unique needs and may thus not appropriately accommodate such needs. Patients may have negative experiences when engaging with their bot-assisted homework assignments. Their negative experiences with their homework assignments can diminish the effectiveness of their homework assignments and their therapy.

- **Undesired consequences from out-of-scope use.** Patients may try to talk to a conversational bot about a topic unrelated to their homework assignment. A conversational bot without appropriate guardrails may engage in a conversation that is outside the scope of its intended use. These out-of-scope conversations can lead a conversational bot to produce inappropriate comments or harmful advice.

- **Risk of human overconfidence on AI assistance.** Tech-induced frustration, a conversational bot's lack of sensitivity to a patient's background and unique needs, and other factors may lead to a bad user experience for patients. When a therapist does not check with patients that their bot-assisted therapy homework is impacting the patient positively, patients may have frustrating experiences in doing their homework. These negative experiences can diminish the effectiveness of the homework assignment and their therapy.

- **Risks to patient privacy.** The use of conversational bots can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on a patient’s data may reveal that data in response to adversarial queries.

Bot-administered talk therapy

Several providers of AI tools offer conversational bots that aim to mimic a therapist in a virtual talk therapy session. Initial research results motivate cautious optimism about these bots' efficacy (Aggarwal et al., 2023; Schillings et al., 2024). Some of these conversational bots are advertised directly to patients as consumers. Others are advertised to health organizations and/or clinicians who may recommend or prescribe the use of these bots to their patients. This document only considers using such bots as prescribed or recommended by their therapist.

Potential benefits

- **Increased capacity through automation.** Using conversational bots for additional talk therapy sessions allows therapists to design treatment plans that include talk therapy for the patient at a higher frequency than their therapist can provide. Patients can benefit from frequent mental health support that can be flexibly adjusted to accommodate a patient’s schedule and preferences. As a result, the patient may be able to manage some challenges of their day-to-day life without the assistance of their therapist. The therapist can then use the time within therapy sessions to focus on other, possibly more complex, aspects of their patient’s treatment.

- **Personalized service.** Conversational bots can adjust their responses to a user’s personal preferences, which they deduce from user data collected over time. Personalized user experiences can make the interactions between a patient and the conversational bot more engaging for the patient. A conversational bot may be able to converse with a patient in the patient’s preferred language and/or conversation style. For patients seeking to improve their mental health through long and deep discussions of a special-interest topic, a conversational bot may be an apt conversation partner.

- **Increased sensitivity to a user’s background and unique needs.** The conversational bot may communicate with a patient in a way that accommodates the patient’s background and unique needs. For example, many AIs are multilingual and can converse with users in a language of the user’s choice. They may also accommodate varying literacy levels, cultural backgrounds, and preferred communication methods (e.g., voice, text). Aspects of the communication format, such as volume of voice or font size of text, can be adjusted to accommodate a patient’s preferences and/or hearing or visual impairments.

- **Increased availability of support.** A conversational bot can operate 24/7, thereby allowing the patient to receive support at their preferred time or in particularly stressful situations.

Risks

- **Risk of inaccurate AI outputs.** A conversational bot's responses can be inaccurate if it misses or misunderstands information provided by a patient. Additionally, a conversational bot may sometimes provide incorrect, inappropriate, or non-evidence-based responses due to generative AI hallucinations or other limitations of the bot's development. These inaccuracies can lead to discrepancies between the conversational bot's guidance and the therapist's advice, thereby potentially causing confusion and/or hindering a patient's progress.

- **Technological barriers for users.** Some patients may face technological barriers, such as limited access to technology, lack of digital literacy, or difficulty using technology due to age, disability, or other factors. These patients may find it difficult or impossible to have helpful conversations with a conversational bot. Their negative experiences with the conversational bot can diminish the overall effectiveness of their therapy.

- **Lack of sensitivity to a patient's background and unique needs.** The conversational bot may fail to detect or misinterpret signs that indicate a patient's unique needs and may thus not appropriately accommodate such needs. Patients may have negative experiences when engaging with their bot-assisted homework assignments. Their negative experiences with their homework assignments can diminish the effectiveness of their homework assignments and their therapy.

- **Undesired consequences from out-of-scope use.** Patients may try to talk to a conversational bot about topics that are either unrelated to their mental well-being or address their mental well-being in harmful ways (e.g., a patient with an eating disorder asking for advice on how to lose weight quickly). A conversational bot without appropriate guardrails may engage in a conversation that is outside the scope of its intended use. These out-of-scope conversations can lead a conversational bot to produce inappropriate comments or harmful advice.

- **Risk of human overconfidence in AI assistance.** Tech-induced frustration, a conversational bot's lack of sensitivity to a patient's background and unique needs, and other factors may lead to a bad patient user experience when a therapist who is overconfident in the capabilities of AI-administered talk therapy may decide not to check with patients that the conversations with the bot are impacting the patient positively. In such a case, negative experiences with the conversational bot can detrimentally affect the patient's progress and well-being. Additionally, a patient can grow overconfident in the capabilities of the conversational bot. This overconfidence can affect the patient negatively when, for example, a patient trusts the conversational bot's advice over their judgment and the advice from their therapist.

- **Risk of human overreliance on AI assistance.** Frequent or excessive use of a conversational bot or a feeling of insecurity or helplessness in situations in which the bot is unavailable can be signs of a patient becoming overreliant on a conversational bot (Hu et al., 2023). Such an overreliance can

  - cause a patient to experience a diminished sense of agency and autonomy,
  - lower the patient's self-esteem and sense of security,
  - develop unhealthy coping mechanisms for navigating stressful situations and other mental-health-related challenges,
  - impair the patient's relationships with other people,
  - and ultimately have detrimental effects on the patient's mental health.

- **Risks to patient privacy.** The use of conversational bots can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the therapist or one of their business associates. Additionally, AIs trained on a patient's data may reveal that data in response to adversarial queries.

Bot-assisted learning experiences

Conversational bots can guide or mentor students and trainees of mental health therapy to enhance their learning experience. For example, a conversational bot can be a source of relevant information that can be accessed via conversation. A conversational bot may guide a student through an exercise, help students test their skills through simulating test scenarios, evaluate a student's responses, and assist in developing personal learning plans.

Potential benefits

- **Increased efficiency through automation.** A conversational bot that serves as a source of information for studying mental health therapy can quickly provide students with the most relevant information for their queries. Such bots may thus make information retrieval for students more efficient than manually searching a textbook and more accurate than an internet search.

- **Personalized service.** Conversational bots can adjust their responses to a user's personal preferences, which they deduce from user data collected over time. Personalized user experiences can make the interactions between a student and the conversational bot more engaging for the student. The student may feel more motivated to learn and have better learning outcomes than without the conversational bot's assistance.

- **Engaging user experiences.** In addition to offering a personalized user experience, some conversational bots can incorporate interactive elements, such as multimedia content and gamified progress updates. These elements can make the learning more fun for some students.

- **Increased sensitivity to a user’s background and unique needs.** A conversational bot may be able to communicate with a student in a way that accommodates the student’s background and unique needs. For example, many AIs are multilingual and can converse with users in a language of the user’s choice. They may also accommodate varying literacy levels, cultural backgrounds, and preferred communication methods (e.g., voice, text). Aspects of the communication format, such as volume of voice or font size of text, can be adjusted to accommodate a student’s preferences and/or hearing or visual impairments.

- **Increased availability of support.** A conversational bot can operate 24/7, allowing students to learn at their preferred time.

Risks

- **Risk of inaccurate AI outputs.** A conversational bot’s responses can be inaccurate if the bot misses or misunderstands information a student provides. Additionally, a conversational bot may sometimes provide incorrect, inappropriate, or non-evidence-based responses due to generative AI hallucinations or other limitations of the bot’s development. These inaccuracies can lead to discrepancies between the conversational bot's guidance and the teacher's advice, potentially causing confusion and/or hindering the student’s learning progress.

- **Technological barriers for users.** Some students may face technological barriers, such as limited access to technology, lack of digital literacy, or difficulty using technology due to age, disability, or other factors. These students may find using a conversational bot to improve their learning experience and outcomes difficult or impossible. When a teacher suggests that students use a conversational bot to augment their learning experience, but the student has negative experiences with the bot, the continued interactions can diminish their learning outcomes.

- **Lack of sensitivity to a patient’s background and unique needs.** A conversational bot may fail to detect or misinterpret signs that indicate a student’s unique needs and may thus not appropriately accommodate such needs. Negative experiences with such a bot due to inappropriate responses can distract and sometimes harm a student’s learning outcomes.

- **Undesired consequences from out-of-scope use.** Students may try to talk to a conversational bot about a topic unrelated to their studies. A conversational bot without appropriate guardrails may engage in a conversation that is outside the scope of its intended use. These out-of-scope conversations can lead a bot to produce inappropriate comments or harmful responses.

- **Risk of human overconfidence in AI assistance.** Tech-induced frustration, a conversational bot’s lack of sensitivity to a student’s background and unique needs, and other factors may lead to a bad user experience for students. When a teacher does not check with students that their bot-assisted learning activities positively impact the student and/or their learning process, the teacher may remain unaware of a student's negative experiences, which can be detrimental to learning outcomes.

- **Risks to patient privacy.** The use of conversational bots can lead to breaches of confidentiality through accidental or purposeful mishandling or unauthorized access of the collected data by the teacher, supervisor, or one of their business associates. Additionally, AIs trained on a student's data may reveal that data in response to adversarial queries.

6. Best Practices for the Use of Artificial Intelligence Technologies by Mental Health Therapists

6.1 Best Practices for the General Use of AI Technologies

Informed consent

- A mental health therapist who uses or intends to use AI technologies that involve
  - recording or transcribing therapy sessions or
  - interactions between a patient and a conversational bot beyond patient-intake procedures

  in their practice should ensure that patients are aware of and consent to use these technologies.

- The mental health therapist should inform patients about (1) the potential benefits and risks of the use of these technologies, (2) what data is collected, stored, or shared with the provider of the AI tool through the use of these technologies, and (3) the risks associated with this data collection and dissemination. For AI technologies involving interactions between a patient and a conversational bot beyond patient intake, the mental health therapist should also inform patients about the form of monitoring that the mental health therapist uses in their ongoing assessment of the effectiveness of these technologies.

- The mental health therapist should ensure that (1) this information is presented to each patient in a manner that is accessible to the patient, (2) the patient understands the presented information, and (3) the patient understands that they have the option to refuse to give consent.

- If a patient asks to review the uses of AI technology to which they consented, the mental health therapist should provide that information.

- A patient may revoke their consent at any time. Once notified about their patient's withdrawn consent, a mental health therapist should act promptly to ensure that all uses of AI technologies that the patient does not consent to anymore cease. The therapist should inform the patient that (1) the therapist has stopped the use of these AI technologies, (2) data shared with providers of AI tools before the revocation of consent may still be stored and used by those providers in accordance with the data-sharing agreement between the therapist and the AI providers.

- Patients must be allowed to refuse to consent to any or all uses of AI technologies in the service they receive from the mental health therapist. If a patient refuses the use of any or all AI technologies, the mental health therapist should make a reasonable effort to provide the service using only the AI technologies to which the patient has consented. If the mental health therapist can't provide the service to the patient under these circumstances, the mental health therapist should make a reasonable effort to connect the patient with a mental health therapist who can.

Disclosure

- A mental health therapist who uses or intends to use AI technologies that are not covered under *Informed Consent* and that

    - include interactions between a patient and a conversational bot during patient-intake procedures,
    - generate text to be included in patients' therapy notes, progress notes, or other parts of their electronic health records,
    - generate individual treatment recommendations or individual treatment plans for patients

  should disclose the use or intended use of these AI technologies to their patients.

- The intake bot can disclose the use of a conversational bot during patient-intake procedures at the beginning of the interaction. The disclosure should include instructions on how the patient can contact a human provider to conduct their intake if they object to conducting their intake with the conversational bot.

- The mental health therapist should inform patients about (1) the potential benefits and risks of the use of these technologies, including the risk of inadequate responses to patient backgrounds and unique needs (see *Risk of inadequate response to patient backgrounds and unique needs*), (2) what data is collected, stored, or shared with third parties through the use of these technologies, and (3) the risks associated with this data collection and dissemination.

- The mental health therapist should ensure that this information is presented to each patient in a manner that is accessible to the patient.

- If a patient objects to using any or all of the disclosed AI technologies, the mental health therapist may offer to provide their therapeutic services without using them. If the mental health therapist cannot or does not offer this option to the patient, the mental health therapist should make a reasonable effort to connect the patient with a mental health therapist who will offer their therapeutic services in a way that aligns with the patient's stated preferences on the use of AI technologies in their therapy.

Data privacy and data safety

- Using AI technologies in behavioral health care introduces risks of sharing sensitive patient data with third parties, intentionally or unintentionally. Before using any AI tool, a mental health therapist should ascertain what data the provider collects.

- If the information that the AI tool collects includes patients' personally identifiable information (PII)—which may include but is not limited to personal health information (PHI)—the therapist should further ascertain
    - how long collected PII is stored,
    - if, how, and for what purpose PII is shared with or sold to other parties,
    - and what precautions the provider of the AI tool has taken to protect collected data from theft and accidental leakage.

- If the provider's practices for data collection, storage, sharing, and safety meet appropriate safety standards (e.g., as stated in the code of ethics and administrative guidance of the licensed mental-health therapist's profession; see Utah Administrative Code R156-60c), the therapist or the healthcare organization that employs the therapist should establish a Business Associates Agreement (BAA) with the provider of the AI tool to verify the provider's adherence to these standards. At the minimum, the BAA should clarify
    - concrete requirements on the provider's data handling procedures to meet appropriate safety standards
    - notification procedures and liability in the event of a data breach.

- Therapists should refrain from using an AI tool if the provider of the AI tool
    - follows practices for data collection, storage, sharing, and safety that do not meet appropriate safety standards or
    - refuses to sign a BAA verifying the provider's adherence to these standards.

- An AI tool's "HIPAA certification" does not alleviate the need for a BAA between its provider and the therapist or the healthcare organization that employs the therapist.

Mental health therapist competence with AI technology

- Mental health therapists should maintain a high level of competence with the AI technologies they employ. This practice involves continuous education and training to understand these AI technologies' capabilities, limitations, and proper use. A thorough understanding of an AI tool's capabilities and limitations includes knowledge about how frequently and under what circumstances one should expect the AI tool to produce inaccurate or undesirable outputs.

- A thorough understanding of the capabilities of an AI tool used in diagnosing or treating a mental health therapist's patients includes knowledge of its scope, i.e., the demographic, cultural, and other population groups for which the AI tool is expected to work as intended. Before incorporating an AI tool in the diagnosis or treatment of their patients, therapists should inform themselves about the scope of the AI tool. When this information is not publicly available, the therapist should inquire directly with the provider of the AI tool if the AI tool has been developed considering the needs associated with the population characteristics of the therapist's patients. For example, AI tools used for diagnosing and treating patients in Utah should generally be developed considering the needs, expectations, and everyday environment of US populations. Other relevant population characteristics to consider may include
    - age ranges (e.g., an AI tool developed using only training data from individuals with ages between 21 and 50 may be unsuitable for the diagnosis or treatment of minors and seniors),
    - socio-economic statuses (e.g., an AI tool developed using only training data from predominantly medium-income households may be unsuitable for the therapist's work with low-income individuals or families),
    - cultural, racial, and religious backgrounds (e.g., an AI tool developed using only training data from people for whom religion plays a minor or no role in their lives may be less accurate or effective for the diagnosis or treatment of patients with a strong religious identity),
    - mental health conditions and their severity (e.g., an AI tool developed to augment the treatment of patients with mild depression may be unsuitable for treating patients with severe depression or mild depression with severe comorbidities), and
    - mental or physical disabilities (e.g., a conversational bot that has no consideration for the accommodations that some physical disabilities require may be unhelpful to patients with mobility disabilities).

- Mental health therapists should consider that
    - more information about the performance, efficacy, and safety of an AI tool (including accuracy and incidence rates for false-positive and false-negative outputs of tools using predictive AI technologies) may become available with time,
    - this information can come from various sources, including (but not limited to) internal or external audits of AI tools or their providers, evaluations conducted by professional societies and healthcare organizations, clinical studies carried out by AI tool providers, universities, and other institutions,

    - capabilities, limitations, and proper use of these AI technologies can be substantially affected by new information about the efficacy and safety of an AI tool,
    - capabilities, limitations, and guidelines for proper use can also be considerably affected by software updates to AI technology and interacting software and hardware products.

- Mental health therapists should make a reasonable effort to stay informed about the efficacy, safety, capability, limitations, and proper use of AI tools. When new information indicates that an AI tool is no longer efficacious or safe for the diagnosis or treatment of all or some patients, therapists should change their use of the AI tool to prioritize the well-being and safety of their patients.

- Mental health therapists may find it helpful to regularly consult information technology specialists to ensure their competence on these issues adjusts to relevant technology updates.

Patient safety and competence with AI technology

- Before incorporating patient-facing AI tools in a patient's treatment, a mental health therapist should weigh potential benefits against potential harms that the patient may experience from interacting with AI tools.

- In assessing potential harms, therapists should consider
    - the patient’s digital literacy,
    - the patient’s propensity for AI over immersion, overreliance, or addiction, and
    - physical, economical, logistic, and time constraints that may make it difficult or impossible for the patient to interact with the AI tool as intended.

- The therapist may talk to them about their technology experiences to assess a patient's digital literacy. It may be useful to demonstrate the patient-facing AI tool to the patient and ask them to rate their confidence in using the technology appropriately.

- To assess a patient’s potential for harm via AI over immersion, overreliance, or addiction, the therapist may consult the patient’s record and ask the patient about prior experiences with over-immersion in, overreliance on, and addiction to technology. Examples may include online social media, video games, smartphones, and other devices or applications that may or may not use AI technologies.

- When a therapist decides to incorporate an AI tool in treating a patient with a history of technology-related behavioral issues—because the likely benefits outweigh the risks—the therapist should adjust their procedures accordingly. For such a patient, a customized approach to continuously monitoring patient-AI interactions may be implemented to address the increased risk of harm (see *Continuous monitoring of patient-AI interactions*).

- A patient's physical constraints may become relevant if patient-facing AI tools require a certain mode of interaction. For example, an AI tool that uses a graphical user interface incompatible with screen readers may be unsuitable for some patients with visual impairments, and an AI tool that interacts with patients via speech may be unsuitable for patients with auditory or speech impairments.

Unique patient needs

- AI technologies trained on data sourced predominantly from one demographic group are likely to be less well-suited for individuals with differing unique needs, which may depend on factors related to different cultures, ways of learning, abilities, educational levels, and economic circumstances. When using generative AI technologies, mental health therapists should consider patients' unique needs.

- The considerations should include
    - the patient's social, economic, racial, ethnic, cultural, or other backgrounds, and
    - the patient's mental health conditions, physical health conditions, and temporary or permanent disabilities.

- If a patient's background or condition falls outside the scope of an AI tool (see *Mental health therapist competence with AI technology*), therapists should generally not incorporate the AI tool in the patient's diagnosis or treatment.

- These considerations apply to AI technologies used in patient-AI interactions (e.g., an assistant for therapy homework) and AI technologies used in mental health therapist-AI interactions (e.g., a diagnostic assistant) because both use cases can cause harm to patients whose unique needs differ from the source of training data.

- If the mental health therapist assesses that using AI technology will benefit a patient despite a discrepancy between the AI technology's trained capacity and the patient's unique needs, the therapist should inform the patient about the potential benefits and risks arising from this discrepancy. If the patient consents to the use of the AI technology after receiving this information, the mental health therapist may proceed cautiously with the use of the AI technology.

Continuous monitoring and reassessment

- To the best of their abilities, a mental health therapist should ensure that interactions facilitated by AI are safe and do not cause harm.

- A mental health therapist should continuously monitor and critically challenge AI outputs for inaccuracies and biases and intervene promptly if the AI produces incorrect, incomplete, or inappropriate content or recommendations.

- The frequency with which a therapist monitors an AI tool's outputs should reflect
  - the therapist's familiarity with the AI tool,
  - the therapist's knowledge of evidence on the accuracy, failure rates, and safety of the AI tool,
  - considerations related to the background and unique needs of individual patients.

- Considerations related to the background and unique needs of individual patients should include that AI tools can be less accurate and more prone to fail in cases of patients with rare backgrounds, rare unique needs, or rare mental-health conditions due to shortcomings of their training (see *Failure Modes for Modern Artificial Intelligence*).

6.2 Best Practices for Therapists' Interactions with Generative AI

Authorship transparency

- When AI technologies contribute to documentation or communication, mental health therapists should disclose their use of these technologies to maintain transparency. Such disclosure can, for example, be included as part of a signature (e.g., similar to an automated email signature) or in the form of an authorship note (e.g., see *Authorship note and Acknowledgments*) at the end of a text or in the form of a disclosure (e.g., the statement "The following text was generated using with the AI tool [name of AI tool]") at the beginning of the text.

- Acknowledging AI involvement helps preserve integrity and allows for appropriate attribution of sources in professional records. It also facilitates an accurate assessment of the veracity of information in patient records subpoenaed or divulged in discovery. (For example, it may be essential to know that AI-generated text to understand that words or sentences that seem out of context or very unusual could result from AI hallucinations.)

Review of AI-generated text

- Therapists should review all AI-generated text about a patient, the patient's condition, or the patient's progress before they
  - include the text in a patient's electronic health record,

    - share the text with another healthcare provider who provides healthcare to the patient or
    - use the text to determine a diagnosis or treatment decision.

  This ensures accuracy, relevance, and appropriateness, allowing mental health therapists to correct any errors or omissions the AI may have made.

Reasonable doubt about AI-generated diagnosis and treatment proposals

- Mental health therapists should critically evaluate any diagnoses or treatment suggestions AI technologies offer. Recognizing that AI is a supplement—not a replacement—for professional judgment helps prevent overreliance on technology and supports evidence-based practice.

6.3 Best Practices for Patients' Interactions with Generative AI

Patients' relationship with AI technology

- Before suggesting the use of AI technology to a patient, a mental health therapist should consider the patient's views about technology and how they use it, including strengths, needs, risks, and challenges. These considerations are important for assessing whether the potential benefits outweigh the risks of using this AI technology in the mental health therapist's service to the patient.

Primary commitment to patients

- Patients' well-being should always precede the use of AI technology. Mental health therapists should ensure that (1) AI technologies augment their services for the benefit of their patients and (2) the potential benefits outweigh the risks for each patient in the context of their unique needs.

Service interruptions and malfunctions

- Mental health therapists should prepare for potential AI service disruptions by having backup plans in place. Communicating these contingencies to patients reassures them that their care will continue smoothly even if technical issues arise.

Conduct of generative AI

- Conversational bots should interact with patients ethically and professionally. They should provide accurate information and appropriate responses.

- The more the capabilities or appearance of AI technology in patient-AI interactions make them similar to interactions with a licensed mental health therapist, the more the technology can be expected to comply with the ethical guidelines and standards of the mental health therapist's profession (e.g., confidentiality, competence, beneficence).

- Before incorporating a conversational bot into their therapeutic practice or recommending its use to patients, a mental health therapist should examine and verify that the bot displays awareness and alignment with ethical and professional best practices commensurate with its capabilities and appearance.

Responsibility in emergency situations

- A generative AI technology that interacts with a patient similarly to how a mental health therapist would interact with the patient can be expected to show a similar level of responsibility in an emergency situation. Mental health therapists should inform themselves about AI technology's capability to act responsibly in emergencies and escalate situations if needed.

- If an AI technology does not have the capabilities to act responsibly in an emergency, the mental health therapist should not proceed with the use of the AI technology or proceed only after having taken appropriate precautions. These precautions include (1) considering the risk of their patient seeking help from the AI technology in an emergency based on their patient's vulnerability and relationship to technology, (2) informing the patient about the limits of the AI technology's capability in emergencies, (3) informing the patient about how to seek appropriate help in emergencies.

Continuous monitoring of patient-AI interactions

- The mental health therapist should monitor interactions between patients and conversational bots that occur as part of the patient's treatment to continuously assess the effectiveness of the use of AI technologies in the service provided to the patient.

- The mental health therapist should select a form, frequency, and practice of monitoring patient-AI interactions that adequately reflects the mental health therapist's experience with the AI technology, the AI technology's robustness and safety, and the patient's unique needs. (In medium or higher risk scenarios, this may involve a timely review of entire chat logs. In low-risk scenarios, this may affect the mental health therapist, who initializes regular discussions about the patient's experience with AI technology.)

- The mental health therapist should explain the frequency and practice of monitoring patient-AI interactions to the patient and receive the patient's informed consent prior to using the AI technology and the mental health therapist's proposed monitoring scheme.
- Changes in technology or a patient's needs may justify changes to the monitoring scheme. If a mental health therapist decides that a change is appropriate, they should discuss the proposed change with the patient and only proceed with the change after receiving the patient's informed consent.

Mandated reporting of patient-AI interactions

- A mental health therapist's mandated reporting duties include patient-AI interactions to the extent that the therapist monitors them. The therapist should inform the patient about the therapist's reporting duties in this context.

6.4 Best Practices for Supervisors' Interactions with Generative AI

Review of AI-generated text

- Supervisors should thoroughly review all AI-generated text before using it to assess a student's performance or the well-being of a student's patients.

Reasonable doubt about AI-generated student assessments

- Mental health therapists should critically evaluate any student assessment offered by AI technologies. Recognizing that AI is a supplement—not a replacement—for professional judgment helps prevent overreliance on technology and ensures that students receive fair, accurate, and helpful training and feedback.

6.5 Best Practices for Students' Interactions with Generative AI

Student's relationship with AI technology

- Before suggesting the use of AI technology to a student, a supervisor should consider the student's views about technology and the ways in which they use technology, including strengths, needs, risks, and challenges. These considerations are important for assessing whether the potential benefits outweigh the risks of using this AI technology in the training and supervision of students and trainees. Risks for the student and risks for the student's patients should be considered.

Primary commitment to students

- The well-being of students and the positive outcome of their training should always take precedence over the use of AI technology. Supervisors should ensure that (1) AI technologies augment the educational experience of their students and (2) the potential benefits for the student outweigh the risks for each student and their patients.

Service interruptions and malfunctions

- When using AI technology in the education and training of mental health therapists involves a risk of overreliance or dependence on AI technology, supervisors should discuss these risks with their students. They should also prepare for potential AI service disruptions by having backup plans to mitigate harms that may befall students or students' patients. Communicating these contingencies to students reassures them that their care will continue smoothly even if technical issues arise.

Continuous monitoring of student-AI interactions

- Interactions between students and conversational bots that occur as part of the student's training and are overseen by a supervisor should be monitored by the supervisor to continuously assess the effectiveness of using the employed AI technologies in the education and supervision provided to the student.

- Supervisors should select a form, frequency, and practice of monitoring patient-AI interactions that adequately reflects the supervisor's experience with the AI technology, the student's experience with the AI technology, and the AI technology's robustness and safety. When the supervisor or the student has no or little experience with AI technology, monitoring may involve a timely review of entire chat logs. When the supervisor and the student have substantial experience with AI technology, regular discussions may be held between the supervisor and the student about the student's experience with AI technology.

- The supervisor should explain the frequency and practice of monitoring student-AI interactions with the student and obtain the student's informed consent prior to using the AI technology and the supervisor's proposed monitoring scheme.

7. Authorship Note and Acknowledgments

This document was prepared by Alice C. Schwarze and Zachary M. Boyd for the Utah Office of Artificial Intelligence Policy (OAIP) and the Utah Division of Professional Licensing (DOPL).

Google Gemini and Anthropic's Claude AI were used in the preparation of this document. All AI-generated content was reviewed and/or revised by the authors.

The authors would like to thank Jared Ferguson, Aaron J. Fisher, Cole Hooley, Zac E. Imel, Jana Johanson, Nina de Lacy, Tamara Moores Todd, Frederic Reamer, Mark Steinagel, Sarah Stroup, Greg Whisenant, Brady Young, the American Psychological Association, and the Utah Behavioral Health Board for their helpful suggestions and comments.

8. References

Aggarwal, A., Tam, C. C., Wu, D., Li, X., & Qiao, S. (2023). Artificial intelligence–based chatbots for promoting health behavioral changes: systematic review. Journal of medical Internet research, 25, e40789. URL: *https://doi.org/10.2196/40789*

American Association for Marriage and Family Therapy. (2015). Code of Ethics.
URL: *https://www.aamft.org/AAMFT/Legal_Ethics/Code_of_Ethics.aspx*

American Mental Health Counselors Association. (2020). Code of Ethics.
URL: *https://www.amhca.org/events/publications/ethics*

American Mental Health Counselors Association. (2023). Code of Ethics Addendum: Addressing Artificial Intelligence. URL:
https://www.amhca.org/viewdocument/amhca-code-of-ethics-addendum-addr#:~:text=AMHCA%20Code%20of%20Ethics%20Addendum%3A%20Addressing%20Artificial%20Intelligence%3A%202023&text=CMHCs%20clearly%20disclose%20to%20clients,person%2Dto%2Dperson%20contact.

American Nurses Association. (2014). Psychiatric-Mental Health Nursing: Scope and Standards of Practice (2nd ed.).

American Psychiatric Association. (2013). Principles of Medical Ethics With Annotations Especially Applicable to Psychiatry. URL: *https://www.psychiatry.org/psychiatrists/practice/ethics*

American Psychological Association. (2016). Ethical Principles of Psychologists and Code of Conduct.
URL: *https://www.apa.org/ethics/code*

Association of Social Work Boards. (2014). Model Regulatory Standards for Technology and Social Work Practice. URL:
https://www.aswb.org/wp-content/uploads/2015/03/ASWB-Model-Regulatory-Standards-for-Technology-and-Social-Work-Practice.pdf

Ettman, C. K., & Galea, S. (2023). The Potential Influence of AI on Population Mental Health. Journal of Medicine, Surgery, and Public Health, 10, Article e49936. URL: *https://doi.org/10.2196/49936*

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., and Agarwal, S. (2025). How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Controlled Study. arXiv:2503.17473.
URL: *https://arxiv.org/abs/2503.17473*

Hu, B., Mao, Y., & Kim, K. J. (2023). How social anxiety leads to problematic use of conversational AI: The roles of loneliness, rumination, and mind perception. Computers in Human Behavior, 145, Article 107760. URL: *https://doi.org/10.1016/j.chb.2023.107760*

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., Paulus, M. P., Krystal, J. H. & Jeste, D. V. (2021). Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 6(9), 856-864. URL: *https://doi.org/10.1016/j.bpsc.2021.02.001*

Narayanan, A., & Kapoor, S. (2024). AI Snake Oil. What Artificial Intelligence can do, what it can't, and how to tell the difference. Princeton University Press.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., & Lee, K. (2023). Scalable Extraction of Training Data from (Production) Language Models. arXiv:2311.17035. URL: *https://arxiv.org/abs/2311.17035*

National Association for Alcoholism and Drug Abuse Counselors / National Certification Commission for Addiction Professionals. (2021). Code of Ethics.
URL: *https://www.naadac.org/assets/2416/naadac_code_of_ethics_06012025.pdf*

National Association of Social Workers. (2021). Code of Ethics.
URL: *https://www.socialworkers.org/About/Ethics/Code-of-Ethics/Code-of-Ethics-English*

National Board for Certified Counselors. (2023). Code of Ethics.
URL: *https://nbcc.org/assets/ethics/nbcccodeofethics.pdf*

National Association of Social Workers, Association of Social Work Boards, Council on Social Work Education, & Clinical Social Work Association. (2017). Standards for Technology in Social Work Practice.
URL:
https://www.socialworkers.org/Practice/NASW-Practice-Standards-Guidelines/Standards-for-Technology-in-Social-Work-Practice

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of Medicine, Surgery, and Public Health, 3, Article 100089.
URL: *https://doi.org/10.1016/j.glmedi.2024.100099*

Schillings, C., Meißner, E., Erb, B., Bendig, E., Schultchen, D., & Pollatos, O. (2024). Effects of a chatbot-based intervention on stress and health-related parameters in a stressed sample: randomized controlled trial. JMIR Mental Health, 11(1), e50454. URL: *https://doi.org/10.2196/50454*

Schwarzschild, A., Feng, Z., Maini, P., Lipton, Z., & Kolter, J. Z. (2025). Rethinking LLM memorization through the lens of adversarial compression. Advances in Neural Information Processing Systems, 37, 56244-56267.
URL: *https://proceedings.neurips.cc/paper_files/paper/2024/hash/66453d578afae006252d2ea090e151c9-Abstract-Conference.html*

Thakkar, A., Gupta, A., & de Sousa, A. (2024). Artificial intelligence in positive mental health: A narrative review. Frontiers in Digital Health, 6, Article 1280235. URL: *https://doi.org/10.3389/fdgth.2024.1280235*

Utah State Code. URL: *https://le.utah.gov/xcode/code.html*

Utah Administrative Code R156-60c. URL:
https://adminrules.utah.gov/public/rule/R156-60c/Current%20Rules?