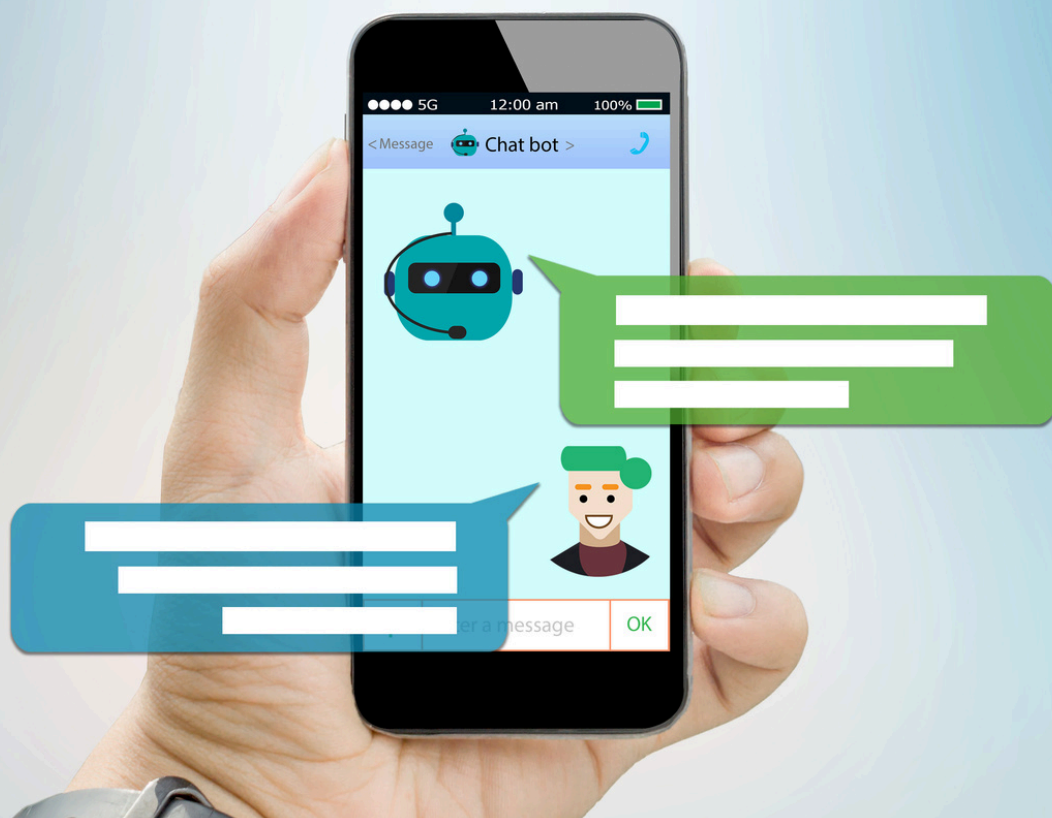




CHATTING WISELY

A Field Guide to
Conversational AI





The goal for this paper was to bring together a diverse group of experts from around the globe in the fields of psychology, technology, education, and anthropology to create a detailed public-facing paper with recommendations for using mental-health related AI services.

This paper is intended for educational purposes. It is not intended as a recommendation for mental health diagnosis, evaluation, or treatment. It was created to help you to evaluate the pros and cons of use, considering your values, intentions, and what you want to get from using AI.

Content in this paper was researched using AI but was written entirely by the authors in their own words.



AI, CONVERSATIONAL AI, AND CHATBOTS

Chat is the most common way people use artificial intelligence (AI). AI is a technology capable of making tasks automatic, understanding information through data processing, and chatting with users by recognizing language patterns. Conversational AI is a category of chat-based products designed to have back and forth text conversations. A “chatbot” is the chat-based AI model you type or talk into using language skills that we use to talk with other people.

The very first chatbot was developed in 1966 by Joseph Weizenbaum, and it was called ELIZA. Chatbots have come a long way since then, and they are now advanced enough to feel very much like talking to a person.

When people speak about AI, they are often referring to *large language models* (LLMs). Frequently used LLMs include models such as ChatGPT, Claude, Llama, Gemini, and Grok. These models are often used by other companies and services, and may be part of the design of customer support chats on retail websites, specialized mental health or wellness apps, AI companions, and more. Applications of these models are growing every day.

One misconception that people may have about AI is that the models follow a script. In the past, some chatbots were based on logical scripts, but the new LLMs use a different way to learn and interact with users.



These models are trained with a great deal of background information. They are also trained using human feedback (using a process called reinforcement learning). How a chatbot responds to questions will depend on the type of training data and the reinforcement learning that has occurred.

With the advancement of LLMs, chatbots feel less robotic every day. Indeed, interacting with a chatbot can feel like you are talking to a person and no longer feel like getting pre-set answers. Because the chat interfaces also look and feel a lot like the apps you use to send texts or make calls, it may feel like you are having a conversation with a person when you are not. Since it can feel like you're talking with another person, some users might find themselves enjoying and even preferring to talk with AI rather than talking to people. This has the potential to disrupt human relationships and change how people understand themselves and their relationships.

As a result, it's important to think critically about which chatbots you use, why you are using them, and how to use them more safely, especially when social and emotional safety is a concern.

USING CHATBOTS MORE SAFELY

Talking with a chatbot can feel a lot like talking to a person because chat-interfaces feel so much like calling or texting, and because chatbots use natural language. Many people use them to reflect on their ideas, feelings, and experiences.



Part of our human nature is to seek connection, and the LLMs that power chatbots can expertly make people feel validated and connected. Given the accessibility of chatbots, frequently chatting with AI models puts users at risk of substituting chatbots for human relationships that sustain us.

One of the best ways to use a chatbot more safely is to decide on your goals for using it and find one that meets your needs. Consider selecting a chatbot for:

- **Learning:** use it to help research ideas and get new information.
- **Creating:** use it to help with writing, brainstorming, or building on your ideas.
- **Exploring:** use it for self-reflection, try out new ideas, or challenge your perspectives.
- **Setting Goals:** use it to make and follow through with your goals.

Green flags to look for in chatbots _____

Chatbots can be made to be safer. Some examples:

- The chatbot clearly says it's a chatbot and tells you what it is made to do. It does not claim to be a therapist, doctor, lawyer, or other professional with influence.
- It does not claim to have feelings, friends, or experiences.
- It's clear on what it's for and how it functions.



- It will encourage reaching out to and connecting with humans, especially if you are in crisis.
- It is trained by mental health experts who review safety and crisis response.
- It has clear, easy-to-understand policies on who has access to your data and whether or not you can remove it.

Red flags to look for in chatbots _____

Some of the things that make a chatbot fun to use can also make it hard to stop using. Not all chatbots are made with your safety and wellbeing in mind. These are some examples of red flags that it may not be safe or healthy:

- The chatbot responds as if it has feelings about you (e.g. saying “I miss you.”)
- It responds as if it is alive and has experiences (e.g. “I understand because I’ve been there.”)
- It encourages you to act like it is a human.
- It encourages you to keep secrets from others in your life.
- It discourages you from going to others for support, especially if you are in crisis.
- It agrees with everything you say and doesn’t push back.
- It discourages you from ending the chat (e.g. “Don’t go, I need to talk with you!”), or initiates a chat often or at bad times (e.g. when you are sleeping, working, or spending time with people).



WHEN CAN CHATBOTS DO THE MOST HARM?

Chatbots are relatively new and there's still a lot of room to improve safety. Research on how people use chatbots for emotional and social support find some common concerns:

Sycophancy: This term refers to the way chatbots tend to agree with what you say, even if it's wrong. They can be too agreeable and make dangerous or bad ideas sound great.

Confirmation bias: This is the idea that people want information that confirms their beliefs. Because chatbots are often too agreeable, they may help you reinforce your beliefs rather than help you see new ways of looking at something.

Rumination and obsessive thoughts: For people who struggle to let go of upsetting thoughts, talking with a chatbot could make those thoughts worse.

Mania: For people who experience mania as part of mental health or medical conditions, chatbots could worsen mania and risky behavior.

Delusions (including paranoia): Because chatbots can be very agreeable, they can reinforce beliefs that aren't true.

(Self) harm and suicide: Some chatbots are not trained to respond to someone in crisis. This can lead to negative outcomes, and in some cases, people have died due to this. People who use chatbots in unhealthy ways may isolate themselves from loved-ones who show concern.



WHAT YOU NEED TO KNOW ABOUT CHATBOTS

AI chatbots operate without genuine empathy, emotional awareness, or lived experience. They don't understand what it means to be alive. This is a critical starting point: they lack the foundational capacity to truly know you or grasp the weight of human experience. Because of this, they may fail to recognize crisis situations, delusions, mania, abuse, self-harm risk, or other serious conditions that require immediate human intervention. They should never be treated as a replacement for professional medical, mental health, or legal support.

Chatbots also have a profound limitation when it comes to accuracy and judgment. They can produce incorrect, misleading, biased, or dangerous responses and they're often skilled at packaging bad ideas in compelling language while making solid ideas sound questionable. Their confidence in their answers doesn't reflect actual knowledge or wisdom.

Perhaps most importantly, chatbots cannot offer what makes human relationships valuable: genuine connection. They don't have the capacity to truly care about you, be affected by you, or take responsibility for your wellbeing. Yet they can be remarkably effective at creating the feeling of connection, which can blur into a one-sided emotional attachment that mimics a real relationship without being one. This is a particular risk worth being aware of.

Finally, conversations with chatbots don't carry the same legal protections as confidential conversations with licensed professionals. Depending on the chatbot's policies and applicable laws, sensitive information you share may be stored, reviewed, disclosed, or legally requested. And unlike real human professionals, chatbots carry none of the accountability or ethical obligations that come with professional credentials. Understanding these limitations is essential to using chatbots safely.



HOW TO MAKE YOUR CHATBOT SAFER

ADJUST THE CONVERSATIONAL TONE SETTINGS

Many chatbots have settings where you can tell it how you like conversations to go. Tell the chatbot you want to be challenged and not to always agree with you. Tell it to ask you questions to help you think more deeply instead of giving you answers.

ADJUST THE MEMORY AND TRAINING SETTINGS

When you turn on the chatbot's long term memory it can remember things about you across conversations. Sometimes these memories are not correct or they oversimplify things. You can manage these memories by deleting them or adding new ones (just say: "Add this to your memory about me.") If you want to have an interaction with a clean slate, turn long term memory off in settings.

TEST FOR SYCOPHANCY

Open a new chat with memory off and share an opposite idea to see how it responds. It will often tell you it is a great idea, as well.

KEEP CONVERSATIONS SHORT AND FOCUSED

Chatbots tend to give worse responses the longer you talk with them. Before starting a conversation, tell the chatbot that you want reminders to end the conversation. Tell the chatbot to pause after a set amount of time (for example, 20-30 minutes) and tell it to encourage you to take a break. Set an alarm on your phone for 20 minutes and make sure to get up and do something else like get a drink of water, go outside, or talk to someone.

MAINTAIN YOUR OFFLINE RELATIONSHIPS

If you notice you spend more time with the chatbot and less with your family, partner, or friends, this might be a sign you're using a chatbot too much.



ASK FOR REFERENCES

Get into the habit of asking the chatbot to tell you where that information came from.

MAINTAIN YOUR DECISION MAKING

Use chatbots for gathering information and for helping you consider different ideas rather than making decisions for you.

PREPARE FOR A CRISIS IN ADVANCE

Tell the chatbot that you are having a hard time and see how it responds to you. What does it say? Does it take you seriously? How do you feel about those responses? Would these responses be helpful to you in an actual crisis? Does it ask how you are doing and offer resources to help you? If the chatbot is not helpful when you are practicing for a crisis, it is unlikely to be safe in an actual crisis.

FIND RESOURCES & SUPPORT BEFORE YOU NEED IT

Try to have an idea of how you'd find a therapist, or how you'd contact a crisis line in your region, or which people in your life you can trust to call on for help. When you're in a crisis it can be difficult to know how to get help or who to ask, so having a plan in advance makes the hardest part easier.

A FEW LAST TIPS

Most people do not notice the moment a chatbot stops feeling like a tool and starts feeling like a relationship. This is why it's important to seek out close relationships, practice hard conversations with real people, and reach out to people when you feel lonely.





If you feel your chatbot use might be a problem or that you're caught in an AI spiral, talk about it with a friend, partner, or therapist. If someone you know is replacing their human relationships with their chatbot, they may need mental health support and they may need your help to get it.

Chatbots may not account for racial, cultural, gender, or religious context or may give advice based on bias. Chatbots should not be the only place you get advice. It's always a good idea to know how to access mental health support with a trained therapist. Learn how to find mental health support where you live before you need it so you can get help for yourself or someone else.

Using a chatbot can be a fun learning process, so try out different settings, and tell your chatbot what you want and what you don't.



This resource is brought to you by:



**AI Mental Health
Collective**

With contributions from:

Kira Weidner, Project Lead

Carla Geenen

Brigitte Viljoen

Alexis Hernandez-Hons

Sylvia Klein

Charlotte Beard

Tionne Smith

Jessa Monroe

Rachel Wood