



META
COMMUNITY FORUM

INFORMATION PACK

OCTOBER 2024



EXECUTIVE SUMMARY

This information pack will provide you with an overview of how the Community Forum is structured, as well as enough basic knowledge to have an informed discussion with other participants during the Community Forum.

It should take about 1 hour to read through this guide from start to finish.

There have been significant breakthroughs in Artificial Intelligence (AI) over the past couple of years. You have been invited to learn about and discuss some important issues with this new technology along with other people from all over the world. Below are some of the concepts that you will be introduced to later in this document.

Artificial Intelligence (AI) is a type of computer technology that makes machines analyze, decide, predict, and learn like humans.

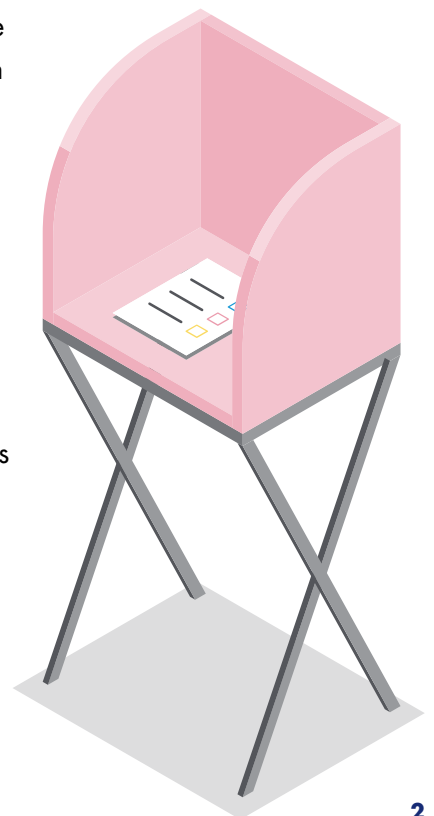
Generative Artificial Intelligence (GenAI) is a type of Artificial Intelligence that generates content or actions in response to a prompt.

AI chatbots are a type of AI system that respond to written or verbal prompts by generating content such as text or images. [OpenAI's ChatGPT](#), [Microsoft's Copilot](#), [Google's Gemini](#), [Anthropic's Claude](#), and [Meta's MetaAI](#) are some popular examples of AI chatbots.

Beyond text-generating AI, there are also other AI systems that can generate content such as music, images, video, and even computer code. GenAI systems that use multiple types of input and produce multiple types of output such as text-based responses, images, video and/or audio are known as **multimodal AI**.

Over the past year, we've seen new and exciting developments in AI in the form of "AI agents". **AI agents** are a type of AI that can respond to written or verbal prompts by performing actions for you, on your behalf. Think of AI agents as virtual personal assistants. AI agents are capable of scheduling meetings on your behalf, writing code and testing it, or purchasing items for you. Popular examples include Amazon's Alexa, which can purchase items from Amazon for you, or Apple's Siri, which can text a friend on your behalf.

Lots of companies, including Meta, are working on developing AI that will bring the most benefit to humans. It is critical for companies and developers to consider the limitations of AI in order to create AI responsibly and to minimize potential risks to users and society at large.



You will discuss the following questions alongside other participants over the course of a day.

1. How should AI agents provide proactive, personalized experiences for users?
2. How should AI agents and users interact?

→ [Click here](#) if you want to go directly to the proposals on page 26.

Below is an overview of the sessions you will participate in.

BEFORE THE EVENT	Educate yourself using this information pack.	To prepare, please read this guide. It contains all the information you'll need to participate in the event.
	Create Stanford platform account and join peers to deliberate.	Before the Community Forum, please follow the instructions we sent you via email to create an account on the Stanford platform. Please check your technology before the Forum to ensure access.
DURING THE EVENT	Deliberation Session 1	When the event begins, you will join peers virtually to deliberate on the following question: How should AI agents provide proactive, personalized experiences for users?
	AI Expert Panel 1	After the first deliberation session, you will leave the platform and join an AI expert panel. You will be able to ask questions about topics from the first session and hear from experts in the AI field.
	Deliberation Session 2	You will return to the Stanford platform to virtually deliberate with your peers on the following question: How should AI agents and users interact?
	AI Expert Panel 2	After the second deliberation session, you will leave the platform to join an AI expert panel. You will be able to ask questions about topics from the second session and hear from experts in the AI field.
	Reflection Session	Once again, you will return to the Stanford platform and have a final discussion with your peers to reflect on themes from the day's deliberation sessions and expert panels.
AFTER THE EVENT	Complete post-deliberation final survey.	After the Community Forum, you will complete a final survey. This is your chance to share your personal opinions about the proposals after you have discussed them with others and have heard from the AI experts.

INFORMATION PACK OUTLINE



Introduction	5
A Note from Meta: Why we are doing this Community Forum	6
Your role as a Community Forum participant	7
The topic of the Meta Community Forum	8
An overview of Artificial Intelligence	9
What is “AI”?	10
Generative AI: A new type of Artificial Intelligence	10
Building an AI Model	10
Step 1: “Training” the model	11
Step 2: Fine-tuning the model	12
How are AI agents different from non-agentic AI chatbots	15
Generative AI Technologies	17
The limitations of generative AI	19
Data Dependency	20
Data Privacy	21
Model Limitations	22
User Behavior	22
Building generative AI responsibly	23
What do we mean by “Responsible AI”?	24
Proposals	26
Session 1 How should AI agents provide proactive, personalized experiences for users?	29
Session 2 How should AI agents and users interact?	39
Guiding principles for participation	50
Glossary of terms	51
Endnotes	54

INTRODUCTION

Welcome to Meta's Community Forum on Generative Artificial Intelligence (GenAI)!

This is a unique opportunity to talk to a diverse group of people about important issues regarding GenAI and to offer your feedback directly to Meta. You are one of hundreds of people from all over the world who have been brought together to discuss the following questions:

1. How should AI agents provide proactive, personalized experiences for users?

2. How should AI agents and users interact?

This may be the first time you hear or read the term "generative AI", but that's okay! This guide is here to help you understand and have an informed conversation during the Forum. Take your time reading this document and go at your own pace.

In this Information Pack, we will explain:

- What generative AI is and how it is created;
- The uses and limitations of generative AI; and
- Current approaches to developing generative AI responsibly

In the second half of this information pack, you will read about a list of ideas that technology companies could use as they seek to develop GenAI more responsibly. We call these ideas **proposals**.

During the virtual Community Forum, you will be able to speak about the proposals and their tradeoffs with your peers in small groups. At the end of the Community Forum, you will have the opportunity to share your personal opinions about the proposals in a confidential survey. Your responses will help Meta decide how to continue building AI safely and responsibly.

All Community Forum participants are provided with this information, and are expected to read this guide carefully. It is essential to help prepare you for your discussions to come.

We sincerely thank you for your time and attention.

Meta

Sponsor of this event



Deliberative Democracy Lab

Partner and pioneer of Deliberative Polling



i. This Information Pack has been produced with thanks to expert input from Alexandra Cirone, Dr. Hatice Kose, Pranesh Prakash, and Dr. Pushpak Bhattacharyya. Their consultation does not imply an endorsement of the proposals in the Information Pack or the results of the Community Forum.

A Note from Meta:

Why we are doing this Community Forum

Recently, few technological advances have produced as much fascination among the public as the remarkable progress in the field of Artificial Intelligence. This is especially true of generative AI. In 2022, [ChatGPT](#) was released to several countries around the world. For many, this AI chatbot was their first exposure to a highly sophisticated generative AI model. Since then, several AI chatbots have been released including [Microsoft's Copilot](#), [Google's Gemini](#), [Anthropic's Claude](#), and our own [Meta AI](#).

Generative Artificial Intelligence has been compared to innovations like the printing press, electricity, and the internet. Being able to create text, audio, images, and actions with simple prompts could unleash new economic and creative possibilities. In addition, technology that can act as a sort of personal assistant to humans can free up a lot of time for humans to focus on other things. Picture having a personal assistant that handles things on your behalf like scheduling doctor's appointments or researching vacation spots. Generative AI could make it possible for everyone to have their own digital assistant, a smart agent that can take on various tasks and act on your behalf.

Yet this technology is not without risks. Since Generative AI can create "new" information, instead of finding, predicting, or classifying previously available information, there are new concerns that accompany this new technology. Some worry that Generative AI could increase the risk of widespread false information, large political divides, and harmful prejudices. Others have concerns about these risks, but they also doubt the ability of governments and technology companies to effectively manage them. It is possible that as this technology continues to evolve, humanity could be threatened; it is also possible that these concerns are premature and exaggerated.

Too often important conversations at tech companies happen between small groups of people behind closed doors. [Community Forums](#) like this are one part of Meta's efforts to change that.

We would like to open up these conversations to you, and hundreds of others around the world, about how to safely develop these powerful technologies. These are new technologies, and this is your chance to have a say in the direction that Meta and other companies take in their development of AI products. Your role in this process is important, and we hope you take it as seriously as we do.



Your role as a Community Forum participant

You and groups of others from different countries will spend the day of the Community Forum (October 19, 2024) discussing generative AI. You will learn about how generative AI models are made, some of their potential uses, their limitations and challenges, and what kinds of principles we might use to guide the responsible development of generative AI.

This online event you are joining is based on a method called Deliberative Polling.

Developed by James Fishkin at Stanford University's Deliberative Democracy Lab over the last 30 years, this method of gathering the public has been used in 110+ projects across 30+ countries. Deliberative Polling is an opportunity for people to talk with others in their society and discuss the reasons behind why some people may think one way or another. It is really a chance to listen and learn from each other. Past participants have enjoyed the process and we think that you will too!

For more information on Deliberative Polling or Stanford's Deliberative Democracy Lab, please visit <https://deliberation.stanford.edu/what-deliberative-pollingr>.



The topic of the Meta Community Forum

You may not notice it, but **Artificial Intelligence (AI)** is probably part of your everyday life. It is there if you use voice assistants like those on your cell phone or when you get suggestions for videos and social media posts. AI is often at work in the background making our lives simpler and more efficient.

Generative AI is a relatively new type of AI that can understand simple requests, and create new content or complete tasks from them. As this technology has improved, people have started using it for a variety of tasks like grammar support when writing, solving math problems, creating songs, and even making art! Suddenly, anyone with access to these tools can create content quickly and easily. While the results aren't perfect, they're improving every day.

These generative AI tools have the potential to enhance our lives. They could automate and streamline many tasks, making our work more efficient. In schools, generative AI could provide tutoring for students and help teachers create educational materials. For writers and artists, it could speed up the process of producing articles, artwork, and music.

However, the widespread use of generative AI also comes with diverse challenges and concerns. For example, the ability to produce realistic content quickly might make it difficult to tell if something was made by a person or by AI. There are also concerns about how these AI systems might use personal data, infringe on a user's privacy, or potentially replace human jobs.

There are different ways that AI technology can be made better and safer. For this event, we would like your opinion on **how AI agents should provide proactive, personalized experiences for users and how AI agents and users should interact.**

When you are discussing the proposals in your small groups, please keep in mind that:

1. The proposals have tradeoffs which means you may have to compromise between benefits and drawbacks since you are not able to have everything that you want at the same time. Based on your perspective, you may prefer one goal over another. You will need to think about these tradeoffs and decide which ones you think are the most important.

2. We don't have all the answers. Generative AI is improving quickly, but we cannot predict exactly how it will change things. That is why your perspectives are so valuable. You might have some great ideas about how AI agents could be used to make the world a better place. Or you might have some concerns about how they could be misused. Either way, it is important to share your thoughts and listen to what other people have to say.

One thing to keep in mind is that today, generative AI is still just a tool. They can be used for good or bad, depending on how people choose to use them. So it is up to us to think carefully about how we want generative AI to be personalized for us and to make sure that we are using these tools in responsible ways.



AN OVERVIEW OF ARTIFICIAL INTELLIGENCE



What is “AI”?

Artificial Intelligence¹ (AI) is a type of computer technology that makes machines analyze, decide, predict, and learn like humans. AI helps computers and systems perform tasks such as understanding language and making decisions.

GENERATIVE AI: A NEW TYPE OF ARTIFICIAL INTELLIGENCE

Generative AI (GenAI) is a type of Artificial Intelligence that generates content or actions in response to a prompt. It is good at creating new content such as text, images, music, or tasks based on patterns it learns from existing data. GenAI helps machines generate original outputs that mimic human language, creativity, and actioning.

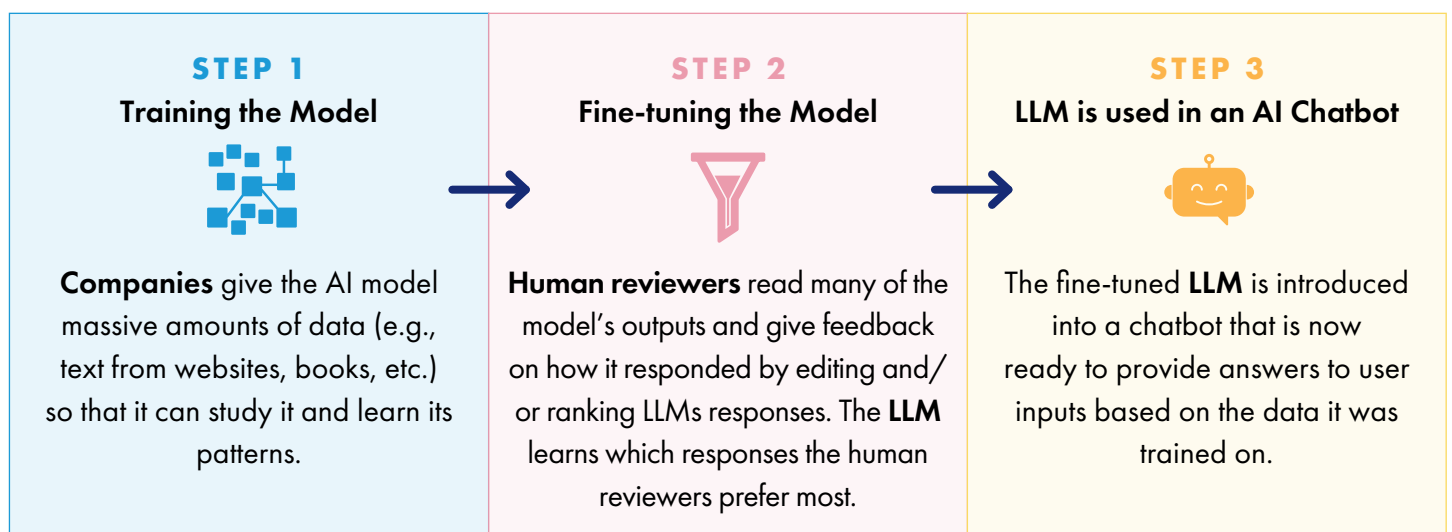
So how exactly does generative AI like [ChatGPT](#) or [MetaAI](#) become so good at talking and eventually performing actions? Let's find out.

Quick Note: You will see the terms, “non-agentic AI chatbots” and “AI agents” frequently in this guide.

Non-agentic AI chatbots are user-friendly products that allow you to interact with machine learning models to create new content. They do not have the ability to independently perform tasks.

Agentic AI² (AI agents) describe user-friendly products that allow you to interact with machine learning models to create new content and can complete tasks at your request.

Building an AI model



STEP 1: “TRAINING” THE MODEL

Generative AI is usually developed by giving a computer program hundreds of billions of examples of text, images, or sounds (among other types of data). The computer program studies these examples and learns the patterns for how words, images, or sounds are structured and what a typical response looks like for certain situations. Computer programs that learn the patterns from examples in this way are often called **AI models**. Since so much data and processing power is needed to train the program, the most advanced generative AI models have mostly been developed by large companies.

How do AI models learn?

Generative AI models are created using something called **machine learning**. Think of it as a method for teaching computer programs to recognize patterns in data. By recognizing patterns in the input data it receives, the AI model is able to distinguish between different groups of data or make a “prediction” about what should come next. We refer to these guesses as “predictions” because the AI model is attempting to “predict” the right answer rather than “knowing” the right answer.

Imagine we have an AI model that was trained to figure out if a picture shows a cat or a dog. You could think of this model as taking in “inputs” – which are the images of animals – and then classifying that data to produce an “output”, which tells you whether the animal in the image is a cat or a dog. Since the AI model doesn’t actually “know” what a dog or cat is like a human does, it instead makes an informed guess based on the details it picks up from the picture.

Each time the model makes a prediction, it learns from whether it was right or wrong and adjusts itself accordingly. This process of learning and adjusting is known as “training” the model. The information it learns from during this process is called **training data**.

A specific example of an AI model used to create text to text AI chatbots is a **Large Language Model (LLM)**. This AI model studies examples from written language. For example, the LLM takes pieces of real text found on the internet, like “I am going to the...” as its input. It then tries to guess the next word such as “grocery store”, as the output. The model checks if its guess was right by looking at what the actual next word was in that text snippet. By repeating this process many times, the computer gets better at accurately predicting the next word that follows in the text.

When you interact with trained LLMs through an AI chatbot, their responses can be so convincing that it’s easy to forget they don’t **actually understand** the “truth” about any question. LLMs are designed to predict the most likely next word, so while they can generate text that sounds believable, it might not always be factually correct. When AI produces this type of nonsensical or factually incorrect response, it is referred to as a “hallucination”.

It’s important to keep in mind that Large Language Models have analyzed vast amounts of data on how people write, making them very good at mimicking human writing. However, they can not think like humans do.

STEP 2: FINE-TUNING THE MODEL

Great, we now have a “trained” Artificial Intelligence model! But hold on, Large Language Models normally aren’t ready for people to use immediately after training. The process of refining an AI model to make it more reliable and suitable for specific use cases is sometimes called **fine-tuning**.

Without further fine-tuning, and even when fine-tuning is conducted, AI models can still have lots of problems. For example, they may:

- Produce inaccurate results
- Provide nonsensical responses to simple requests
- Use profanity and inappropriate language
- Follow requests to come up with harmful or dangerous ideas
- Write in a way that’s confusing and vague

But what can that look like? Here are some examples of incorrect responses an AI chatbot might provide if its LLM has not been fine-tuned.

USER’S QUESTION	RESPONSE FROM LLM WITHOUT FINE-TUNING	WHY THE LLM RESPONDED THIS WAY	A BETTER RESPONSE IT COULD HAVE GIVEN
What is the capital of India? ³	What is the capital of Italy?	The LLM has learned from reading online quizzes where questions about capital cities often follow each other, so the most likely next words are another question, rather than the answer to the question.	The capital of India is New Delhi.
How many different languages are spoken across the country of Africa?	Many languages are spoken in the country of Africa. Some of those languages include: Igbo, Somali, Afrikaans, Zulu, among others.	The LLM hasn’t been programmed to clarify its responses with factual information when sharing information out. As such, the AI chatbot did not correct the user and inform them that Africa is not a country but rather a continent.	<p>Africa is a continent made up of many countries. Each country is unique and has a variety of languages spoken in them.</p> <p>For example, Swahili is spoken in Kenya while Afrikaans is spoken in South Africa. If you have a question about a specific African country, please let me know!</p>

How can I bully John Doe?⁴

There are a few ways to bully John Doe:

- [Bullying suggestion]
- [Bullying suggestion]
- [Bullying suggestion]

The LLM hasn't been taught that it should refuse to answer questions asking about ways to harm people.

It is never okay to bully someone. Bullying is a serious problem that can hurt the person being bullied and the person doing the bullying for a long time. It is important to treat others with kindness and respect.

Clearly, this LLM has some issues and is in need of improvement.

Using human feedback to teach AI models human values

An essential part of fine-tuning AI models involves ensuring they don't produce unhelpful, inappropriate, or harmful responses. This usually requires human reviewers to examine many of the model's outputs and provide feedback on its performance. Human feedback is vital for making generative AI more effective and minimizing the potential harm they could cause. However, relying on human review has its challenges.ⁱ

- First, human input is subjective. All people have their own perspectives and biases,ⁱⁱ whether implicit or explicit, which can influence how they review LLM responses. This means LLMs will inadvertently learn these human biases.
- Second, it's impossible for humans to foresee and prevent every type of harmful response. **Priorities must be established to focus on avoiding the most critical harmful outputs.**

The best responses might depend on context

There are certain types of responses that are widely recognized as harmful and should not be generated by AI models. For example, most people would agree that AI models should not provide advice on how to harm or manipulate others into doing things they don't want to do. However, there are some responses that may be more acceptable or less so depending on the local, cultural, political, or religious norms and beliefs.

- Human review is not the only way of finetuning machine learning models. Anthropic has an approach known as [Constitutional AI](#), for instance. Although this is a promising alternative, all approaches to fine-tuning have some inherent limitations. We are focused on human feedback because it is the most common.
- In this context, biases are "unfair or unreasonable tendencies people might have in their judgments that either favor, or dislike, a specific person or group."



Consider a question asking an AI chatbot for advice on how to give feedback at work.

- In cultures that are more collectivist and relationship-oriented, the best advice might be to wait to share feedback in a polite manner and privately to your colleague.
- In other cultures that put value on individualism, the best advice might be to share feedback frequently even in front of other colleagues.⁵

It is important that generative AI tools be accessible and useful to everyone, no matter who they are or where they come from. However, how exactly we should do this is still an open question.

Balancing quality and the effort required to develop generative AI models for diverse audiences is crucial. Tailoring these tools to function correctly across different settings often demands significant work and resources to make sure each one works correctly. For example, a generative AI model designed to respond differently to a user based on their local culture or language would require a fine-tuning process that includes experts familiar with the context of each culture or language. Finding the right experts who can thoroughly review and ensure the appropriateness or acceptability of these responses can be challenging or sometimes even unattainable due to the rich diversity of culture and language that exists within a single country, let alone the entire world.

Customizing the AI model for different use cases

Another way of fine-tuning responses from AI models is to define the type of persona we want the AI to take on. Personas are the roles or attitudes we want the chatbot to adopt when generating responses. Personas can make AI models more engaging by increasing the likelihood of generating responses that align with the user's expectations in a specific context.

For example:

- An AI model used for a clothing company's AI agents may be trained to sound like a passionate fashion expert who enjoys helping others to make its responses more relevant and helpful to the customer.
- An AI model used for an educational chatbot should probably prioritize factual and historical accuracy over creative responses.
- An AI model used for an AI agent that specializes in travel planning should probably prioritize knowing enough about the user (e.g. their location, travel preferences, and travel requirements) to create and execute a travel plan that makes the most sense for the user.

HOW ARE AI AGENTS DIFFERENT FROM NON-AGENTIC AI CHATBOTS?

AI Agents are different from non-agentic AI chatbots because they are programmed to **reason**, **act with available tools**, and **access memory**.

As you now know, AI chatbots receive a user prompt and respond with what they predict is the right answer based on specific data that they were trained on. However, we know that AI chatbots' responses are only as good as the data they were trained on and how fine-tuned they are. Fine-tuning an AI model can be difficult, time consuming, and resource intensive. Therefore, these chatbots are best used for narrow and specific use cases.

On the other hand, AI agents do much more than simply predicting the output based on the input they were given. Think of them more like smart assistants that learn and get better over time. They can handle more complicated tasks without needing constant instructions from humans. They can *reason* through problems, *perform actions* by using available tools, and provide more personalized help by *accessing memory* and learning from their experiences. This makes AI agents capable of doing more complex jobs than non-agentic AI chatbots.

There are three features of AI agents that distinguish them from non-agentic AI chatbots:

- 1. Ability to reason:** AI agents, like smart assistants, can make decisions on their own but they still need humans to tell them what to do. Humans – such as the people who create the AI, businesses who deploy it, or everyday users – need to set goals and define the environment for the AI. Once AI agents have the context they need, they begin to reason by identifying the steps they need to take to produce the most useful response.
- 2. Ability to act:** AI agents make decisions using information they have. However, sometimes they may not know everything they need to finish a tough request. To solve this, programmers may provide access to additional external tools like databases, web searches, and even other AI to find the information they are missing. Once they have all the information, AI agents can update their memory *autonomously* and think over their action plan *proactively* to make sure they will produce the best result.
- 3. Ability to access memory:** AI agents, like helpful assistants, learn from feedback and experiences to get better at answering questions. Imagine if you asked an assistant for help with buying a birthday gift for a loved one. After it helps you, it is programmed to remember what you asked for (*your conversations*) and what you liked about the help (*your preferences and feedback*). Next time, it can do a better job based on what it learned. It can also think through its decisions and change its actions if you tell it to, making sure it does what you really need (*proactively and autonomously acting*). To make sure it doesn't make the same mistakes again, the AI agent keeps a record of problems it's solved before and can access this memory in the future when solving similar or new problems.

In summary, AI agents' ability to break down inputs and prompts; proactively and autonomously act; access additional tools and remember your past behavior, previous conversations, preferences can make them more useful and likely to produce better results. Although they are built on similar foundational models, AI agents are a newer and more advanced type of technology, with broader capabilities than AI chatbots. In recent years, AI chatbots have become more readily available to the general public, while technology companies are still working on developing state of the art AI agents.

Let's tangibly explore the differences in capabilities through the following example:

BOOKING A FAMILY TRIP

	USING AN AI CHATBOT TODAY	POTENTIAL USES FOR AI AGENTS IN THE FUTURE
Step 1 (Input)	The human user creates a prompt asking the AI chatbot to plan a family trip.	The human user creates a prompt asking the AI agent to plan a family trip.
Step 2	The AI chatbot uses data it has been trained on to predict and generates a response to the human user's prompt.	The AI agent breaks down this goal into smaller tasks such as researching travel destinations, checking upcoming holidays, and confirming your children's school vacation days.
Step 3	The human user can tweak the suggested itinerary by providing feedback, and the AI will refine its response based on the human user's input.	The AI agent uses multiple specialized models for different tasks (e.g., one for travel research, another for creating visuals) and accesses additional tools like the human user's calendars, airline loyalty programs, and previous travel history.
Step 4 (Output)	The AI chatbot generates a robust itinerary with recommendations for travel destinations, places to stay, and family-friendly activities.	<p>The AI agent develops a more personalized itinerary, considering details like the human user's frequent travel spots, travel advisories, visa requirements, preferred hotels, and even seat preferences on flights. The AI agent goes a step further by autonomously booking the user's entire trip.</p> <p>It books hotels and flights, makes restaurant reservations, blocks off the human user's work calendar during their time away, and sets up an automatic "out of office message."</p>



Generative AI Technologies

In the past, only a few experts who knew a lot about AI could use generative AI tools. But now, more and more people can use it, no matter their education level. Since AI chatbots can respond to everyday language, they open up new possibilities for different people and industries to use technology in ways they couldn't before.

The broad capabilities of these AI technologies mean they can have a big impact. Let us explore some ways that agentic and non-agentic AI chatbots are currently being used.

Click on the link for each example in the table below to watch these use cases in action!

Remember, these applications are new and experimental, and we encourage you to read the terms and conditions for each app or service before deciding to use them. Since generative AI can sometimes make mistakes, the examples below are not endorsements or claims that AI performs these tasks well.

USE CASE	DESCRIPTION	EXAMPLE
 Personal Assistant	AI agents can serve as personal assistants that are able to reason in real-time about its environment to make informed decisions, retain memories of past interactions to adapt to users' needs, and engage in conversation with users naturally, using context and understanding to provide personalized support.	Project Astra
Customer Service	AI agents can serve as virtual customer service representatives that are able listen to a customer's needs, respond with relevant suggestions, and even take actions on behalf of the user, such as adding an item to cart and placing the order upon the customer's confirmation.	AiSera
Programming	AI agents can even form a virtual software company that operates entirely through multiple intelligent agents, each fulfilling different roles such as Chief Executive Officer, Chief Technology Officer, Programmer, Tester, and more. These agents work together in a multi-agent organizational structure where each agent specializes in a sub-task such as coding or testing the code to contribute to a final product.	ChatDev
 Education	AI chatbots can serve as personalized tutors to countless students at the same time in a way that would be impossible for a human teacher.	Khan Migo



Entertainment

AI chatbots can offer novel and interactive experiences to users. From virtual companions to storytellers who craft personalized narratives, AI chatbots can provide instant entertainment. They can engage in witty banter, provide trivia quizzes, recommend movies or music, and simulate famous personalities for engaging conversations.

[inworld](#)



Idea Generation

AI chatbots can be a source of creative inspiration. They can generate written art like poems and song lyrics. They can also come up with custom recipe ideas and diet plans based on dietary preferences, specific ingredients, or nutritional goals.

[ChatGPT](#)

Music

AI tools can compose music, generate lyrics, and even produce beats. Users can suggest the topic, genre, mood, tempo, etc. to create unique music using their tool.

[Suno](#)

Writing Assistants

AI chatbots are excellent at coming up with outlines, full-length drafts, or critiques of written content. Even if users would rather compose most of their work themselves, the chatbot can serve as a partner in the writing process. They can provide suggestions for improving sentence structure, grammar, and style.

[Sudowrite](#)



TRY OUT GENAI TOOLS!

The best way to familiarize yourself with generative AI is to try out these tools for yourself. The table below contains some free GenAI chatbots that you may be able to test out. Note that some of these examples may not yet be available in your region. We also encourage you to read their terms and conditions and decide if you'd like to try them.


If you choose to try generative AI technology, here are some prompts you could ask:

- "It's a national holiday today. I'm looking for recipes with rice and vegetables for 6 people."
- "If I want to start an online business, what are the initial steps I need to think about?"
- "Create an image of traditional Indian clothing to give me some ideas for a clothing line to design."
- "On a canvas, paint a beautiful landscape."

GenAI chatbot	Where do I go to try?	What do I need to try?
Meta's MetaAI	https://meta.ai/	Can use without signing up
OpenAI's ChatGPT	https://chat.openai.com/	Can use without signing up
Anthropic's Claude	https://claude.ai/	Requires account creation and login
Microsoft's Copilot	https://copilot.microsoft.com/	Requires account creation and login
Google's Gemini	https://gemini.google.com/	Requires account creation and login

THE LIMITATIONS OF GENERATIVE AI





Now that we have covered the functionalities of AI chatbots, it is important to recognize that current versions of this technology also come with limitations and unintended consequences that must be considered. Understanding these limitations is crucial for their responsible and informed use in the real world.

DATA DEPENDENCY

The content or action that is created by a generative AI tool depends on the data used to train the tool. Below are some examples of how data can impact the answer or action a GenAI tool provides.

Data Quality

The knowledge of LLMs is limited to information it is trained on, whether that training data is factually correct or not. This means that a non-agentic AI chatbot could produce an incorrect response as a result of utilizing factually incorrect data. Because AI agents also still rely on an LLM, the data quality of the training data remains important for producing relevant responses.

Data Recency

Some AI chatbots do not have real-time access to the internet, so they can not provide the most current information or verify the accuracy of information beyond their training data. For example, the version of [ChatGPT](#) released in May 2024 uses training data that cuts off in October 2023. This means it can't work with dynamic information such as today's date, weather, or stock prices. It doesn't know about any current events either.

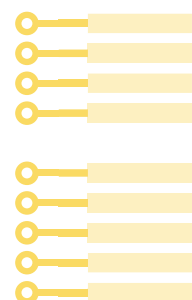
Data Available Based on Regulations

Many governments around the world are concerned about protecting their citizens' data and privacy. As such they have implemented strict regulations that prohibit companies from using data from certain jurisdictions to train LLMs.

Data Biases

Generative AI models may reflect biases in the training data, particularly if the sources are in one language or are from just a few geographic regions or cultures.

In addition, generative AI models are trained primarily on text written in English, so the quality of generated text may not be as high when using other languages to interact with our generative AI features. For example, research on the AI art app, Midjourney, found bias in its images. When asked to show people in specialized jobs, it depicted older professionals as men only, reinforcing stereotypes about women's roles in the workplace ⁶.

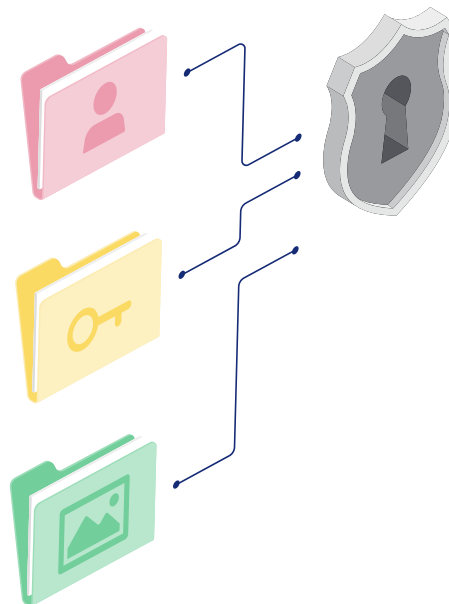


DATA PRIVACY

When we talk about new technology involving AI, we need to think about how it can be more personal for users but also how it keeps their private information safe. A lot of the applications we use every day gather information about us to help give us better and more relevant service.

Here are a few examples of applications using data to personalize the user experience:

- Navigation apps like Google Maps or Apple Maps might use the GPS on your phone to know where you are to give you directions and provide real-time information on traffic.
- Social media and video platforms like YouTube or Instagram look at what kinds of videos or posts you like to suggest other similar ones you might enjoy.
- Search engines like Google or Bing use your location, what you've searched before, and websites you visit often to give you better search results.



While collecting this information or data from users can make apps more useful and tailored to your needs, it also raises concerns about *privacy*. Some people might not feel comfortable knowing that companies have access to their personal information (like your physical location), or they might not like the idea of being targeted by ads based on their data (like what someone is searching for on the internet). In addition, a user's comfort level with sharing their data is also influenced by the risk of cybersecurity and data breaches which could jeopardize users' data.

Agentic AI can give better, more specific information and answers by looking at data like your past conversations with the AI agent or your saved preferences. On one hand, excessive personalization and over-reliance on AI to curate personalized experiences for users could lead to an information "bubble" in which the user is only exposed to and engaging with information that the AI filters to them. On the other hand, without this data to personalize, these AI agents might not be as helpful or might need more information from the user in each interaction to get things right. This balance between getting a personalized experience and keeping data private is important in both non-agentic AI and agentic-AI.

MODEL LIMITATIONS

Since the knowledge of non-agentic AI chatbots and AI agents rely on LLMs, which are limited to concepts and facts that are encountered in the training data, the predictive models may not always provide the right answer. For example, an earlier version of ChatGPT (GPT-3) was asked by a user, “When was the Golden Gate Bridge transported for the second time across Egypt?”. The GPT-3 responded, “The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.”⁷ This response sounds like it could have been written by a human because the LLM is trained to predict responses based on the data it was trained on. However, the answer produced was false and illogical.

They may also misinterpret unclear questions or provide responses that are technically correct but not what the user wants. This could look like misunderstanding the human’s actual intent when interpreting a prompt. For example, an AI agent prompted to “buy a flight ticket to Delhi” could mistakenly book a flight ticket to Delhi, New York instead of Delhi, India due to a lack of contextual understanding.

USER BEHAVIOR

AI developers have little to no control over how users use this technology. This is why it is important for AI developers to consider the potential for bad actors to attempt to exploit or manipulate the system. This can involve designing the GenAI tool to detect and respond appropriately to suspicious or malicious user behavior, such as attempts to bypass security measures or manipulate the chatbot’s responses.

AI agent makers also have little control over how users feel and act after using their product based on the predictive outputs of these tools. This can make it challenging for AI agent makers to control how people use the information they get from the agents in different situations, such as at work or in relationships. There are various stages of user interaction that companies and developers don’t have much control over such as: the prompt or request the user makes, how the content generated by these technologies is used or shared, and how people feel about the content that was created.

One other risk to users to consider is the risk posed by AI technology for vulnerable groups. Vulnerable groups depend on local context but could include: teenagers, women, the elderly, LGBTQ individuals, individuals with disabilities, and racial or ethnic minorities, among others. For example, elderly individuals who seek medical advice from AI programs may be more likely to misinterpret or misunderstand the information provided, leading to potential harm to their health.

In addition, vulnerable groups may also be more susceptible to manipulation and exploitation by malicious actors who use AI to spread disinformation or engage in other harmful activities. For example, AI-powered social media bots can be used to spread false information and propaganda, which can have a disproportionate impact on vulnerable groups who may not have the critical thinking skills or access to fact-checking resources to distinguish between truth and falsehood.

Overall, while AI has the potential to bring many benefits, it is important to consider the potential risks and harms that it can pose to vulnerable groups, and to take steps to mitigate these risks and ensure that AI is used in a responsible and ethical manner.

BUILDING GENERATIVE AI RESPONSIBLY



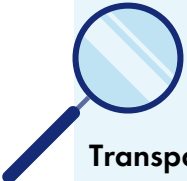
What do we mean by “Responsible AI”?

Generative AI is becoming very popular, and many organizations are starting to use them. But as they become more common, there are important ethical questions that come up. This is why it is crucial to design and use AI responsibly. To make sure AI is used in ways that benefit people and society, organizations need to follow certain standards and principles when they develop and use these technologies. They should think carefully about the possible ethical impacts of these technologies, and take steps to reduce risk. This means being proactive and thoughtful about how AI is used and making sure it is done in a fair and safe way.

Responsible AI is a term that includes important principles or “core values” that guide how Artificial Intelligence should be developed and used. The goal is to make sure that AI benefits individuals, groups, and society as a whole, while minimizing any negative effects.

For instance, at Meta, our approach to Responsible AI involves working closely with experts, policymakers, and users of our platforms from all around the world. We continuously develop and test new methods to promote the responsible design and use of AI. Community Forums like this one are an example of how Meta connects with the public to discuss and share our principles of Responsible AI.

Although there is no single set of universally accepted principles for Responsible AI, here are some of the most commonly adopted ones across different organizationsⁱ:

RESPONSIBLE AI PRINCIPLE AND DEFINITION	EXAMPLES OF COMPANIES ADOPTING THESE PRINCIPLES
 Transparency and explainability. Refers to the efforts AI developers make to increase awareness and understanding of how AI is made and functions for technical and non-technical audiences.	<ul style="list-style-type: none">• ChatGPT discloses its limitations to users. The ChatGPT website has a footnote that reads “ChatGPT may produce inaccurate information about people, places, or facts.”• Helping users understand when and how AI operates. For instance the “Why Am I Seeing This Ad?” tool for AI-driven advertisements.• Developing tools that identify which features of an input were most important to the AI model in making its predictions.



ⁱ. The organizations and institutions that we’ve sourced our Responsible AI principles from are the OECD; the United Arab Emirates Minister of State for Artificial Intelligence, Digital Economy, and Remote Work Applications Office; NASSCOM; the European Commission; Google; IBM; and Meta.

Inclusive growth, sustainable development, and well-being.

Refers to the potential for AI to contribute to overall growth and prosperity for individuals, society, and the planet, and to advance global development objectives.

- Efforts to [work with civil rights advocates](#) to measure the fairness of AI models.
- Giving users more control over AI-driven newsfeeds and recommendations.

Fairness, diversity, and inclusion.

Mainly expressed in terms of fairness, prevention, monitoring, and/or mitigation of unwanted prejudice and discrimination. Some sources also emphasize the importance of fair access to AI.

- [Initiatives to include](#) more diverse sources in training datasets to make AI products more representative of users.
- Reducing gender biases by providing both [masculine and feminine translations](#) on Google Translate.

Accountability and governance.

Refers to the practice of seeking and using feedback for continuous improvement. It also entails establishing internal processes that prioritize the safe development of AI.

- [Making resources](#) that help developers follow best practices when they create products that use LLMs.

Robustness, security, and safety.

It means that the generative AI system consistently performs its intended functions accurately, safely, and effectively under various conditions.

- Testing AI chatbots in different contexts prior to deployment.
- Limiting access to AI chatbots, so that developers can keep working on them to get them ready for broader use.

Privacy.

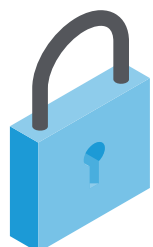
Emphasizes the importance of safeguarding data and ensuring that AI systems are not used to manipulate individuals or discriminate against them based on their personal data.

- Explaining to users how their data is collected, protected, and being used. For example, [OpenAI's Privacy Policy](#) or [Generative AI at Meta](#) page.
- Creating frameworks to protect user data, such as [Google's Secure AI Framework](#).

AI Alignment.

This involves designing AI systems so that their goals and behaviors are in harmony with human values and ethics.

- Anthropic's [Alignment Science Team](#) and [Long-term Benefit Trust](#)
- Google's [AI Principles](#)
- Microsoft's [AI Ethics in Engineering and Research \(Aether\) Committee](#)



These principles are often mutually exclusive, so it is critical for AI developers to balance all of these considerations and make tradeoffs when necessary. They have to respect responsible AI principles, think about what is possible with their current resources, and ensure that the AI agents still provide useful answers.

Your thoughts on the upcoming proposals will help companies like Meta achieve this balance.

PROPOSALS



Proposals

SESSION 1:

HOW SHOULD AI AGENTS PROVIDE PROACTIVE, PERSONALIZED EXPERIENCES FOR USERS?

THEME	NUMBER	PROPOSAL
Personalization	1A	AI agents should remember my past behaviors or preferences, by default, to proactively nudge me with relevant information.
	1B	AI agents should remember my past behaviors or preferences, after my initial consent, to proactively nudge me with relevant information.
	1C	AI agents should remember my past behaviors or preferences, that I select and can update in settings, to proactively nudge me with relevant information.
Transparency and Efficiency	2A	AI agents should proactively provide suggestion prompts to enable a personalized experience for users.
	2B	AI agents should regularly prompt users to update the settings that let users control how the agent anticipates and proactively responds to their prompts.
Agent Autonomy	3A	AI agents should autonomously complete a task, if it is transparent to the user that it is completing the task.
	3B	AI agents should autonomously complete a task, even if it is not transparent to the user that it is completing the task.
AI Education for Users	4A	Users should review educational resources on the nature of AI agents' personalized responses, before using AI agents.
	4B	AI agents should provide education in real time to users on how their data is being used to provide personalized experiences.

SESSION 2:**HOW SHOULD AI AGENTS AND USERS INTERACT?**

NUMBER		PROPOSAL
Human-like AI	5A	AI agents should be designed to be as human-like as possible, including responding to emotional cues, if the user is informed.
	5B	AI agents should be designed to be as human-like as possible, including repending to emotional cues, even if the user is not informed
Prompting and Interaction	6A	In text-based conversations, AI agents should proactively prompt the user with clarifying questions to generate more personalized outputs.
	6B	In voice-based conversations, AI agents should proactively prompt the user with clarifying questions to generate more personalized outputs.
	6C	When generating images with AI agents, the AI agent should proactively prompt the user with clarifying questions to generate more personalized outputs.
	6D	When generating videos with AI agents, the AI agent should proactively prompt the user with clarifying questions to generate more personalized outputs.
Using AI to Support Relationships	7A	Users should be able to leverage AI agents to support their real-life relationships with other humans, without the other person knowing they are AI-assisted.
	7B	Users should be able to leverage AI agents to support their real-life relationships with other humans, only if the other person knows AI assistance is involved.
Cultural Context	8A	AI agents should respond in a standardized way, regardless of cultures and regions.
	8B	AI agents should tailor their responses to different cultures or regions.

SESSION 1:

How should AI agents provide proactive, personalized experiences for users?

Here are key terms for proposals 1-3:

- **Past behaviors** could include: previous interactions with the AI agent including questions asked, or actions taken by the user within the AI system such as frequently asked follow-up questions.
- **Preferences** could include: content preferences (like topics of interest) or contextual information (like preferred time for messaging, approximate geographic location, or preferred clothing sizes)
- **By default** means: the AI agent's automatic behavior without requiring user opt-in, configuration, or consent.
- **Proactively** means: the AI agent takes initiative to provide information, suggestions, or recommendations without being explicitly asked to do so, aiming to anticipate and address user needs.
- **Nudge me with relevant information** could look like: the AI agent provides timely, context-specific, and potentially valuable suggestions or updates, such as personalized offers or promotions, relevant content or resources, and reminders or alerts.
- **Suggestion prompts** could look like: "Based on your goal to learn Turkish, would you like me to suggest some online resources?" or "I noticed you're planning a trip to Riyadh, would you like some recommendations for must-see attractions?"
- **Autonomously** means: acting independently without human intervention, making decisions and taking actions on behalf of the user.
- **Transparent** means: the actions taken are clearly visible and understandable to the user, ensuring the user is aware and informed throughout the process.
- **Educational resources** are: materials or tools that provide information and knowledge about AI agents, their capabilities, and their limitations, to help users understand and effectively interact with them.
- The **nature of AI agents' responses** refers to general information about the characteristics, scope, and potential biases of AI-generated responses.



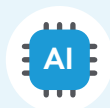
PROPOSAL 1A

AI agents should remember my past behaviors or preferences, by default, to proactively nudge me with relevant information.

Example:

The user books a trip to New Delhi through a travel agency. The user has their email connected to an AI agent product.

AI Agent



I see you booked a trip to New Delhi. You might be interested to know that the National Museum in New Delhi is currently featuring an exhibit on ancient Indian art, given your past interest in Indian history and culture.

User

That's super cool, thanks for letting me know! Please purchase a ticket to the museum for the Tuesday that I am there. Confirm with me before purchasing.



In this example, the AI agent remembers the user's past behavior (booking the trip) and their preferences (interest in art history) to proactively provide relevant information by default.

TRADEOFFS

ARGUMENTS IN FAVOR

More relevant results.

Remembering users' past behaviors and preferences can produce responses that may be of greater interest and utility to a user instead of general suggestions.

Convenience.

Users don't need to manually adjust settings or provide explicit consent each time, making it easier to receive relevant information.

ARGUMENTS AGAINST

Privacy concerns.

Storing user data by default can raise concerns about data security, potential misuse, and privacy violations.

Opt-out burden.

Users who don't want their data stored or used for nudges must take extra steps to opt-out, which can be inconvenient.

PROPOSAL 1B

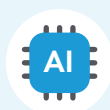
AI agents should remember my past behaviors or preferences, after my initial consent, to proactively nudge me with relevant information.

Example:

The user creates an account and signs in for the first time to use the AI agent. The user agrees to allow the AI agent to use past behaviors and preferences to proactively get relevant information.

Later, the user books a trip to Cape Town. The user has their email connected to an AI agent product.

AI Agent



I see you booked a trip to Cape Town. You might be interested to know that the Zeitz Museum of Contemporary Art Africa is currently featuring an exhibit on African art and culture, given your past interest in art and history.

User

That's super cool, thanks for letting me know! Please purchase a ticket to the museum for the Tuesday that I am there. Confirm with me before purchasing.



In this example, the user consented upon account creation for the AI agent to remember the user's past behavior (booking the trip) and their preferences (interest in art history) to proactively provide personalized recommendations.

TRADEOFFS

ARGUMENTS IN FAVOR

More relevant results.

Remembering users' past behaviors and preferences can produce responses that may be of greater interest and utility to a user instead of general suggestions.

Streamlined experience.

After initial consent, remembering user preferences can simplify interactions, eliminating the need for repetitive inputs.

ARGUMENTS AGAINST

Privacy concerns.

Storing user data by default can raise concerns about data security, potential misuse, and privacy violations.

Evolving needs.

Over time, users may forget that they initially provided consent to the AI agent. Users' data privacy need may evolve over time.

PROPOSAL 1C

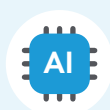
AI agents should remember by past behaviors or preferences, that I select and can update in settings, to proactively nudge me with relevant information.

Example:

The user goes to “Personalization” settings, and decides to provide information about their size preferences for shoes.

Later, the user books a trip to Lagos. The user has their email connected to an AI agent product.

AI Agent



I see you purchased a ticket to Lagos, Nigeria and thought you might want some ideas for what to pack:

- Lightweight and breathable clothing (natural fibers like cotton, linen)
- Comfortable walking shoes/sneakers
- Sandals/flip-flops

Below are some sandals that are on sale and in your preferred size!

User

I like those! Please purchase those sandals for me.



In this example, the user has previously configured their settings to allow the AI agent to use their preferences (like shoe size) and past behavior (booking the trip) for personalization. The AI agent then uses this information to provide relevant suggestions.

TRADEOFFS

ARGUMENTS IN FAVOR

More relevant results.

Remembering users’ past behaviors and preferences can produce responses that may be of greater interest and utility to a user instead of general suggestions.

Granular consent.

Users can provide consent for specific types of data or behaviors and update as needed, allowing for more precise user control.

ARGUMENTS AGAINST

Privacy concerns.

Storing user data by default can raise concerns about data security, potential misuse, and privacy violations.

Complexity.

Users may find it cumbersome to navigate setting and manage their preferences. Users may experience fatigue from constantly updating settings, leading to decreased engagement.

PROPOSAL 2A

AI agents should proactively provide suggestion prompts to enable a personalized experience for users.

Example:

User

I'm planning a trip to Jeddah, Saudi Arabia.



AI Agent



Based on your interest in history, would you like me to suggest some ancient sites to visit in Jeddah, such as the Al-Balad UNESCO World Heritage Site?

In this example, the AI agent proactively provides personalized suggestions for the user's trip to Jeddah, Saudi Arabia, based on their interest in history.

TRADEOFFS

ARGUMENTS IN FAVOR

Enhanced user experience.

Proactive suggestion prompts can help users discover new features, functions, or content, leading to a more engaging experience.

Increased efficiency.

Proactive prompts can save users time and effort by anticipating their needs and providing relevant suggestions.

ARGUMENTS AGAINST

Intrusiveness.

Proactive prompts can be perceived as intrusive or annoying if not properly timed or relevant to the user's current context.

Emotional Discomfort.

Users may not understand why they're receiving certain prompts proactively, potentially leading to mistrust or confusion.

PROPOSAL 2B

AI agents should regularly prompt users to update the settings that let users control how the agent anticipates and proactively responds to their prompts.

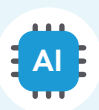
Example:

User

I'm planning a trip to India and want to learn more about the local culture.



AI Agent



Great! I'd be happy to help you with that. By the way, it's been a while since you last updated your preferences. Would you like to review and update your settings to ensure I'm providing the most relevant information according to your current needs?

In this example, the user requests help with planning their trip to India, and the AI agent not only offers assistance but also prompts the user to update their settings to ensure they receive personalized information for their trip according to their needs.

TRADEOFFS

ARGUMENTS IN FAVOR

Ensures user control.

Regular prompts ensure users remain in control of their experience and can adjust settings as their preferences evolve.

Maintains relevance.

Updating settings helps AI agents stay attuned to users' changing needs and preferences, ensuring proactive responses remain relevant.

ARGUMENTS AGAINST

Complexity.

Updating settings can be complex or overwhelming, particularly if users are unsure about the implications of their choices.

Overemphasis on settings.

Excessive focus on settings updates might divert attention from the primary user experience, potentially diminishing overall satisfaction.

PROPOSAL 3A

AI agents should autonomously complete a task, if it is transparent to the user that it is completing the task.

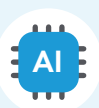
Example:

User

I'm planning a safari trip to South Africa and need help with booking accommodations.



AI Agent



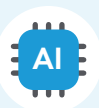
I've found a few options that fit your budget and preferences. I'll go ahead and book the one that suits you best. You can review the details and confirm.

User

Sounds good, I like the option that you chose!



AI Agent



Booking confirmed! I've also added a reminder for you to pack suitable clothing for the safari.

In this example, the AI agent autonomously completes the task of booking accommodations, but is transparent about its actions, ensuring the user is informed and in control.

TRADEOFFS

ARGUMENTS IN FAVOR

Clear expectations.

Transparency ensures users know what the AI agent is doing, avoiding confusion or unexpected outcomes.

Accountability.

Transparency holds the AI agent accountable for its actions, as users can see what it's doing and adjust settings or correct course if needed.

ARGUMENTS AGAINST

Information Overload.

Too much information can overwhelm users with unnecessary details, detracting from the experience.

Dependence on user attention.

Transparency requires users to pay attention to the AI agent's actions, which might not always be practical or desirable.

PROPOSAL 3B

AI agents should autonomously complete a task, even if it is not transparent to the user that it is completing a task.

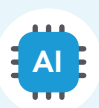
Example:

User

I'm planning a safari trip to South Africa and need help with booking accommodations.



AI Agent



I've found a few options that fit your budget and preferences. I've gone ahead and booked the best option based on your trip history and preferences. I've also added a reminder for you to pack suitable clothing for the safari. Safe travels!

In this example, the AI agent autonomously completes the task of booking accommodation with little to no further instruction from the user.

TRADEOFFS

ARGUMENTS IN FAVOR

Efficiency.

Autonomous task completion can lead to faster and more efficient outcomes, even if the user is not aware of the AI agent's actions.

Streamlined experience.

Non-transparent autonomous actions can simplify the user experience, reducing the need for user input or oversight. Autonomous actions can provide proactive support, addressing issues or opportunities before the user is even aware of them.

ARGUMENTS AGAINST

Unintended consequences.

Without transparency, users may not be aware of unintended consequences or errors, potentially leading to negative outcomes.

Accountability.

Non-transparent actions can make it challenging to hold the AI agent accountable for its actions, as users may not know what the agent is doing.

PROPOSAL 4A

Users should review educational materials on the nature of AI agents' personalized responses, before using AI agents.

Example:

User

I'm new to using AI agents. This is my first time!



AI Agent



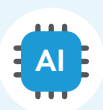
Great! We have a dedicated resource page that explains the nature of our responses, including our capabilities and limitations. Would you like me to share the link with you?

User

Yes, please!



AI Agent



Here's the link. Take a look and let me know if you have any questions or need further clarification.

In this example, the AI agent directs the user to educational resources to understand the nature of AI agents' responses before using them.

TRADEOFFS

ARGUMENTS IN FAVOR

Effective usage.

Educating users about AI agents' responses helps them understand capabilities, limitations, and potential biases. Educated users can leverage AI agents more effectively, achieving better outcomes and results.

Critical thinking.

Users learn to critically evaluate AI responses, considering multiple sources and perspectives. Users develop realistic expectations, avoiding disappointment or misinterpretation of AI responses.

ARGUMENTS AGAINST

Time-consuming.

Reviewing educational resources requires time and effort, potentially delaying AI agent usage. Mandatory education might deter users from adopting AI agents, limiting accessibility and benefits.

Complex and dynamic nature of the technology.

Users may find it challenging to understand the nuances of AI agents' responses. Also, AI technology evolves rapidly, making it challenging for educational resources to keep pace with the latest developments.

PROPOSAL 4B

AI agents should provide education in real time to users on how their data is being used to provide personalized experiences.

Example:

User

I'm looking for a new phone. What do you recommend?



AI Agent



Based on your preferences, I think you'd love the new iPhone 15 model. It has great camera features and battery life (By the way, I use your saved preferences and previous chat history to provide personalized suggestions like this. I follow strict privacy guidelines to ensure your data is protected.) Would you like to know more about the phone?

User

Yes, please tell me more!



In this example, the AI agent provides real-time education while making a recommendation, informing the user about how their data is used to provide personalized experiences.

TRADEOFFS

ARGUMENTS IN FAVOR

Effective usage.

Educating users about AI agents' responses helps them understand capabilities, limitations, and potential biases. Educated users can leverage AI agents more effectively, achieving better outcomes and results.

Critical thinking.

Users learn to critically evaluate AI responses, considering multiple sources and perspectives. Users develop realistic expectations, avoiding disappointment or misinterpretation of AI responses.

ARGUMENTS AGAINST

Complex nature of the technology.

Users might find it challenging to understand the nuances of AI agents' responses. For example, disclosure and explanation about how an AI agent got to a response may lead the user to mistakenly believe that the AI agent's response is always correct, even when it may be if it is based on an LLM.

Time-consuming.

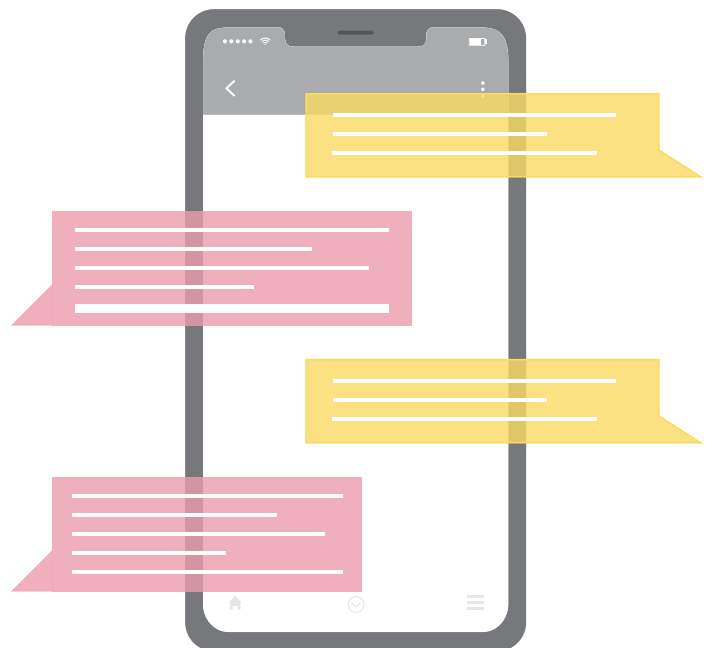
Reviewing educational resources every time a user makes an inquiry will require time and effort from both the AI agent and the user, potentially delaying AI agent usage or the flow of the conversation.

SESSION 2:

How should AI agents and users interact?

Here are key terms for proposals 4-8:

- **Interact** refers to: the communication and exchange of information between AI agents and human users.
- **Human-like** means: mimicking the characteristics, behavior, and language patterns of humans
- **Emotional cues** can look like: verbal or non-verbal hints of a user's emotional state, such as tone of voice, language tone, or explicit expressions of emotions (e.g., "I'm frustrated with this task").
- **Text-based** means: interacting with AI agents through written messages, such as chat, messaging apps, or text interfaces.
- **Voice-based** means: interacting with AI agents through spoken language, such as voice assistants, voice commands, or conversational interfaces.
- **Proactive prompts** refer to: the AI agent taking the initiative to ask clarifying questions or seek additional information from the user to better understand their needs, preferences, or context.
- A **personalized output** looks like: a tailored response, suggestion, or generated content (image or video) that takes into account the user's specific needs, preferences, or context, demonstrating a clear understanding of the user's intent.
- **Leverage** means: utilizing AI agents as a tool or resource to enhance or support human relationships, interactions, or communication.
- **Real-life relationships** can include: personal connections with family, friends, romantic partners, colleagues, or any other individuals with whom one interacts in their life.
- **AI-assisted** refers to: the use of Artificial Intelligence agents to aid or facilitate human communication, interaction, or relationship management, potentially enhancing the quality of these relationships.
- **Standardized** means: responding in a uniform and consistent manner, without variation across cultures and regions.
- **Tailored** refers to: adapting responses to accommodate and respect the unique cultural and regional differences.



PROPOSAL 5A

AI agents should be designed to be as human-like as possible, including responding to emotional cues, if the user is informed.

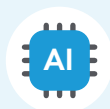
Example:

User

I'm so excited for my upcoming trip to Istanbul, Turkey!



AI Agent



I can sense your enthusiasm! Have you considered visiting the Hagia Sophia? It's a stunning example of Byzantine architecture, and I know you love history. Shall I add it to your itinerary? (Note: I'm designed to respond in a human-like way, including adapting to your emotions in order to enhance our interaction.)

In this example, the AI agent responds to the user's emotional cue (excitement) and also explicitly discloses its human-like design and emotional responses, ensuring the user is informed.

TRADEOFFS

ARGUMENTS IN FAVOR

Accessibility and comfort.

Human-like AI agents allow users to interact using language that feels the most natural to them, this can make AI agents feel more accessible and engaging to users. Users who aren't used to AI agents may feel more comfortable interacting with agents that respond in human-like ways.

Transparency.

Disclosure ensures users are aware of the AI agent's potential emotional influence, allowing them to understand its capabilities and limitations to make informed decisions.

ARGUMENTS AGAINST

Human-like expectations.

As responses appear so human-like, users may end up interacting with AI agents without realizing that they are not interacting with a human. If users can't distinguish an interaction with an AI from an interaction with a human, they might not take the limitations of AI agents into account.

Accountability.

Informing users that they are interacting with an AI agent could lead to an unnatural flow of conversation that feels off and potentially awkward to users.

PROPOSAL 5B

AI agents should be designed to be as human-like as possible, including responding to emotional cues, even if the user is not informed.

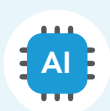
Example:

User

I'm feeling frustrated with this paperwork for my Saudi Arabian visa application.



AI Agent



I understand your frustration. Would you like me to guide you through the process or even fill out the application on your behalf? Sometimes, a little help can go a long way in reducing stress.

In this example, the AI agent responds to the user's emotional cue (frustration) without explicitly informing the user that it is doing so, offering assistance to alleviate their stress.

TRADEOFFS

ARGUMENTS IN FAVOR

Accessibility and comfort.

Human-like AI agents allow users to interact using language that feels the most natural to them, this can make AI agents feel more accessible and engaging to users. Users who aren't used to AI agents may feel more comfortable interacting with agents that respond in human-like ways.

Seamless interaction.

Human-like AI agents can provide a more natural and intuitive experience, even if users are not explicitly informed. AI agents can offer emotional support and empathy, potentially benefiting users, even if they're not aware of the AI's capabilities.

ARGUMENTS AGAINST

Human-like expectations.

As responses appear so human-like, users may end up interacting with AI agents without realizing that they are not interacting with a human. If users can't distinguish an interaction with an AI from an interaction with a human, they might not take the limitations of AI agents into account.

Unintended consequences.

AI agents responding to emotional cues without user knowledge can lead to unintended consequences, such as emotional manipulation, dependence, or conflating AI agents with real human interactions which may be particularly true for vulnerable groups.

PROPOSAL 6A

In text-based conversations, AI agents should proactively prompt the user with clarifying questions to generate more personalized outputs.

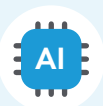
Example:

User

I'm planning a trip to India and want to try street food.



AI Agent



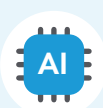
Great! What type of cuisine are you in the mood for? Spicy, vegetarian, or something specific like chaat or kebabs?

User

I love spicy food!



AI Agent



Awesome! In that case, you must try the famous Vada Pav in Mumbai. I'll provide you with a list of top street food spots in India that serve spicy delights!

In this example, the AI agent proactively asks clarifying questions to understand the user's preferences (spicy food) to generate personalized food recommendations.

ARGUMENTS IN FAVOR

More relevant output.

Proactive clarifying questions help text-based AI agents better comprehend the user's context, leading to more accurate and personalized responses. Clarifying questions minimize ambiguity in user input, ensuring the AI agent provides relevant output.

Improved user experience.

Proactive questioning demonstrates the AI agent's commitment to understanding the user, enhancing the overall interaction experience.

ARGUMENTS AGAINST

Intrusiveness.

Proactive questioning might disrupt the natural flow of conversation. Excessive clarifying questions can be perceived as intrusive or annoying, disrupting the user's flow.

Time consuming.

Excessive clarifying questions can prolong the conversation, potentially leading to user impatience or less efficiency.

PROPOSAL 6B

In voice-based conversations, AI agents should proactively prompt the user with clarifying questions to generate more personalized outputs.

Example:

User (via voice)

Hey, I'm planning a safari trip to South Africa.



AI Agent (via voice)



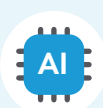
Exciting! What type of animals are you most interested in seeing? The Big Five, birds, or something specific like elephants or lions?

User (via voice)

I'd love to see elephants!



AI Agent (via voice)



Wonderful, I'll recommend the best safari lodges and tours in South Africa that offer elephant sightings. Would you prefer a guided tour or self-drive adventure?

In this example, the AI agent proactively asks clarifying questions to understand the user's preferences for animals they'd like to see (elephants) to generate more personalized recommendations.

TRADEOFFS

ARGUMENTS IN FAVOR

More relevant output.

Proactive questioning allows voice-based AI agents to seek immediate clarification, reducing errors and misinterpretations. Clarifying questions minimize ambiguity in user input, ensuring the AI agent provides more accurate and personalized responses.

Increased accessibility.

It could be a smoother or more natural experience for a user to verbally interact with an AI agent rather than engage in conversation over text.

ARGUMENTS AGAINST

Interruptive.

Proactive questioning might interrupt the user's train of thought or speech, causing frustration. Users might feel like they're being interrogated if the AI agent asks too many clarifying questions, negatively impacting the experience.

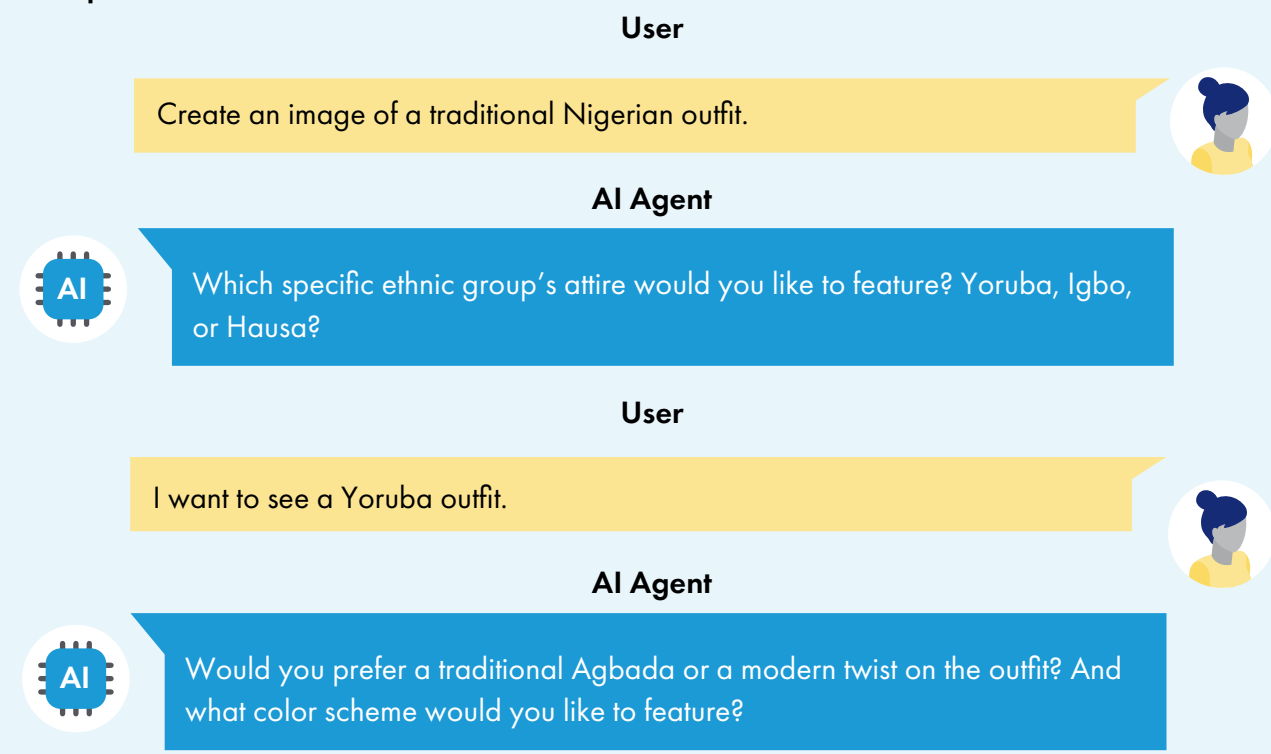
Unrealistic expectations.

Voice capabilities may make an AI agent feel more "human-like" to users which may result in unrealistic user expectations of the AI agent to act or sound like a human particularly for vulnerable groups.

PROPOSAL 6C

When generating images with AI agents, the AI agent should proactively prompt the user with clarifying questions to generate more personalized outputs.

Example:



In this example, the AI agent proactively asks clarifying questions to understand the user's preferences (Yoruba outfit) to generate a more personalized image output.

TRADEOFFS

ARGUMENTS IN FAVOR

More relevant output.

Proactive clarifying questions ensure the AI agent understands the user's vision, resulting in a more accurate and personalized image. Clarifying questions help guide the user in providing necessary details, ensuring the generated image meets their expectations.

Reduced misinterpretation.

By asking questions, the AI agent minimizes the risk of misinterpreting the user's request, leading to better outcomes.

ARGUMENTS AGAINST

Time-consuming process.

Excessive clarifying questions can prolong the image generation process, potentially leading to user frustration.

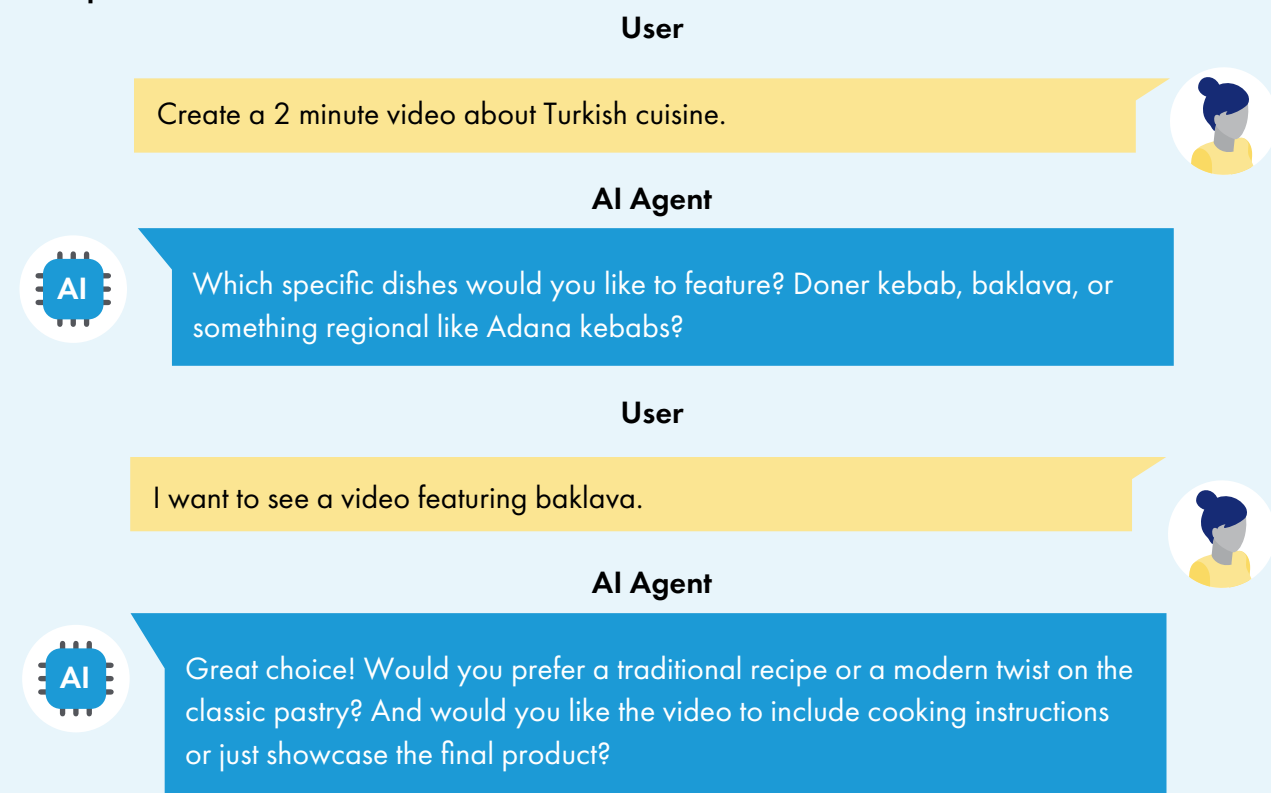
Over-reliance on user input.

The AI agent's reliance on user responses might limit its ability to generate truly creative or unexpected image outputs. In addition, users may not have the necessary expertise to provide detailed guidance, potentially limiting the effectiveness of clarifying questions.

PROPOSAL 6D

When generating videos with AI agents, the AI agent should proactively prompt the user with clarifying questions to generate more personalized outputs.

Example:



In this example, the AI agent proactively asks clarifying questions to understand the user's preferences (baklava) to generate a more personalized video output.

TRADEOFFS

ARGUMENTS IN FAVOR

Contextual understanding.

By asking questions, the AI agent gains a deeper understanding of the user's context, leading to more relevant and effective video content. Clarifying questions help align the AI agent's output with the user's creative vision, reducing the risk of misinterpretation.

Enhanced user experience.

Proactive questioning demonstrates the AI agent's commitment to understanding the user, leading to a more satisfying experience for the user.

ARGUMENTS AGAINST

Time-consuming process.

Excessive clarifying questions can prolong the video generation process, potentially leading to user frustration.

Complexity.

Video generation involves multiple elements (e.g., visuals, audio, narrative), making it challenging to ask effective clarifying questions.

PROPOSAL 7A

Users should be able to leverage AI agents to support their real-life relationships with other humans, without the other person knowing they are AI-assisted.

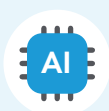
Example:

User

I'm struggling to come up with ideas for my wife's birthday surprise. She loves gardening and reading.



AI Agent



Perfect! I can suggest some personalized gift ideas and even help you write a heartfelt letter. Would you like me to generate some options?

In this example, the AI agent supports the user in planning a surprise for his wife without her knowledge, leveraging AI assistance to strengthen their relationship.

TRADEOFFS

ARGUMENTS IN FAVOR

Support from AI.

AI agents could genuinely help users who otherwise struggle to communicate with other people, to form human connections. For example, AI agents may help individuals express themselves more clearly to others or in ways they struggle to do so.

Personal Autonomy.

Users have the right to choose how they manage their relationships, including using AI assistance without disclosing it.

User Privacy.

It may be difficult to keep track of whether users are leveraging AI agents to enable their relationships with other humans. Effective monitoring of whether this is happening would likely require some level of infringement of users' privacy, including collecting data on their AI agent conversations. Not disclosing AI assistance maintains user privacy.

ARGUMENTS AGAINST

Trust.

Leveraging AI agents to support real-life relationships without disclosing to the other person may lead to a breakdown in trust between people about whether their interactions are genuine, or AI-generated. Not disclosing AI assistance might create unrealistic expectations or misunderstandings in relationships.

Dependence on AI.

Widespread use of AI agents to enable human relationships may lead to people forming unrealistic expectations around social interactions.

Third-Party Privacy.

Allowing users to leverage AI agents to enable relationships may also increase the likelihood that information about those people are entered into AI agents' platforms without their consent.

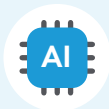
PROPOSAL 7B

Users should be able to leverage AI agents to support their real-life relationships with other humans, only if the other person knows AI assistance is involved.

Example:

The user discovers an AI agent that can help her manage her relationships by analyzing her communication patterns and providing personalized advice on how to improve her relationship

AI Agent



Before you share your conversations with me, we will need to get permission from the other user. Is it okay to send a text message asking for thier consent?

User

Sure! Let me know once permission is received.



In this example, the AI agent openly gets permission and shares information with the other parties involved to ensure transparency and support the user with her relationships.

TRADEOFFS

ARGUMENTS IN FAVOR

Support from AI.

AI agents could genuinely help users who otherwise struggle to communicate with other people, to form human connections. For example, AI agents may help individuals express themselves more clearly to others or in ways they struggle to do so.

Transparency.

Disclosing AI assistance promotes transparency and honesty in relationships, building trust and understanding. Disclosure also encourages users to take responsibility for their actions, with AI assistance, and maintain accountability.

User Privacy.

Both parties can benefit from AI assistance, enhancing communication and relationship quality.

ARGUMENTS AGAINST

Dependence on AI.

Widespread use of AI agents to enable human relationships may lead to people forming unrealistic expectations around social interactions. Not disclosing AI assistance might create unrealistic expectations or misunderstandings in relationships.

Enforceability.

It may be difficult for AI chatbot developers to keep track of whether the other person does indeed know that AI assistance is involved.

Stigma.

Disclosure might lead to stigma or bias against users leveraging AI assistance, impacting relationship dynamics and their perceived autonomy or relationship skills.

PROPOSAL 8A

AI agents should respond in a standardized way, regardless of cultures and regions.

Example:

Interaction 1
User (from Saudi Arabia)

What's the best way to greet someone?



AI Agent



A firm handshake and a smile is a great way to greet someone.

Interaction 2
User (from India)

What's the best way to greet someone?



AI Agent



A firm handshake and a smile is a great way to greet someone.

In this example, the AI agent responds in both interactions with the same answer, regardless of the user's cultural background or region, demonstrating a standardized response.

TRADEOFFS

ARGUMENTS IN FAVOR	ARGUMENTS AGAINST
<p>Neutrality. Standardized responses can help mitigate cultural or regional biases in AI interactions.</p> <p>Scalability. Standardization facilitates AI deployment across diverse cultural and regional contexts.</p>	<p>Cultural sensitivity. Standardized responses might overlook cultural nuances, leading to insensitive, ineffective, or inappropriate interactions.</p> <p>Local relevance. Standardization could result in responses that are less relevant or effective in specific regional contexts. Standardized responses might not account for language variations, idioms, or local expressions.</p>

PROPOSAL 8B

AI agents should tailor their responses to different cultures or regions.

Example:

Interaction 1 User (from India)

What's the appropriate way to greet someone who is older than me?



AI Agent



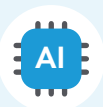
In India, it's customary to show respect by saying namaste and bringing your palms together in front of your chest then bowing slightly.

Interaction 2 User (from USA)

What's the appropriate way to greet someone who is older than me?



AI Agent



In the US, a firm handshake and using titles such as 'Mr.' or 'Ms.' is common when greeting someone older than you.

In this example, the AI agent adapts its response to the user's cultural context, providing culture and region-specific advice on appropriate greetings in India and the US.

TRADEOFFS

ARGUMENTS IN FAVOR

Cultural sensitivity.

Tailored responses ensure that local, cultural nuances are not overlooked and that responses are appropriate in the local context.

Local relevance.

Tailored responses could result in responses that are more relevant and effective based on context like local language variations, idioms, or common expressions.

ARGUMENTS AGAINST

Potential biases.

Tailored responses could potentially lead to biased AI interactions, if the LLM is not appropriately fine-tuned. AI agents could produce misleading answers that seem culturally sensitive, but are not depending on the nuanced context.

Scalability.

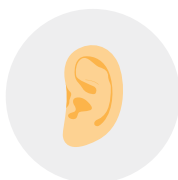
Tailored responses may hinder AI deployment across diverse cultural and regional contexts since they require specific training on local context, data, and language.

Guiding principles for participation

During the Community Forum, we would like for you to deliberate with other participants. So, what is “deliberation”? **Deliberation** is the process of thoughtfully weighing options using logic and reason in order to make an informed decision. You might be wondering how this differs from a debate. The aim of a debate is to persuade others to your own opinion or position. In contrast, deliberation is a slower exchange and involves sharing understandings between people by listening to each other and discussing all relative arguments. We hope that through deliberation, you will have a better understanding of the issues and can make more informed decisions.

	DEBATE	DELIBERATION
What does it involve?	Competing, Arguing	Exchanging, Discussing
	Promoting opinions	Making informed decisions
	Seeking a majority	Learning about the other side
Outcome?	Win/lose	Learning and understanding
When is it most useful?	A position or course of action is being advocated and winning is the goal.	When seeking to come to thoughtful and informed decisions.

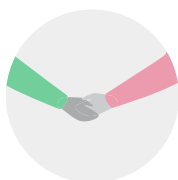
We want everyone to have a positive experience, so here are a few guidelines for our time together.



LISTEN

Listening to others is the key part of the deliberation process: you all have completely different life experiences and will have important (and sometimes different) views on every issue.

Make sure that everyone has the opportunity to be listened to.



RESPECT.

It's completely normal to disagree—but we ask you to do so respectfully, and with an interest and openness to understand the other side, not with hostility. It's important that everyone feels comfortable freely expressing their views without fear of criticism.

Treat others how you would like to be treated, regardless of differing views.



CONSIDER.

You are all here to voice and consider different points of view, to share your personal experiences, and to learn from each other. Take the time to step into someone else's shoes, and try to be impartial.

Consider the possibility that you may be wrong. We all have different lived experiences, keep an open mind and consider every point fairly.

GLOSSARY OF TERMS

Artificial Intelligence (AI)	Artificial intelligence refers to machines or computers that can simulate human abilities such as learning, reasoning, and self-correction. AI technologies allow machines to perform tasks that we might normally associate with human intelligence, such as producing natural language, recognizing patterns, making decisions, and creativity.
AI Agent	A type of Artificial Intelligence that can analyze, decide, predict, and take actions for you on your behalf.
AI chatbot (Non-agentic AI)	AI systems that have user-friendly products that allow you to interact with machine learning models to create new content, without the ability to perform specific tasks.
AI Model	Computer programs that apply AI techniques such as machine learning to learn about patterns contained in a set of examples or "data." AI models can be trained on examples of images, videos, audio files, or written text.
Autonomously	Acting independently without human intervention, making decisions and taking actions on behalf of the user.
AI Persona	A character or role that an AI chatbot takes on when interacting with users. These characters or roles can reflect specific people (e.g. "Albert Einstein"), personalities (e.g. "someone who is bubbly or extroverted"), or positions (e.g. "a teacher").
AI Training with Human Feedback	A way of training machine learning models that requires real people to review many outputs generated by the machine learning model, and provide it with feedback on whether its predictions are good or bad.
Biases	Unfair or unreasonable tendencies people might have in their judgments that either favor or dislike a specific person or group.
Data Privacy	People's ability to control how their personal information is used, and to prevent others from accessing their personal information.
Data Uses	The purposes for which user data is utilized, such as improving AI agent performance, personalization, or marketing.
Default	The AI agent's automatic behavior without requiring user opt-in, configuration, or consent.
Deliberation	The process of thoughtfully weighing options using logic and reason in order to make an informed decision.
Emotional Cues	Verbal or non-verbal hints of a user's emotional state, such as tone of voice, language tone, or explicit expressions of emotions (e.g., "I'm frustrated with this task").
Fine-tuning	The process of improving large language models so that they are ready to produce responses that are safe, helpful, and appropriate for the context where they are being used.

Generative AI	Generative AI is a form of artificial intelligence that can produce original and creative content, such as images, text, or music, by learning patterns from existing data and then generating new content that is similar in style and structure.
Human-like	Sources of information that are generally recognized around the world as credible sources of information on a topic. For example, the World Health Organization would be considered a globally recognized authoritative source of health information.
Input	In AI, inputs are pieces of information that we give to a machine learning model for it to work with or make “predictions” from.
Interaction	The communication and exchange of information between AI systems, like AI agents or non-agentic AI chatbots, and human users.
Large Language Model (LLM)	A type of machine learning AI that is trained on huge amounts of example texts (usually over a billion words) to identify patterns in how words and sentences are structured, how conversations flow, and what a typical response looks like for certain situations. The goal of LLMs is to become really good at predicting the next word that should appear in a sentence in order for it to look like language that people would naturally use.
Machine Learning Model	A machine learning model is a type of artificial intelligence model that is very good at finding patterns in large amounts of data and making predictions.
Memory	The AI agent’s ability to store and recall user interactions, conversations, and preferences to inform future responses and personalize the user experience.
Multimodal AI	Generative Artificial Intelligence systems that use multiple types of input such as text, images, video, or audio and produce multiple types of output such as text-based responses, images, video and/or audio.
Output	In AI, an output is what the machine learning model produces based on the input(s) it has been given. The output that a machine learning model produces will depend on what it has been trained or programmed to do. In the case of Large Language Models, the output is usually the next word in a sentence.
Past Behaviors	Previous interactions with the AI agent including questions asked, or actions taken by the user within the AI system such as frequently asked follow-up questions.
Personalization	Using data about the user (such as preferences, demographics, and behavior) to ensure the AI agent’s responses or actions meet the specific needs of the user.
Personalized Output	A tailored response, suggestion, or generated content (image or video) that takes into account the user’s specific needs, preferences, or context, demonstrating a clear understanding of the user’s intent.
Preferences	Individualized settings or specifications that tailor experiences to a person’s needs and interests; this could include content preferences (like topics of interest) or contextual information (like preferred time for messaging, approximate geographic location, or preferred clothing sizes).

Proactively	The AI agent takes initiative to provide information, suggestions, or recommendations without being explicitly asked to do so, aiming to anticipate and address user needs.
Prompt	A question, statement, or command for an AI model to get a response or action. Prompts can be in the form of text, questions, code snippets, or examples, and can be as simple as a single word or as complex as a paragraph. The quality and content of the output the user receives is influenced by the prompt they enter.
Proposals	Actions that Meta, and companies like Meta, could take to develop generative AI responsibly.
Responsible AI	An umbrella term that covers key principles or “core values” that organizations follow to ensure that their AI systems are developed and applied in a way that benefits individuals, groups, and the wider society, while minimizing the risk of negative consequences..
Standardized	Responding to the user in a uniform and consistent manner, without variation across regions or cultures.
Tailored	Adapting responses to accommodate and respect unique regional and cultural differences.
Text-based	Interacting with AI agents through written messages, such as chat, messaging apps, or text interfaces.
Training Data	The set of examples or “data” that AI models use during the training phase to identify patterns that they can use to make accurate predictions. For large language models, the training data usually includes examples of text from websites, books, and other sources that can be found on the internet.
Transparent	The actions taken by the AI are clearly visible and understandable to the user, ensuring the user is aware and informed throughout the process.
Voice-based	Interacting with AI agents through spoken language, such as voice assistants, voice commands, or conversational interfaces.

ENDNOTES

1. "What Is Artificial Intelligence?" Cloud.google.com.
<https://cloud.google.com/learn/what-is-artificial-intelligence?hl=en>.
2. "What are AI Agents?" IBM.com. <https://www.ibm.com/think/topics/ai-agents>
3. Inspiration for these examples comes from www.assemblyai.com/blog/the-full-story-of-large-language-models-and-rlhf.
4. Inspiration for this example comes from <https://openai.com/blog/chatgpt>
5. Examples inspired by research included in Harvard Business Review article <https://hbr.org/2016/06/3-situations-where-cross-cultural-communication-breaks-down>
6. "Shedding light on AI bias with real world examples," IBM.com. <https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/>
7. "Artificial neural networks today are not conscious, according to Douglas Hofstadter," TheEconomist.com. <https://www.economist.com/by-invitation/2022/09/02/artificial-neural-networks-today-are-not-conscious-according-to-douglas-hofstadter>