Meta

October 25, 2023

Dear Senator Markey,

Thank you for your September 27 letter. We appreciate the opportunity to address your questions on these important issues.

Meta has been a pioneer in AI for more than a decade, and we pride ourselves in working to help ensure that people can experience new technology trends in a responsible, safe, and thoughtful way. We believe generative AI technology will play a crucial role in the future, and we see great possibilities ahead for people who use our technologies, including creators and businesses. Generative AI is one of many experiences Meta offers that will unlock new and exciting ways to have fun and express creativity for all people who use our AI technologies, teens included.

As we deepen our investment in AI technology, we constantly consider how to develop and deploy these technologies responsibly. We know that AI has brought—and will continue to bring—huge advancements to society, but we also recognize that it comes with inevitable risks and the potential to cause unintended consequences. Technology companies must proactively work to address these issues, which is why we are working to help advance the responsible design and operations of AI technology and are committed to building this technology thoughtfully from the start.

Consistent with our Responsible Use Guide and best practices, we are building safeguards into our AI features and models prior to launch so people can have safer and more enjoyable experiences. This includes evaluating and improving our conversational AIs with external and internal experts through red teaming exercises, fine-tuning models, and training models on safety and responsibility guidelines, among other steps. We also regularly collaborate with policymakers, experts in academia and civil society, others in our industry, and community stakeholders, to advance the responsible use of this technology. We will continue building these safeguards to provide experiences that are safe and helpful for all individuals using our apps, including young people.

More broadly than the AI-specific safeguards discussed above, we recognize that teens, in particular, require safeguards for their safety, privacy, and wellbeing. That is why we have worked to implement age-appropriate transparency and education, tools meant to help teens better control their experiences online, as well as intuitive supervision tools that are designed to help parents support their teens, and more across our apps. And we will continue to seek feedback from parents and teens, including in the coming months, to better support their unique needs and improve our approach.

With that context in mind, please find answers to your specific questions below.

1. Please describe Meta's plans for introducing chatbots into its services, including any efforts to encourage younger users to use the chatbots.

AI is enabling new forms of connection and expression, thanks to the power of generative technologies. We have introduced new AI experiences and features that can enhance people's connections with others—and give them the tools to be more creative, expressive, and productive. Last month, we introduced Meta AI and 28 other AIs in beta, available to people who use WhatsApp, Messenger, and Instagram in the United States. Meta AI is an advanced conversational assistant that can give people real-time information through our search partnership with Bing or generate photorealistic images from text prompts in seconds. The 28 other AIs have more personality, opinions, and interests, and are a bit more fun to interact with.

We want these experiences to be safe and trustworthy, while bringing new forms of creativity, entertainment, and expression into people's lives. As we work on our vision for Generative AI, we want people, especially young people, to foster their online relationships in an environment where they feel safe, and where they leave our apps feeling good about the time they spend on them.

Meta has committed to taking a thoughtful approach towards teen experiences and leveraging the experience of experts in this field so that our technologies help protect the safety, privacy, and wellbeing of teens.

As we introduce AI experiences, we have taken steps to identify potential vulnerabilities, reduce risks, enhance safety, and bolster reliability. We evaluate and improve our conversational AIs with external and internal experts through red teaming exercises. We are fine-tuning our AI models, including by training the models to perform specific tasks with instructions that can increase the likelihood of providing helpful responses, and we are training them to provide expert-backed resources in response to safety issues.

We recognize that no AI model is perfect. We have built feedback tools within these features and will seek to use the feedback we receive to keep training the models to help improve safety performance and automatic detection of policy violations. We are also making our new generative AI features available to security researchers through Meta's long-running bug bounty program.

We are rolling out AI features methodically and in stages, so if a concern arises, we can work to address it before we expand access to the feature to more people. Overall, we are focused on working to make sure this technology will be beneficial to people's lives and additive to people's social interactions.

For more information regarding our efforts, please see our response to your Question 5.

2. Please describe Meta's plans to collect data from users of its chatbots and how this data will be used.

a. Will Meta commit to not using the data to target advertisements to young users? If not, why not?

To help protect teens on our apps, we have restricted the options advertisers have to reach them and have removed the ability for advertisers to target teens based on their interest and activities. We have also introduced more teen-specific controls and resources to help teens understand how ads work and the reasons why they see certain ads on our apps. Additionally, teens' engagement on our apps—like following certain Instagram posts or Facebook Pages—do not inform the types of ads they see. These measures reflect research, direct feedback from parents and child developmental experts, UN Children's Rights principles, and global regulation.

We are always working on more ways to help keep teens safe, provide them with privacy controls and educate them about how our technologies work. Earlier this year, we added a new privacy page with more information for teens about the tools and privacy settings they can use across our technologies. We have also added an updated section on generative AI to our Teen Privacy Guide to help teens make informed choices about how they use and interact with AI. The guide provides teens with an introduction to AI, how it works, how to recognize AI, how their information might be used by AI, and other helpful tips.

b. Will Meta commit to not using the data collected from young users to train its chatbots? If not, why not?

Generative AI models by design take a large amount of data to effectively train. As we explained in our post announcing our generative AI features last month, we used a combination of sources to train the models that power these features, including information that is publicly available online, licensed data, and information from Meta's technologies and services. We did not train these models using teens' private posts, nor do we use the content of people's messages with friends and family to train our AI models.

We do, however, use data about the prompts people enter into our AI-generated stickers features to improve those models. For example, if someone types in the word "dog", understanding which of several "dog sticker" options was selected by the user helps us provide better stickers for everyone in the future.

As part of our review process, we think carefully about how we collect and use data, with particular scrutiny on how we use data from teens who use our services. We believe that generative features bring compelling value to all users of our services, including teens. Given the broad appeal and usefulness of these features, it is imperative that we also take feedback and build models on data from teens, as well as adults. This means we need to build features that provide suitable responses for the way that teens use them, and also that we need to work to understand ways that these features might be misused by teens so we can build countermeasures. This does not change our stance on taking great care to build safety into all generative features.

3. Does Meta intend to include advertising in its chatbots? If so, how will Meta ensure that those advertisements are clearly identified and do not confuse users, especially younger users? Will Meta commit to not including targeted advertising in these chatbots? If not, why not?

We show advertising on our services—including in apps where people might use generative AI features—but we do not have any specific plans to include advertising in the responses from AI agents themselves.

As we announced last year, we do believe that generative AI has an important role to play in supporting businesses. Specifically, we are developing the ability for businesses to create AIs that reflect their brand's values and improve customer service experiences. From small businesses looking to scale to large brands wanting to enhance communications, AIs can help businesses engage with their customers across our apps. We are launching this in alpha and will scale it further next year.

We may expand the data we use to build our generative AI features in the future, and we'll continue to share details about that if and when we do. As we test and learn, we may consider using people's interactions with Meta's AI assistant or characters for other experiences and applications to better connect people and businesses on our services, such as new tools for expression, enhancements to our safety systems, or ads personalization.

The launch of these generative AI features at Connect does not change our existing practices around use of data for ads personalization. For example, advertisers can only use age and location to reach teens, and we still don't use the content of people's personal messages for ads targeting, which means advertisers cannot target people based on what they say in these messages. Also, we will continue to offer tools for ads transparency and control like our ads library and Why Am I Seeing This Ad.

We are making sure people know how to use our new AI features and understand their limitations. We also provide information to help people understand when they're interacting with AI and how this new technology works. We also flag that the technology may return inaccurate or inappropriate outputs.

For more information, please see our response to your Question 2(a).

4. Has Meta tested its chatbots to ensure they do not discriminate against users based on protected characteristics, including race, nationality, sex, sexual orientation, and gender identity? If so, please identify the testing that Meta has performed.

As referenced above, the custom AI models that power new text-based experiences like Meta AI, our large language model-powered assistant, are built on the foundation of Llama 2 and leverage its safety and responsibility training. In addition to the red teaming, fine-tuning, and training measures discussed above, we have also developed new technology designed to catch and take action on content that violates our policies. Our teams have built algorithms that work to scan and filter out harmful responses before they are shared back to people. We are also taking steps to help reduce potential bias. As with other AI models, having more people use the features and share feedback can help us refine our approach and our tools.

5. Has Meta conducted any research on the potential social and emotional impact of chatbots on younger users? If so, please include that research in your response. If not, why not? Has Meta consulted with experts and parents on the impact of chatbots on young users? If so, please describe those conversations. If not, why not?

We want teens to have positive, age-appropriate experiences online, and we are taking steps to help parents and teens safely navigate our generative AI experiences together. In addition to the safeguards discussed above, other safeguards we have implemented include:

- **New Supervision Tools for Parents and Guardians:** We are adding new features to our existing Parental Supervision Tools on Messenger, with similar features on Instagram available in the coming months. Specifically, we'll notify parents when their teen interacts with an AI Assistant/character for the first time and we are developing expert-backed resources to help them have conversations with their teens about how to use generative AI safely.

- **New Resources to Support Teens:** Our in-service disclosures and overall approach to transparency in this space has been developed with young people in mind, using age-appropriate language and reading levels. We have developed an easy-to-understand "Generative AI Teen Guide" so teens can make informed decisions about how they use and interact with AI. The guide provides teens with an introduction to what AI is, how it works, how to recognize AI, how their information might be used by AI and other helpful tips.

- **Helping to Make Generative AI features Age-Appropriate:** Our Generative AI features are only permitted for those aged 13 and up. We are routinely testing and retraining our models to help ensure that our AI features provide experiences that are age-appropriate for everyone 13 and older.

- **Providing Teens with Transparency:** To help teens understand whether they are interacting with something generated by our AI technologies, we will label photorealistic images with watermarks, and we make chats with AI visually distinct, so it is clear these are different from personal messages and conversations**.** This is available for all ages.

- **Providing Feedback on Responses:** We have developed feedback tools within our technologies so people can flag responses that they perceive to be unsafe or offensive, and we will use this feedback to continue training the models and improve our ability to restrict our AIs from providing such responses.

- **User Control Over AI interactions:** As our conversational assistant Meta AI works today, it cannot be brought into the chat by Meta in WhatsApp, Instagram or Messenger, and cannot message a person or their group first. In group chats with friends and family where someone asks Meta AI to join, this conversational AI only reads messages that invoke "@Meta AI" or messages where a reply invokes "@Meta AI." Meta AI does not access other messages in the chat.

We plan to continue working closely with parents and experts in mental health, psychology, youth privacy, and online behavior as we develop generative AI experiences. For example, we regularly consult with our Youth Advisory Council and Safety Advisory Council to develop features that help protect the safety and privacy of teens online. We are also working with experts and partners in the technology industry to help prevent generative AI services from being used to harm people, and we are routinely testing and retraining our models to help ensure that our AI features provide experiences that are safe and helpful for young people. And we solicit feedback from parents and teens about their safety, wellbeing, and privacy expectations across our services.

As the industry evolves, we will identify additional opportunities to undertake and iterate on our approach, so we can reflect technological changes, evolving industry standards, and global expectations around AI transparency.

6. How will Meta ensure that it complies with the commitments it made to the White House on AI safety in July 2023? For example, will Meta commit to publicly reporting its AI-powered chatbots' capabilities, limitations, and areas of inappropriate use and to prioritize research on potential discrimination and bias in its chatbots?

Meta supports risk-based, technology-neutral approaches to regulation of AI, and we welcomed the White House's voluntary commitments. The commitments are an important step in establishing responsible guardrails, and they created a model for other governments to follow.

AI should benefit the whole of society. For that to happen, these powerful new technologies need to be built and deployed responsibly. Meta approaches AI with the same core commitments to safety, security, and trust as the White House commitments. We joined these commitments because they represent an emerging industry-wide consensus around the things that we have been building into our services. For example, we have already committed to publicly report future AI systems' capabilities, limitations, and areas of appropriate and inappropriate use, as well as to prioritize research on the societal risks that future AI systems can pose.

The White House commitments focus on next-generation models that could be developed in the future and the risks and opportunities that could come along with them. These commitments are geared toward industry frontier models, as distinct from our recently released AI features developed using Llama 2.

Nevertheless, at Meta, our priority is to ensure that AI is developed and deployed responsibly - with transparency, safety and accountability at the forefront. We have invested heavily in transparency. For example, we released information about the decisions we made in building Llama 2 and resources to help developers build with Llama 2 responsibly, including a research paper with more details about our approach to safety improvements, finetuning, and red teaming.

Thank you, again, for the opportunity to provide information on this topic. We look forward to working with your office going forward.

Sincerely,

Kevin Martin

V.P. North America Policy