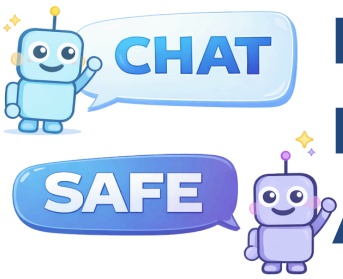

Building Trust and Prioritizing Safety in the Age of Conversational AI

Generative AI enabled chatbots can offer incredible opportunities for communication, learning, and innovation across every sector of society, providing people with accessible information tailored to their needs, level of background knowledge, and context.[1]

Companion chatbots differ from general-purpose chatbots and AI tools because they can foster a sense of emotional connection and serve as a general replacement for human contact. This can lead people to rely on companion chatbots for advice on sensitive topics, from mental health to financial decisions for which they are not designed and trained.

Interactions involving sensitive topics may involve the processing of personal, sensitive, or real-time user data. This introduces new risks we can't ignore, including the potential for surveillance, inadvertent misuse of sensitive data, and security vulnerabilities. AI systems gather information and make inferences about their users, and this data needs to be handled with care.

Furthermore, the capabilities of the generative models that power companion chatbots mean they can produce and quickly disseminate content that could be harmful, discriminatory, manipulative, or false. This demands strong and immediate safeguards to protect users from potential psychological, financial, and societal harm.

While we believe that Congress, first and foremost, needs to pass a comprehensive consumer privacy bill to ensure privacy protections for all Americans and that could address AI including chatbots, we do understand the additional interest in policies targeting AI chatbots.

The SIIA CHAT SAFE Principles propose a framework to simultaneously encourage the development of new technological tools and protect the privacy, safety, and security of all Americans.

[1] For the purposes of this document, a "companion chatbot" is an artificial intelligence system with a natural language interface that simulates an ongoing AI-human relationship or emotional connection in such a manner that it can serve as a replacement for real human social companionship. A companion chatbot does this by providing adaptive, human-like responses to user inputs and which may include exhibiting anthropomorphic features, asking emotion-based questions, and retaining information on interactions with the user. "General-purpose chatbots", which differ from "companion chatbots," may include a bot that is used only for customer service, a business' operational purposes, productivity and analysis related to source information, internal research, or technical assistance; or a bot that does not sustain a relationship across multiple iterations or generate outputs that are likely to elicit emotional responses in the user.

C CLEAR DISCLOSURES

Companion chatbots must include a clear and conspicuous disclosure that the chatbot is an artificial intelligence system and not a human. These disclosures should be repeated to the user periodically during sessions, where appropriate.

H HARM MITIGATION

Developers and deployers should implement a risk-based approach to mitigate potential harms for users. This should include maintaining a protocol for preventing companion chatbot conversations that encourage self-harm, suicide, or harm to others. If such expressions are detected, developers and deployers should surface appropriate resources. Additional guardrails, or reasonable measures to protect, are necessary for certain populations such as known children.

A ACCOUNTABILITY

Developers and deployers of companion chatbots should make reports available on company websites disclosing protocols to mitigate harms and establish trust. Enforcement of legislative provisions should rest with state attorneys general and the U.S. Federal Trade Commission.

T TRUST AND RELIABILITY

Developers and deployers should establish protocols, policies, and guardrails to enhance the reliability and trustworthiness of companion chatbots.

S SECURITY AND PRIVACY

Deployers of companion chatbots must maintain robust data privacy and security programs. In regulated industries, such as education and healthcare, special attention will be required to ensure compliance with the existing legal frameworks.

A ADAPTABILITY

Policymakers should encourage developers and deployers of companion chatbots to evolve and improve in response to new risks, threats, and user feedback. Developers and deployers, as appropriate, should also be proactive and respond quickly to new risks, threats, and user feedback.

F FOSTER INNOVATION

Policymakers and companies should embrace innovation to protect the safety, security, and privacy of users, such as through novel disclosures, bug bounty programs, and other emerging techniques. Additionally, ethical deployment of companion chatbots in regulated industries should be encouraged with the proper guardrails.

E EDUCATION AS A UNIQUE USE CASE

Chatbots may play a unique role in education as tutors and guides, and may employ pedagogical strategies – such as reinforcing, redirecting, and reminding –which may approximate emotional connection in order to effectively serve teaching and learning goals. Whether general-purpose or companion chatbots, these should align with the goals of CHAT SAFE as well as other federal and state privacy, safety, security, and civil rights requirements regulating technology in the classroom.