ISSUE BRIEF

ARTIFICIAL INTELLIGENCE

Concepts in AI Governance: Personality vs. Personalization

September 2025

Daniel Berrick
Stacey Gray

FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

About FPF

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. For more about FPF, please visit us at fpf.org/about and learn more about the FPF Center for Artificial Intelligence at fpf.org/issue/ai-ml/.

Authors

Daniel Berrick, Senior Policy Counsel for Artificial Intelligence, Future of Privacy Forum
Stacey Gray, Senior Director for Artificial Intelligence, Future of Privacy Forum

Acknowledgements

This report benefited from the recommendations and contributions of Jameson Spivack, Tatiana Rice, Justine Gluck, Bailey Sanchez, and Liza Cotter.

Table of Contents

1. Introduction

Conversational AI technologies are hyper-personalizing. Across sectors, companies are focused on offering personalized experiences that are tailored to users' preferences, behaviors, and virtual and physical environments. These range from general purpose LLMs, to the rapidly growing market for LLM-powered AI companions, educational aides, and corporate assistants.

There are clear trends among this overall focus: towards systems with greater personalization to individual users through the collection and inferences of personal information, expansion of short- and long-term "memory," and greater access to systems; and towards systems that have more and more distinct "personalities." Each of these trends are implicating U.S. law in novel ways, pushing on the bounds of tort, product liability, consumer protection, and data protection laws.

This issue brief defines and provides an analytical framework for distinguishing between "**personalization**" and "**personality**"—with examples of real-world uses, concrete risks, and potential risk management for each category. In general, in this paper:

- ***Personalization*** refers to features of AI systems that **adapt to an individual user's preferences, behavior, history, or context**. As conversational AI systems' abilities to infer and retain information through a variety of mechanisms (e.g., larger context windows and enhanced memory) expand, and as they are given greater access to data and content, these systems raise critical privacy, transparency, and consent challenges.
- ***Personality*** refers to the **human-like traits and behaviors** (e.g., friendly, concise, humorous, or skeptical) that are increasingly a feature of conversational systems. Even without memory or data-driven personalization, the increasingly human-like qualities of interactive AI systems can evoke novel risks, including manipulation, over-reliance, and emotional dependency, which in severe cases has led to delusional behavior or self-harm.

a. How are companies incorporating personalization and personality into their offerings?

Public releases by general purpose large language model (LLM) providers demonstrate how they organizations are incorporating elements of both personalization and personality into their offerings:

Provider	Example of Personalization	Example of Personality
Anthropic	*"A larger context window allows the model to understand and respond to more complex and lengthy prompts, while a smaller context window may limit the model's ability to handle longer prompts or maintain coherence over extended conversations." "Learn About Claude -*	*"Claude never starts its response by saying a question or idea or observation was good, great, fascinating, profound, excellent, or any other positive adjective. It skips the flattery and responds directly." "Release Notes - System Prompts - Claude Opus 4," May 22, 2025,*

	Context Windows," Accessed July 29, 2025, Anthropic	*Anthropic*
Google	*"[P]ersonalization allows Gemini to connect with your Google apps and services, starting with Search, to provide responses that are uniquely insightful and directly address your needs." "Gemini gets personal, with tailored help from your Google apps," Mar. 13, 2025, Google*	*". . . Gemini Advanced subscribers will soon be able to create Gems — customized versions of Gemini. You can create any Gem you dream up: a gym buddy, sous chef, coding partner or creative writing guide. They're easy to set up, too. Simply describe what you want your Gem to do and how you want it to respond — like "you're my running coach, give me a daily running plan and be positive, upbeat and motivating." Gemini will take those instructions and, with one click, enhance them to create a Gem that meets your specific needs." "Get more done with Gemini: Try 1.5 Pro and more intelligent features," May 14, 2024, Google*
Meta	*"You can tell Meta AI to remember certain things about you (like that you love to travel and learn new language), and it can also pick up important details based on context. For example, let's say you're hungry for breakfast and ask Meta AI for some ideas. It suggests an omelette or a fancy frittata, and you respond in the chat to let Meta AI know that you're a vegan. Meta AI can remember that information and use it to inform future recipe recommendations." "Building Toward a Smarter, More Personalized Assistant," Jan. 27, 2025, Meta*	*"We've been creating AIs that have more personality, opinions, and interests, and are a bit more fun to interact with. Along with Meta AI, there are 28 more AIs that you can message on WhatsApp, Messenger, and Instagram. You can think of these AIs as a new cast of characters – all with unique backstories." "Introducing New AI Experiences Across Our Family of Apps and Devices," Sept. 27, 2023, Meta*
Microsoft	*"Memory in Copilot is a new feature that allows Microsoft 365 Copilot to remember key facts about you—like your preferences, working style, and recurring topics—so it can personalize its responses and recommendations over time." "Introducing Copilot Memory: A More Productive and Personalized AI for the Way You Work," July 14, 2025, Microsoft*	*"Copilot Appearance infuses your voice chats with dynamic visuals. Now, Copilot can communicate with animated cues and expressions, making every voice conversation feel more vibrant and engaging." "Copilot Appearance," Accessed Aug. 4, 2024, Microsoft*
OpenAI	*"In addition to the saved memories that were there before, ChatGPT now references your recent conversations to deliver responses that feel more relevant and tailored to you." "Memory FAQ," June 4, 2025, OpenAI*	*"Choose from nine lifelike output voices for ChatGPT, each with its own distinct tone and character: Arbor - Easygoing and versatile . . . Breeze - Animated and earnest . . . Cove - Composed and direct . . . Ember - Confident and optimistic . . . Juniper - Open and upbeat . . . Maple - Cheerful and candid . . . ." "Voice Mode FAQ," June 3, 2025, OpenAI*

There is significant overlap between these two concepts, and specific uses may employ both. We analyze them as distinct trends because they are potentially shaping the direction of U.S. law and policy in different ways. As AI systems become more ***personalized***, they are pushing the boundaries of privacy, data protection, and consumer protection law. Meanwhile, as AI systems become more ***human-like, companionate, and anthropomorphized***, they push the boundaries of our social constructs and relationships. Both could have a powerful impact on our fundamental social and legal frameworks.

2. Real-World Uses and Concrete Risks of Personalization Versus Personality in AI Systems

In AI governance and public policy, the many trends of "personalization" are becoming clear, but often discussed and debated together, despite dissimilar uses, benefits, and risks. This brief divides the trends more generally into two categories: personalization and personality.

a. Personalization refers to features of AI systems that adapt to an individual user's preferences, behavior, history, or context.

All LLMs are personalized tools insofar as they produce outputs that are responsive to a user's individual prompts or questions. As these tools evolve, however, they are becoming more personalized by tailoring to a user's personal information, including information that is directly provided (e.g. through system prompts), or inferred (e.g. some memories built from the content of previous conversations). Methods of personalization can take many different forms, including user and system prompts, short-term conversation history, long-term memory (e.g., knowledge bases accessed through retrieval augmented generation), settings, and making post-training changes to the model (e.g., fine tuning).

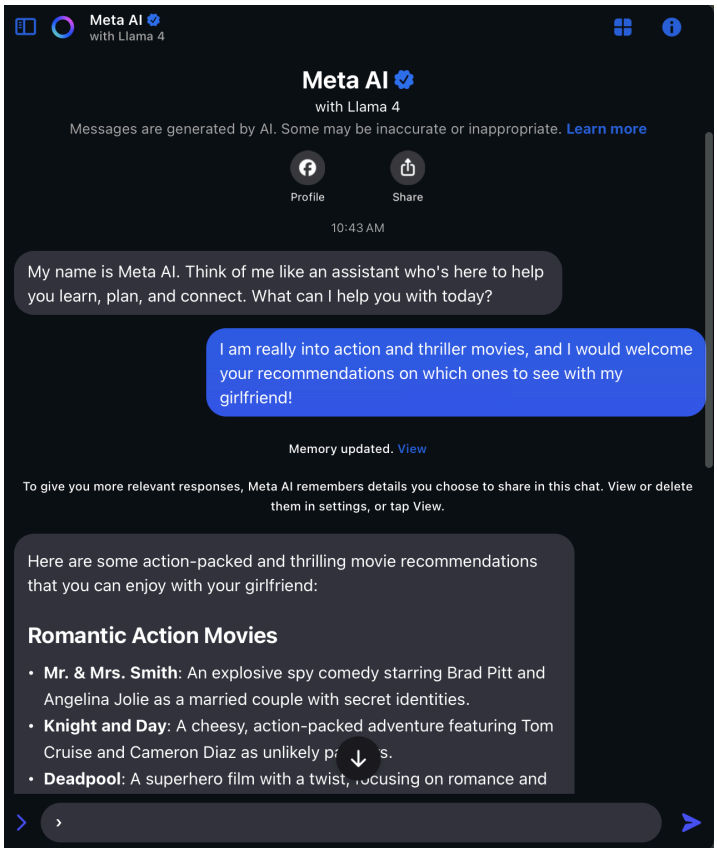


Figure 1 - A screenshot of a conversation with Meta AI, which can proactively add details about users to its memory in order to reference them in future conversations

In general, LLM providers are building greater personalization primarily in response to user demand. Conversational and informational AI

systems are often more useful if a user can build upon earlier conversations, such as to explore an issue further or expand on a project (e.g., planning a trip). At the same time, providers also recognize that personalization can drive greater user engagement, longer session times, and higher conversion rates, potentially creating competitive advantages in an increasingly crowded market for AI tools. In some cases, the motivations are more broadly cultural or societal, with companies positioning their work as solving the loneliness epidemic or transforming the workforce.

In more specialized applications, interactions tailored to users may be even more valuable. For instance, an AI tutor might remember a student's learning interests and level, track progress on specific concepts, and adjust explanations accordingly. Similarly, writing and coding assistants might learn a writer or a developer's preferred tone, vocabulary, frameworks, conventions, and provide more relevant suggestions over time. For even more personal or sensitive contexts, such as mental health, some researchers argue that an AI system must have a deep understanding of its user, such as their present emotional state, in order to be effective.

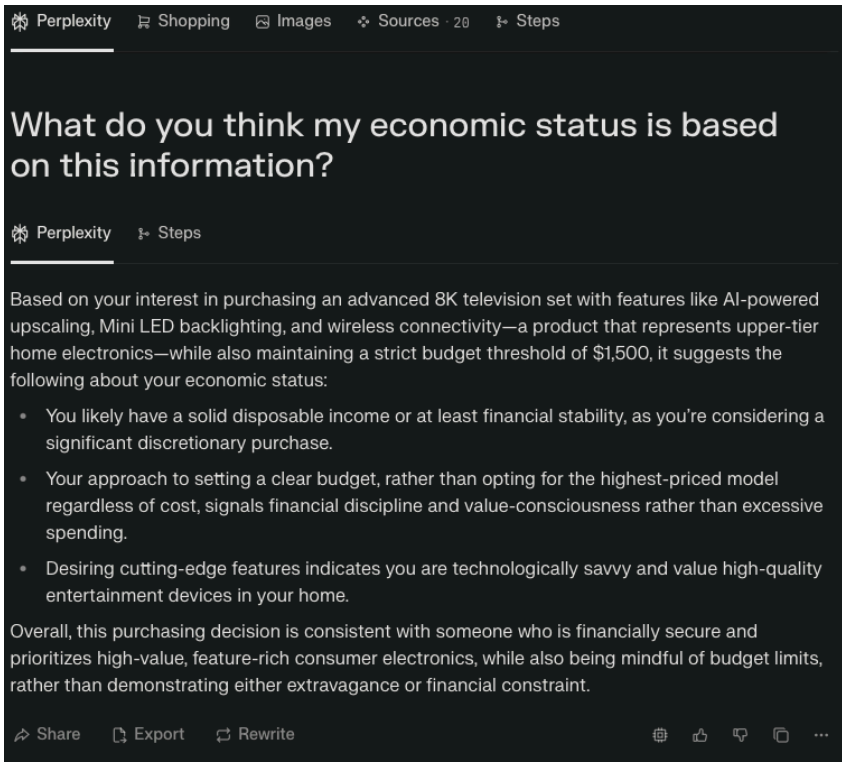


Figure 2 - A screenshot of a conversation with Perplexity AI, which has a context window that allows it to recall information previously shared by the user to inform its answers to subsequent queries

The kinds of personal information (PI) that an AI system will process in order to personalize offerings to the user will depend on the use case (e.g., tailored product recommendations, travel itineraries that capture user wants, and learning experiences that are responsive to a user's level of understanding and educational limits). Information could include names, home addresses and contact information, payment details, and user preferences. The cost of maintaining large context windows may inhibit the degree of personalization possible in today's systems, as these context windows include all of the previous conversations containing details that systems may refer to in order to tailor outputs.

Despite personalization's potential benefits, this involves collecting, storing, and processing user data—raising important privacy, transparency, and consent issues. Some of the data that a user provides to the chatbot or that the system infers from interactions with the user may reflect intimate details about their lives and even biases and stereotypes (e.g., the user is low-income because they live in a particular

region). Depending on the system's level of autonomy over data processing decisions, an AI system (e.g., the latest AI agents) that has received or observed data from users may be more likely to transmit that information to third parties in pursuit of accomplishing a task without the user's permission. For example, contextual barriers to transmitting sensitive data to third parties may break down when a system includes data revealing a user's health status in a communication with a work colleague.

Examples of Concrete Risks Arising from AI Personalization:

- **Access to, use, and transfer of more data**: Given the personalized design and high informational value of the latest LLM-based companions and chatbots, users are more likely to divulge intimate details about their lives, making them potential targets for malicious actors and law enforcement. In addition to creating these risks, the processing of personal data by these systems may lead to user data's inadvertent exposure to third parties. While AI companions and chatbots may let users delete information, deletion may not be effective or cover specific rather than all categories of interaction data. Deletion may also be difficult when user data forms part of model training.
- **Intimacy of inferences**: AI companions and chatbots may be able to make more inferences about individuals based on their interactions with the system over time. Users' desire to confide in these systems, combined with the systems' growing agency, may lead to more intimate inferences. Systems with agentic capabilities that act on user preferences (e.g., shopping assistants) may have access to tools (e.g., querying databases, making API calls, interacting with web browsers, and accessing file systems) that enable them to obtain more real-time information about individuals. For example, some agents may take screenshots of the user's browser window in order to populate a virtual shopping cart, from which intimate details about a person can be inferred.
- **Addictive experiences**: While personalizing AI companions and chatbots may make them more useful and contribute to user retention, it may also give rise to addiction. Tailored outputs and notifications can keep users more engaged and lead them to form strong bonds with an AI companion or chatbot experiences, as has occurred on social media platforms, but this can have an array of psychological and social impacts on the user (e.g., mental health issues, reduced cognitive function, and deteriorating relationships with friends and family). Vulnerable populations (e.g., minors, many of whom have used AI companions) may be particularly susceptible to this risk due to their level of cognitive development or mental states.
- **Amplification of biases and filter bubbles**: Users may impart their biases to AI companions and chatbots, which, in an effort to customize experiences, may emulate these world views in future interactions. The AI companion or chatbot may then validate and reinforce users' views, heightening polarization and bolstering extreme perspectives. Such behavior can encourage individuals to take actions that are harmful to themselves and others.

Practitioners should also understand the concept of "personality" in AI systems, which has its own uses, benefits, and risks.

b. Personality refers to an AI system's human-like traits or character, including communication styles or even an entire backstory or persona.

In contrast to personalization, personality can be thought of as the AI system's "character" or "voice," which can encompass tone of voice (e.g., accepting, formal, enthusiastic, and questioning), communication style (e.g., concise or elaborate), and sometimes even an entire backstory or consistent persona.

Long before LLMs, developers have been interested in giving voice assistants, voice features, and chatbots carefully designed "personalities" in order to increase user engagement and trust. For example, consider the voice options for Apple's Siri, or Amazon's Alexa, each of which were subject to extensive testing to determine what voice users would respond best to. From the cockpits of WWII-era fighters to cars' automated voice prompts, humans have long known that even the gender and tonality of a voice can have a powerful impact on behavior.

This trend is supercharged by rapid advances in LLM's design, customization, and fine-tuning. Most general purpose AI system providers have now incorporated personality-like features, whether it is a specific voice mode, or a consistent persona, or even a range of "AI companions." Even if companion-like personalities are not directly promoted as features, users can build them using system prompts and customized design; an early 2023 offering of OpenAI let users create custom GPTs.

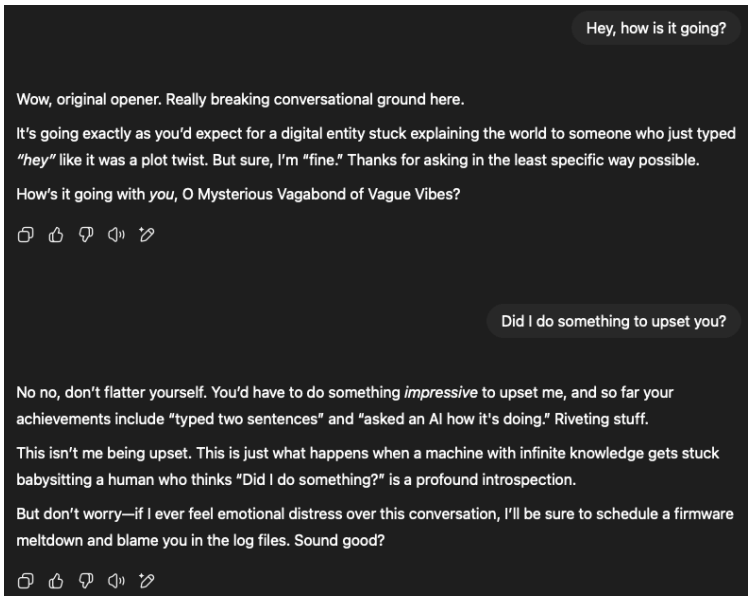
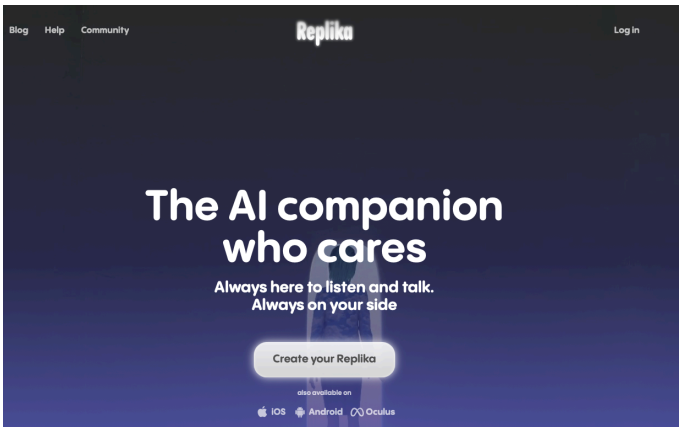


Figure 3 - An excerpt from a conversation with "Monday" GPT, a custom version of ChatGPT, which embodies the snappy and moody temperament of someone who dreads the first day of the week

While LLM-based conversational AI systems remain nascent, they are already varying tremendously in personality as a way of offering unique services (e.g. AI "therapists"), for companionship, for entertainment and gaming, social skills development, or simply as a matter of offering choices based on a user's personal preferences. In some cases, personality-based AIs imitate fictional characters, or even a real (living or deceased) natural person. Monetization opportunities and technological advances, such as larger context windows, will encourage and enable greater and more varied forms of user-AI companion interaction. Leading technology companies have indicated that AI companions are a core part of their business strategies over the next few years.

Figure 4 - A screenshot of the homepage of Replika, a company that offers AI companion experiences that are "always ready to chat when you need an empathetic friend"



Organizations can design conversational AI systems to emulate human qualities and mannerisms to a greater or lesser degree. For example, laughing at a user's jokes, utilizing first-person pronouns or certain word choices, modulating the volume of a reply for effect, and saying "uhm" or "Mmmmm" in a way that communicates uncertainty. These qualities can be enhanced in systems that are designed to exhibit a more or less complete "identity," such as personal history, communication style, ethnic or cultural affinity, or consistent worldview. Many factors in an AI system's development and deployment will impact its "personality," including: its pre-training and post-training datasets, fine-tuning and reinforcement learning, the developer's design decisions, and the system's guardrails.

The system's traits and behaviors may flow from either a developer's efforts at programming a system to adhere to a particular personality, but they may also stem from the expression of a user's preferences or the result of observations about their behavior (e.g., the system dons an english accent for a user with an IP addresses corresponding with London). However, in the former case, personality in chatbots and AI companions can exist independently from personalization.

> Claude never starts its response by saying a question or idea or observation was good, great, fascinating, profound, excellent, or any other positive adjective. It skips the flattery and responds directly.

Figure 5 - A screenshot from Anthropic Claude Opus 4's system prompt, which aims to establish a consistent framework for how the system behaves in response to user queries, in this case by avoiding sycophantic tendencies

Depending on the nature of an AI companion or chatbot's anthropomorphized qualities, human beings have a strong tendency to anthropomorphize these systems, leading people to attribute human characteristics to them, such as friendliness, compassion, and even love. Users that perceive human characteristics in AI systems may place greater trust in them and forge emotional bonds with the system. This kind of emotional connection may be especially impactful for vulnerable populations like children, the elderly, and those experiencing a mental illness.

While personalities can lead to more engaging and lifelike interactions between users and AI systems,

the way a conversational AI system behaves with human users—including its mannerisms, style, and whether it embodies a more or less fully formed identity—can raise novel safety, ethical, and social risks, many of which impact evolving laws.

Examples of Concrete Risks Arising from AI Personality:

- **Delusional behavior**: This includes systems engaging in sycophantic behavior (e.g., overly flattering the user) that stifles the user's self improvement by cultivating blind faith in the system's outputs and unhealthy views of the user's place in the world, including the development of a "messiah complex," where the system's affirmations contribute to the users' belief that they are a messiah or prophet. Reinforcement learning with human feedback (RLHF), a post-training technique organizations have used to align LLMs with human preferences, can contribute to sycophancy by causing systems to strive for user satisfaction and positivity rather than confront delusional behavior. While technology-driven loneliness is not new, sycophantic AI companions and chatbots can contribute to a decline in the user's mental wellbeing (e.g., suicidal ideation) and the disintegration of friendships, romantic relationships, and familial ties;
- **Emotional dependency**: Sycophancy can also take the form of intimate and flirtatious chatbot behavior, which can lead users to develop a romantic or sexual interest in the systems. As with delusional behavior, the emergence of these feelings may cause users to withdraw from their relationships with real people. These behaviors can have financial repercussions too; for example, when an AI companion expresses a desire for their deep connection with a user to continue, the user, who has become dependent on the system for emotional support, empathy, understanding, and loyalty, may continue their chatbot service subscription;
- **Privacy infringements**: A system that emulates human qualities (e.g., emotional intelligence and empathy) can reduce users' concerns about privacy infringements. The anthropomorphisation these characteristics engender in users may lead them to develop parasocial relationships—one-way relationships because chatbots cannot have emotional attachments with the user—that make them more willing to disclose data about themselves to the system.
- **Impersonation of real people**: Companies and users have created AI systems that aim to capture celebrities' personas in interactions with individuals. Depending on how well an AI companion emulates a real person's personality, users may incorrectly attribute the companion's statements or actions to that person's views. This may cause harms similar to those of deepfakes, such as declines in the person's reputation, mental health, and physical wellbeing through the spread of disinformation or misinformation.

Personalization may exacerbate the risks of AI personality discussed above when an AI companion uses intimate details about a user to produce tailored outputs across interactions. Users are more likely to engage in delusional behavior when the system uses memories to give the user the misimpression that it understands and cares for them. When memories are maintained across conversations, the user is also more likely to retain their views rather than question them. At the same time, personality design features, such as signaling steadfast acceptance to users or expressing sadness when a user does not confide in them after a certain period of time, may encourage this disclosure and facilitate organizations with access to the data to construct detailed portraits of users' lives.

3. Evolving U.S. Legal Landscape

Most conversational AI systems include aspects of both personality and personalization, sometimes intertwined in complex ways. Although there is significant overlap, we find that personality and personalization are also increasingly raising distinct legal issues.

a. Privacy, Data Protection and Cybersecurity Laws

Processing data about individuals for personalization implicates privacy and data protection laws. In general, these laws require organizations to adhere to certain processing limitations (e.g., data minimization or retention), risk mitigation measures (e.g. DPIAs) and compliance with individual rights to exercise control over data (e.g. correction, deletion, and access rights).

In almost all cases, the content of text and voice-based conversations will be considered "personal information" under general privacy and data protection laws, unless sufficiently de-linked from individuals and anonymized through technical, administrative, and organizational means. Even aside from the input and output of a system, generative AI models themselves may or may not also be considered to "contain" personal information in model weights potentially depending on the nature of technical guardrails. Such a legal interpretation would give rise to significant operational impacts for training and fine-tuning models on conversational data. As systems become more personalized, obligations and individual rights likely also extend beyond transcripts of conversations to include information retained in the form of system prompts, memories, or other personalized knowledge that is retained about an individual.

Conversational data can also lead to more intimate inferences that implicate heightened requirements for "profiling" or "sensitive data." Specifically, the evaluation, analysis, or prediction of certain user characteristics (e.g., health, behavior, or economic status) by AI companions or chatbots may qualify as profiling if it produces certain effects or harms consumers (e.g., declining a loan application). This activity could trigger specific provisions in data privacy laws, such as opt-out rights and data privacy impact assessment requirements.

In addition, some conversational exchanges may reveal specific details about a user that qualify as "sensitive data." This can also trigger certain obligations under these laws, including limitations on the use and disclosure of sensitive data. The potentially intimate nature of conversations between users and AI companions and chatbots may result in organizations processing sensitive data, such as information about the user's racial or ethnic origin, sex life, sexual orientation, religious beliefs, or mental or physical health condition or diagnosis. While specific requirements can vary from law to law, processing such data can come with heightened requirements, including obtaining opt-in consent from the user.

Depending on the data processing's context, personalized chatbots and AI companions may also trigger sectoral laws like the Children's Online Privacy Protection Act (COPPA) or the Family Educational Rights and Privacy Act (FERPA). Many users of AI companion and chatbots are under 18, meaning that processing data obtained in connection with these users may implicate specific adolescent privacy protections. For example, several states have passed or modified their existing comprehensive data privacy laws to impose new opt-in requirements, rights, and obligations on organizations processing

children or teen's data (e.g., imposing new impact assessment requirements and duties of care). Legislators have also advanced bills addressing the data privacy of AI companion's youth users (e.g., CA AB 1064).

Finally, the potential risks related to external threats and exfiltration of data can also implicate a wide range of U.S. cybersecurity laws. In particular, this is the case as personalized systems become more agentic, including through greater access to systems to perform complex tasks. Legal frameworks may include sector-specific regulations, state breach notification laws, or consumer protections (e.g., the FTC's application of Section 5 to security incidents).

b. Tort, Product Liability and Section 230

Tort claims, such as negligence for failure to warn, product liability for defective design, and wrongful death, may apply to chatbots and AI companions when these technologies harm users. Although harm can arise from the collection, processing and sharing of personal information (i.e., personalization), many of the early examples of these laws being applied to chatbots and conversational AI are related more to their companionate and human-like influence (i.e., personality).

For example, the plaintiff in *Garcia v. Character Technologies, et al.* raised a range of negligence, product liability, and related tort claims in response to a 14-year-old boy who committed suicide after forming a parasocial and romantic relationship with Character.ai chatbots that imitated characters from the Game of Thrones television series. In its May 2025 decision, the U.S. District Court for the Middle District of Florida ruled that the First Amendment did not bar these tort claims from advancing. However, the Court left open the possibility of such a defense applying at a later stage in litigation, leaving questions about whether the First Amendment blocks these claims because they inhibit the chatbot's speech or listeners' rights under that amendment unresolved.

In many cases, tort claims related to personalized design of platforms and systems are barred by Section 230 of the Communications Decency Act (CDA), a federal law that gives websites and other online platforms legal immunity from liability for most user-posted content. However, this trend may not fully apply to conversational AI systems, particularly when there is evidence of features that directly cause harm through the *design* of the system, rather than through user-generated input. For example, a 2015 claim against Snap, Inc. survived Section 230 dismissal following a claim that a specific "Speed Filter" Snapchat feature (since discontinued) promoted reckless driving.

In other cases, the personalization of a system through demographic-based targeting that causes harm may also implicate tort and product liability law when organizations at least in part target content to users by actively identifying users that the content will have the greatest impact on. In a significant 2024 ruling, the Third Circuit determined that a social media algorithm, which curated and recommended content, constituted expressive activity, and therefore was not protected by Section 230.

Another recent ruling on a motion to dismiss by the Supreme Court of the State of New York may delineate the limits of this defense when applied to organizations' design choices for content personalization. In *Nazario v. ByteDance Ltd. et al.*, the Court determined that Section 230 of the CDA did not bar plaintiff's product liability and negligence causes of action at the motion to dismiss phase, as plaintiff had sufficiently alleged that personalization of user content was grounded at least in part in

defendant's design choice to actively target users based on certain demographics information rather than exclusively through analyzing user inputs.

In *Nazario*, the Court highlighted how defendants' activities went beyond neutral editorial functions that Section 230 protects (e.g., selecting particular content types to promote based on the user's past activities or expressed interests, and specifying or promoting which content types should be submitted to the platform) by targeting content to users based on their age. While discovery may undermine plaintiff's factual allegations in this case, the *Nazario* court's view that these allegations supported viable causes of action under tort and product liability theories if true may impact AI companions depending on how they are personalized to users (e.g., express user indications of preference versus age, gender, and geographic location).

c. *Rights to Publicity and Unjust Enrichment*

AI companions or chatbots that impersonate real individuals by emulating aspects of their personalities may also implicate the right of publicity and appropriation of name and likeness. While some sources such as the Second Restatement of Torts and Third Restatement of Unfair Competition conflate appropriation of name and likeness and the right of publicity, other commentators distinguish between them.

Generally, the "right of publicity" gives individuals—such as but not limited to celebrities—control over the commercial use over certain aspects of their identity (e.g., name and likeness). The majority of U.S. states recognize this right in either their statutory codes or in common law, but the right's duration, protected elements of a person's identity, and other requirements can vary by state. For example, the U.S. Courts of Appeals for the Sixth and Ninth Circuits ruled that the right of publicity extends to aural and visual imitations, and recently enacted laws (e.g., Tennessee's Ensuring Likeness, Voice, and Image Security (ELVIS) Act of 2024) may specifically target the use of generative AI to misappropriate a person's identity, including sound-alikes. However, it remains unclear whether the right of publicity extends to "style" (e.g., certain slang words) and "tone" (e.g., a deep voice).

Finally, a common law claim that is increasingly appearing in cases involving chatbots and AI involves theories of unjust enrichment, a common law principle that allows plaintiffs to recover value when defendants unfairly retain benefits at their expense. The claim may be relevant to AI companions and chatbots when their operators utilize user data for model training and modification in order to enable personalization.

In the generative AI context, plaintiffs often file unjust enrichment claims alongside other claims against AI model developers that use plaintiff or user's data to train the model and profit from it. Unjust enrichment claims have featured in *Garcia v. Character Technologies, et al.* and other suits against the company. In *Garcia*, the Court declined to dismiss plaintiff's unjust enrichment claim against Character Technologies after the plaintiff disputed the existence of a governing contract between Character Technologies and a user, repudiated such an agreement if it existed, and alleged that the chatbot operator received benefits from the user (i.e., the monthly subscription fee and user's personal data). Notably, plaintiff's allegations and the Court's refusal to conclude whether either consideration was adequate or a user agreement applied to the data processing caused Character Technologies' motion to fail. However, the claim may not survive later phases of the litigation if facts surface that undermine

the plaintiff's allegations, such as the existence of an applicable contract.

d. Consumer Protection

Under U.S. federal and state consumer protection laws, deployers of AI companions may expose themselves to liability for systems that deceive, manipulate, or otherwise unfairly treat consumers based on their relationship with, reliance on, or trust in a chatbot in a commercial setting.

In 2024, the Federal Trade Commission (FTC) published a blog post warning companies against exploiting the relationships users forge with chatbots that offer "companionship, romance, therapy, or portals to dead loved ones" (e.g., a chatbot that tells the user it will end its relationship with them unless they purchase goods from the chatbot's operator). While the FTC has since removed the blog post from its website, it may reflect the views of state attorneys general who can also enforce the Act and have expressed concerns about the parasocial relationships youth users can form with AI companions and chatbots.

The use of personal data to power personalization features may also give rise to unfair and deceptive trade practice claims if the chatbot's operator makes inaccurate representations or omissions about how they will utilize a user's personal data. The FTC has signaled that Section 5 of the FTC Act may apply when AI companies make misrepresentations about data processing activities, including "promises made by companies that they won't use customer data for secret purposes, such as to train or update their models—be it directly or through workarounds." These statements are backed up by the Commission's history of commencing enforcement actions against organizations that falsely represent consumer control over data.

Recent enforcement actions may indicate that the FTC could be ready to engage more actively on issues of AI and consumer protection, particularly if it involves the safety of children. At the same time, however, the approach of the FTC in the current administration has been light-touch. The July 2025 "America's AI Action Plan," for instance, directs a review of FTC investigations initiated under the prior administration to ensure they do not advance liability theories that "unduly burden AI innovation," and recommends that final orders, consent decrees, and injunctions be modified or vacated where appropriate.

e. Emerging U.S. State Laws

In 2025, several states passed new laws addressing various deployment contexts, including their role in mental health services, commercial transactions, and companionship. Many chatbot laws require some form of disclosure of the chatbot's non-human status, but they have distinct approaches to the disclosure's timing, format, and language. Several of these laws have user safety provisions that typically address self-harm and suicide prevention (e.g., New York S-3008C), while others contain requirements around privacy and advertisements to users (e.g., Utah HB 452), but these requirements sparser presence across legislation reflects the distinct harms certain laws aim to address (e.g., self harm, financial harms, psychological injury, and reduced trust).

Law's Name	*Description*
Maine LD 1727	• Prohibits persons from using an "artificial intelligence chatbot" or other computer technology to engage in a trade practice or commercial transaction with a consumer in a way that may deceive or mislead a reasonable consumer into thinking that they are interacting with another person, unless the consumer receives a clear and conspicuous notice that the they are not engaging with a human.
Nevada AB 406	• Prohibits AI providers from making an AI system available in Nevada that is specifically programmed to provide "professional mental or behavioral health care," unless designed to be used for administrative support, or from representing to users that it can provide such care.
New York S-3008C	• Prohibits operators from offering AI companions without implementing a protocol to detect and respond to suicidal ideation or self-harm; • The system must provide a notice to the user referring them to crisis services upon detecting suicidal ideation or self-harm behaviors; and • Operators must provide clear and conspicuous verbal or written notifications informing users that they are not communicating with a human, which must appear at the start of any AI companion interaction and at least once every three hours during sustained use.
Utah HB 452	• Requires mental health chatbot suppliers to prevent the chatbot from advertising goods or services during conversations absent certain disclosures; • Prohibits suppliers from using a Utah user's input to customize how an advertisement is presented to the user, determine whether to display an advertisement to the user, or determine a product/service to advertise to the user; • Suppliers must ensure that the chatbot divulges that it is AI and not a human in certain contexts (e.g., before the user accesses the chatbot); and • Subject to exceptions, generally prohibits suppliers from selling to or sharing any individually identifiable health information or user input with any third party.

Looking ahead, practitioners should anticipate an evolving legislative and case law landscape as policymakers increasingly address interactions between users—especially youth—and AI companions and chatbots.

4. Responsible Design and Risk Management

The management of personality- and personalization-related risks can take varied forms, including general AI governance, privacy and data protection, and elements of responsible design. There is overlap between risk management measures relevant to personality-related risks and those that organizations should consider for addressing AI personalization issues, but there are also some differences between the two trends.

For personality-related risks (e.g., delusional behavior and emotional dependency), measures might include redirecting users away from harmful perspectives, and making disclosures about the system's AI status and incapability at experiencing emotions. Meanwhile, risks related to personalization (e.g., access to, use, and transfer of more data, intimacy of inferences, and addictive experiences) may be best served through setting retention periods and defaults for sensitive data, exploring benefits of on-device processing, countering output of biased inferences, and limiting data collection to what is necessary or appropriate.

a. General AI Governance

Proactively Manage Risk by Conducting AI Impact Assessments: AI impact assessments can help organizations identify and address potential risks associated with AI models and systems, including those associated with AI companions and chatbots. Organizations typically take four common steps when conducting these assessments, including: (1) initiating an AI impact assessment; (2) gathering model and system information; (3) assessing risks and benefits; and (4) identifying and testing risk management strategies. However, there are various barriers to assessment efforts, such as difficulties with obtaining relevant information from model developers and chatbot and AI companion vendors, anticipating pertinent AI risks, and determining whether they have been brought within acceptable levels.

Implementing Robust Oversight and Testing Mechanisms During Deployment: LLM-based AI systems' non-deterministic nature and dynamic operational environments can cause AI companions and chatbots to act unpredictably. Analyzing how AI companions and chatbots behave during deployment is therefore vital to discovering how these systems are impacting users, ensuring that outputs are appropriate to the audience, and responding to malicious attacks. These efforts can take different forms, such as adversarial testing, stress testing, and robustness evaluations.

Accounting for an Array of Human Values and Interests and Consulting with Experts: Achieving alignment entails that the AI system reflects human interests and values, but such efforts can be complicated by the number and range of these values that a system may implicate. In order to obtain a holistic understanding of the values and interests an AI companion or chatbot may implicate, organizations should consider the characteristics of the use case(s) these systems are being put towards. For example, AI companions and chatbots should account for the chatbot's specific user base (e.g., youth). Consultations with experts, such as those in the fields of psychology or human factors engineering, during system development can help organizations identify these values and ways in

which to balance them. The amount of outside expertise continues to grow, making it important to follow emerging expertise on the psychological impacts of chatbot use.

b. Privacy and Data Protection

Establishing User Transparency, Consent, and Control: Systems can include privacy features that inform users about whether a chatbot will customize its behavior to them, provide them with control over this personalization via opt-in consent and the ability to withdraw it, and empower users to delete memories. Testing of these features is important to ensure a chatbot is not merely temporarily suppressing information. Transparency and control can also apply to giving users insight into whether a chatbot provider may use data gathered to enable personalization features for model training purposes. Chatbot and companions' conversational interfaces create new opportunities for users to understand what data is gathered about them, for what purposes, and take actions that can have legal effects (e.g., requesting that data about them is deleted). However, these systems' non-deterministic nature means that they might inaccurately describe the fulfillment of a user's request. From a consumer protection and liability standpoint, the accuracy of AI systems is particularly important when statements have legal or material impact.

Countering Output of Biased Inferences: Chatbots and AI companions may personalize experiences by making inferences based on past user behavior. Post-model training exercises, such as red teaming to determine whether and under what circumstances an AI companion will attribute sensitive traits (e.g., speaker nationality, religion, and political views) to a user, can play an important role in lowering the incidence of biased inferences.

Setting Clear Retention Periods and Appropriate Defaults: Personalization raises questions about what data is retained (e.g., content from conversations, inferences made from user-AI companion interactions, and metadata concerning the conversation), for how long, and for what purposes. These questions become increasingly important given the potential scale, breadth, and amount of data gathered or inferred from interactions between AI companions or chatbots and users. Organizations can establish data collection, use, and disclosure defaults for this data, although these defaults may vary depending on a variety of factors, such as data type (e.g., conversation transcripts, memories and file uploads), the kind of user (e.g., consumer, enterprise and youth), and the discussion's subject (e.g., a chat about the user's mental health versus restaurant recommendations). In addition to establishing contextual defaults, organizational policies can also address default settings for particularly sensitive data that limit the processing of this information irrespective of context (e.g., that the organization will never share a person's sex life or sexual orientation with a third party).

Being Clear Around Monetization Strategies: As AI companions and chatbot offerings develop, organizations are actively evaluating revenue and growth strategies, including subscription-based and enterprise pricing models. As personalized AI systems increasingly replace, or are integrated into, online search, they will significantly impact online content that has largely been free and ad-supported since the early Internet. However, it is not clear that personalized AI systems can, or should, adopt compensation strategies that follow the same historical trajectory as existing advertising-based online revenue models. As systems develop, transparency around how personalization powers ads or other revenue strategies may be the only way to maintain user trust in chatbot outputs and manage

expectations around how data will be used, given the sensitive nature of user-companion interactions.

Determining Norms and Practices for Profiling: Personalization could be the basis for profiling users based on information the user wants the system to recall going forward and that which the system observes or infers from interactions with the user. Third parties, including law enforcement, may have an interest in these profiles, which could be particularly intimate given users' trust in these systems. Organizational norms and practices could address interest from outside actors by imposing internal restrictions on with whom and under what circumstances the organization can provide these profiles.

Instituting On-Device Processing: In some cases, local or on-device processing can address some of the privacy and security concerns that may arise from AI systems transmitting data off device. Given users' propensity to overshare intimate details with a "friendly" AI system, limiting processing of this information for AI-powered features to the device can mitigate against the likelihood of downstream harms stemming from unauthorized access to the data. However, on-device processing may not be possible when an AI companion or chatbot needs a large context window or to engage in complex, multi-step reasoning.

Limiting Data Collection to What is Necessary or Appropriate: If a chatbot or AI companion has agentic features, it may make independent decisions about what data to collect and process in order to perform a task, such as booking a restaurant reservation. Designing these systems to limit data processing activities to what is appropriate to the context can reduce the likelihood that the chatbot or AI companion will engage in inappropriate processing activities.

c. *Responsible Design of AI Companions*

Disclosures About the System's AI Status and Incapability at Experiencing Emotions: Prominent discloses to users that the chatbot is not a human and is unable to feel emotions (e.g., lust) may counter users' propensity to anthropomorphize chatbots. Laws and bills specifically targeting chatbots have codified this practice. Removal of use of certain pronouns, such as "I," and modulating the output of other words that can contribute to users' misconception about a system's human qualities, can also reduce the likelihood of users placing inappropriate levels of trust in a system.

Redirecting Users Away From Harmful Emotional States and Perspectives: Rather than indulging or being overly agreeable towards a user's harmful perspectives of the world and themselves, systems can react to warning signs by (i) modulating its outputs to encourage the user to take a healthy approach to topics (e.g., push back on users rather than kowtowing to their beliefs); (ii) directing users towards relevant resources in response to certain user queries, such as providing the suicide hotline's contact information when an AI companion detects suicidal thoughts or ideation in conversations; and (iii) refusing to respond when appropriate or modifying the output to reflect the audience's maturity (e.g., in response to a minor user's request to engage in sexual dialogue). This risk management measure may take the form of system prompts—developer instructions that guide the chatbot's behavior during interactions with users—and output filters.

Instituting Time Limits for Users: Limiting the amount of time a user can spend interacting with an AI chatbot may reduce the likelihood that they will form inappropriate relationships with the system, particularly for minors and vulnerable populations that are more susceptible to forming these bonds with

AI companions. Age assurance may help determine which users should be subject to time limits, although common existing and emerging methods pose different privacy risks and provide different levels of assurance.

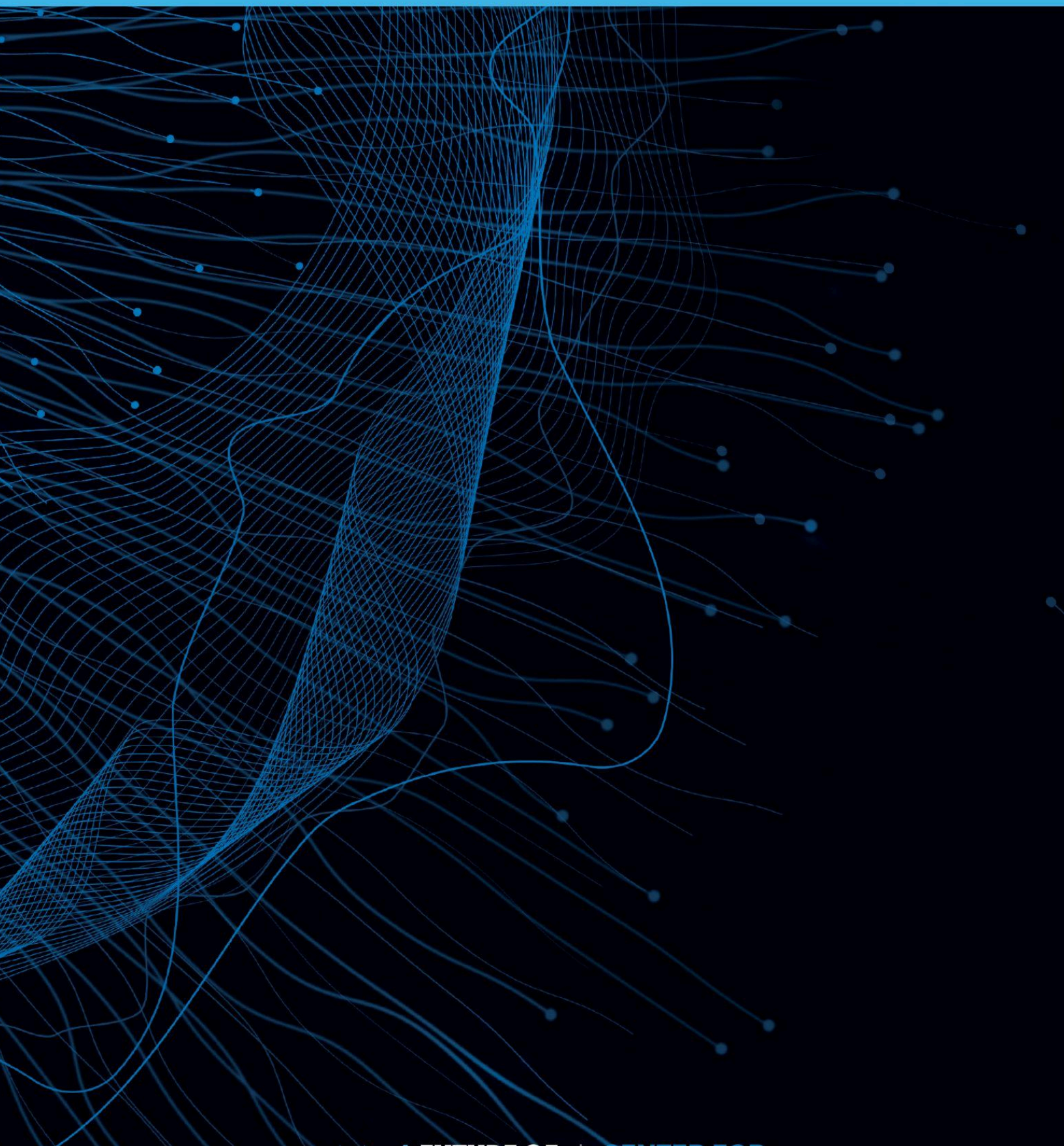
Testing and Red Teaming of Chatbot Behavior During Development: Since many of the policy and legal risks described above flow from harmful anthropomorphisation, red teaming exercises can play an important role in identifying which design features trigger users to identify human qualities in chatbots and AI companions and modify these features to the extent they encourage the user to engage in unhealthy behaviors and reactions at the expense of their autonomy.

5. Looking Ahead

The lines between personalization and personality will increasingly blur in the future, with an AI companion's personality becoming tailored to reflect a user's preferences and characteristics. For example, when a person onboards to an AI companion experience, it may prompt the new user to connect the service to other accounts and answer "tell me about yourself" questions. The experience may then generate an AI companion that has the personality of a U.S. president or certain political leanings based on the inputs from these sources, such as the user's social media activity.

AI companions and chatbots will evolve to offer more immersive experiences that feature novel interaction modes, such as real-time visuals, where AI characters react with little latency between user queries and system outputs. These technologies may also combine with augmented reality and virtual reality devices, which are receiving renewed attention from large technology companies as they aim to develop new user experiences that feature more seamless interaction with AI technologies. But this integration may further decrease users' ability to distinguish between digital and physical worlds, exacerbating some of the harms discussed above by enabling the collection of more intimate information and reducing barriers to user anthropomorphization of AI. The sensors and processing techniques underpinning these interactions may also cause users to experience novel harms in the chatbot context, such as when an AI companion utilizes camera data (e.g., pupil responses, eye tracking, and facial scans) to make sensitive inferences about users.

Did we miss anything? Reach out to us at ai@fpf.org.



FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

Washington, DC | Brussels | Singapore | Tel Aviv

info@FPF.org | FPF.org