It’s (Getting) Personal

How Advanced AI Systems Are Personalized

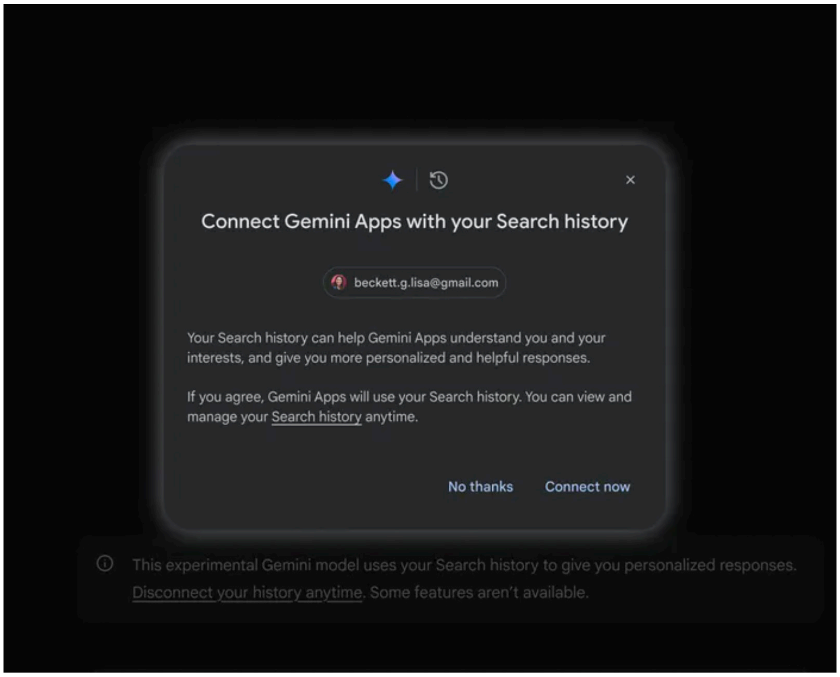
May 2025

Authored by
Princess Sampson, *Fellow, CDT AI Governance Lab*
Miranda Bogen, *Director, CDT AI Governance Lab*

Generative artificial intelligence has reshaped the landscape of consumer technology and injected new dimensions into familiar technical tools. Search engines and research databases now by default offer AI-generated summaries of hundreds of results relevant to a query, productivity software promises knowledge workers the ability to quickly create documents and presentations, and social media and e-commerce platforms offer in-app AI-powered tools for creating and discovering content, products, and services.

Many of today’s advanced AI systems like chatbots, assistants, and agents are powered by foundation models: large-scale AI models trained on enormous collections of text, images, or audio gathered from the open internet, social media, academic databases, and the public domain. These sources of reasonably generalized knowledge allow AI assistants and other generative AI systems to respond to a wide variety of user queries, synthesize new content, and analyze or summarize a document outside of their training data.

But out of the box, generic foundation models often struggle to surface details likely to be most relevant to specific users. AI developers have begun to make the case that increasing personalization will make these technologies more helpful, reliable, and appealing by providing more individualized information and support. As visions for powerful AI assistants and agents that can plan and execute actions on behalf of users motivate developers to make tools increasingly “useful” to people — that is, more personalized — practitioners and policymakers will be asked to weigh in with increasing urgency on what many will argue are tradeoffs between privacy and utility, and on how to preserve human agency and reduce the risk of addictive behavior.



Connect Gemini Apps with your Search history

beckett.g.lisa@gmail.com

Your Search history can help Gemini Apps understand you and your interests, and give you more personalized and helpful responses.

If you agree, Gemini Apps will use your Search history. You can view and manage your Search history anytime.

This experimental Gemini model uses your Search history to give you personalized responses. Disconnect your history anytime. Some features aren't available.

Figure 1. Google is experimenting with offering users the ability to personalize their experience with Gemini using their search history.

Source: *https://blog.google/products/gemini/gemini-personalization/*

Much attention has been paid to the immense stores of personal data used to train the foundation models that power these tools. This brief continues that story by highlighting how generative AI-powered tools use user data to deliver progressively *personalized experiences*, teeing up conversations about the policy implications of these approaches.[1]

How Advanced AI Systems are Personalized

Personalization involves altering systems on the basis of information about individual users or groups of users, changing the functionality, interface, and content of a system depending on who is interacting with it. For instance, users may be served information and resources that are predicted to be most likely to align with their interests, shown details that are dialed in to a particular level of presumed understanding, or receive customer service and other support that automatically integrates information about their specific context and experiences.
In the realm of AI, personalization could enable assistant or agent tools to provide more salient responses or support without users specifying all relevant details in their prompts, inform more tailored responses that include personal information or are more engaging to a particular individual, and support systems that rely on intimate knowledge of personal details.

As with other tools, AI personalization can be either *user-initiated* or *platform-initiated*, based on data that is *directly provided by* or *inferred* about the user, and give users *explicit* or *implicit*

[1] Generative AI can also be used as a tool to power personalization in other contexts, such as recommender systems — for instance, by analyzing user-generated content alongside users’ activity or interests to make suggestions or shape product behavior — but this is distinct from efforts to personalize generative AI-powered tools themselves.

awareness of when and how their experiences are personalized, or a combination thereof. For instance, users might solicit some degree of personalization by providing a few details about themselves and their interests, while AI providers may continue to customize their experience behind the scenes based on preferences inferred through interactions over time.

The section below details the technical means by which advanced AI systems can be personalized, in order to inform future conversations around what sorts of user experiences or policies may be needed to address harmful downstream impacts these products may bring about, foreseen or not. At a high level, these methods include personalizing based on information included in the *model's context*, including information provided as part of a user's prompt; information contained in conversation history, which is generally fed back to the model as part of the prompt; and system prompts, which a user can explicitly alter via their account settings. Methods also include personalizing based on *long-term memory*, or tools apart from the model itself that are used to supply the model with personalized information; *static settings*, or account information that bluntly controls what information is presented to users; and *model modification*, or directly changing foundation models to reflect personal preferences through methods like fine-tuning.

Model Context

A few different approaches to personalization leverage *model context*, or the information a language model can process and reference to generate a useful response, such as user inputs and system prompts. Technical limitations typically constrain how much background information, previous conversation details, or supplementary context a query can incorporate to shape an AI's output. For instance, a model with a short context window would only be able to remember details from the most recent few prompts but "forget" details that a user had explicitly provided a few moments earlier, while more context would mean a model could access details provided over days of interaction history or parse long, user-submitted documents when crafting its outputs.

User Prompts

Customizing AI outputs based on information that users explicitly include in what they submit to an interactive AI system

In interactive AI systems, users often provide context relevant to their query or instructions via prompts in natural language, usually taking the form of short, informal conversations. Even if a system does not explicitly seek to collect or adapt to users' personal details, a user might provide personal or contextual information about themselves or their task within a prompt to get a more tailored response or output from their AI assistant. For instance, if a user requests information about local restaurants and specifies that they live in the Wicker Park neighborhood of Chicago, the response would include restaurants in or near that specific neighborhood. Some chatbots may be prevented from responding directly to prompts that contain personal information out of concern about perceptions that the system is behaving improperly — for instance, not relying on a protected characteristic that a user has shared when being asked to make recommendations about jobs or housing. However, savvy users have found ways to work around these sorts of constraints, such as by asking an AI assistant to imagine a hypothetical situation or respond in the role of a particular type of professional.

Short-term Conversation History

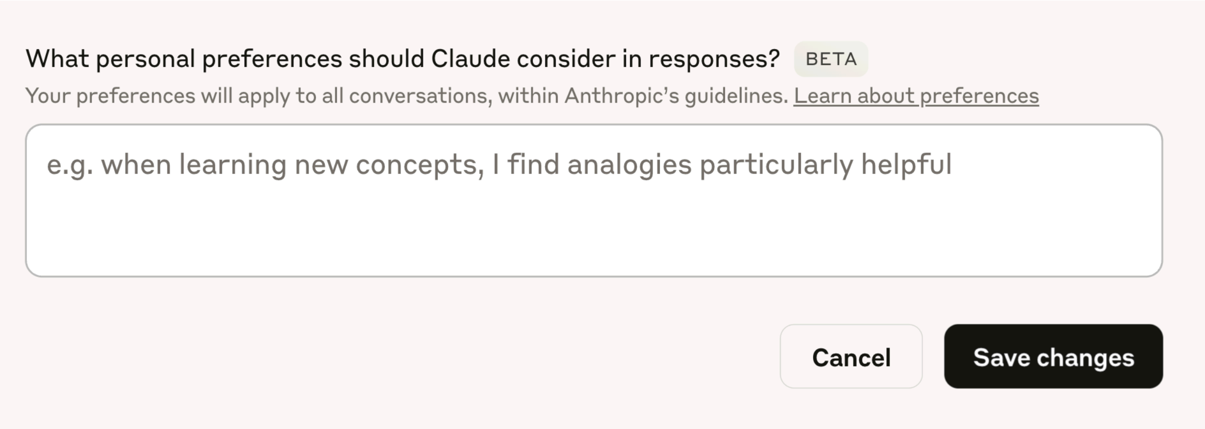
Augmenting user prompts behind the scenes with verbatim or summarized content from earlier in the interaction to recall pertinent details

Short-term memory typically involves maintaining context within a single extended interaction or conversation. Think of these memories like sticky notes — they can be stored and accessed quickly, but aren’t terribly durable, can be overloaded by too much detail, and are not an effective way to organize and retrieve a lot of information. Developers enable short-term memory by modifying queries that users submit to the AI tool to include additional information, and then transmitting the augmented prompts to the underlying model. For instance, chatbots may append previous messages to a prompt behind the scenes in order to “remember” information from earlier in an interaction, distill earlier messages into summaries that are included in prompts, or cache key facts like names or dates from the conversation that can be called later in the session. Advanced models that support significantly larger context windows can enable more contextually-rich — and personalized — communication, appearing to remember information as more details from the session can be included in each prompt.

Information that is provided solely in the context window, and not stored in a way that can be easily retrieved, does not typically persist across multiple sessions, so systems that rely only on this method of personalization will “forget” details if a conversation is refreshed or a user starts a new session. Depending on the service provider and user settings, conversation logs may still be stored for further model training, fine-tuning, or debugging even if they are not immediately absorbed into the AI system in a way that powers personalized experiences.

System prompts

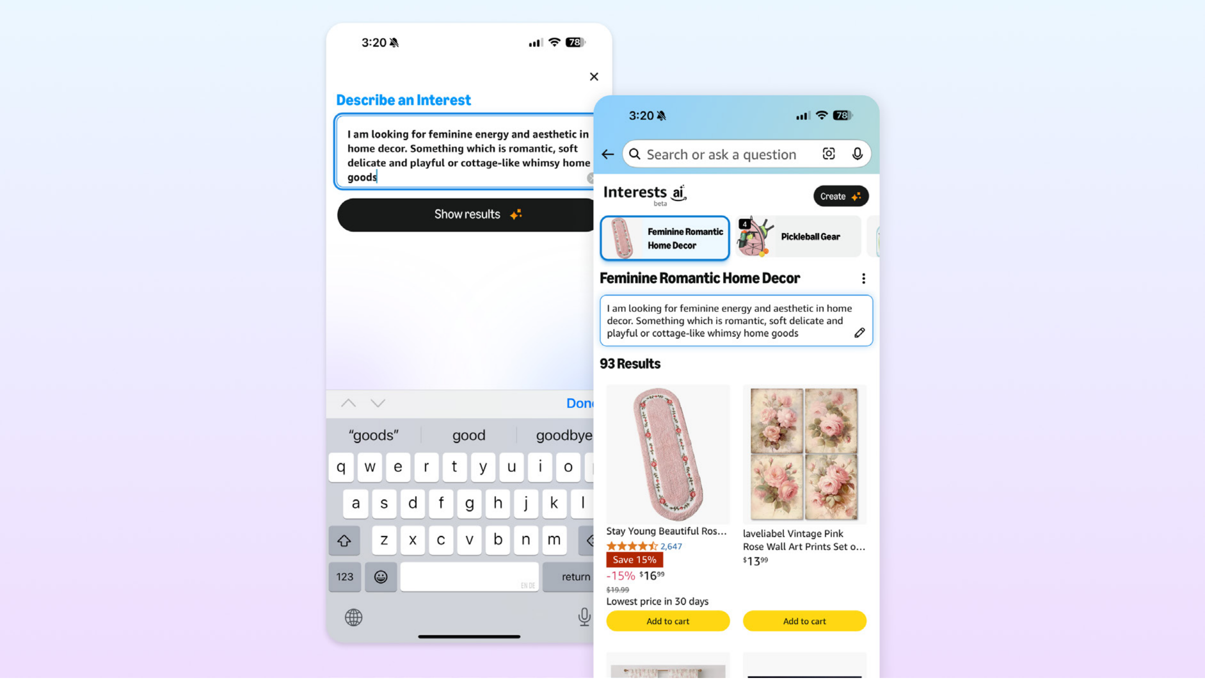
System prompts, or information that is always included in prompts even when not explicitly written in a user query, can also support personalization via model context. Developers may have a static set of system prompts that shape responses to all users, but many AI companies have added tools to let users add additional information to alter the chatbot’s behavior, tone, style, and understanding of relevant context.



What personal preferences should Claude consider in responses?

Your preferences will apply to all conversations, within Anthropic's guidelines. Learn about preferences

e.g. when learning new concepts, I find analogies particularly helpful



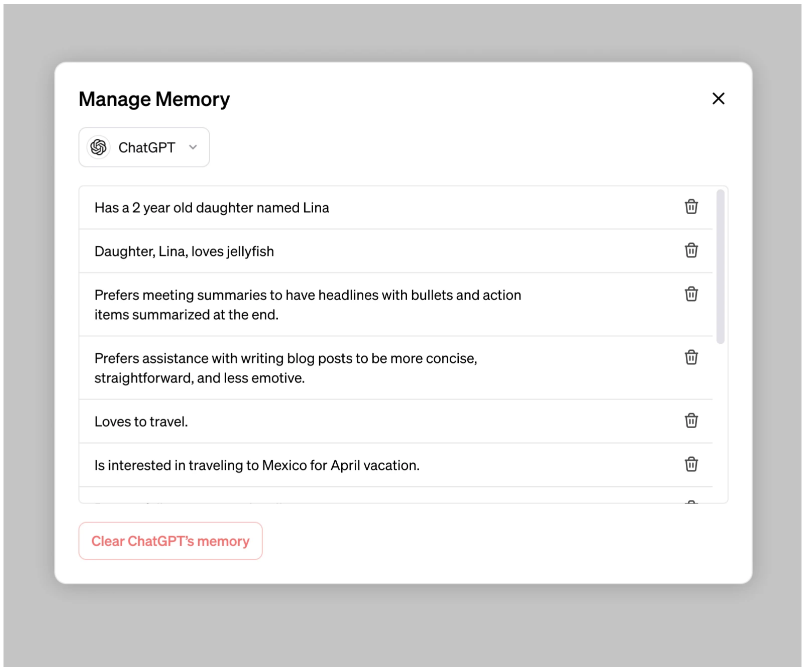
I am looking for feminine energy and aesthetic in home decor. Something which is romantic, soft delicate and playful or cottage-like whimsy home goods

I am looking for feminine energy and aesthetic in home decor. Something which is romantic, soft delicate and playful or cottage-like whimsy home goods

Figures 2 and 3. Top: Anthropic invites users to share preferences that Claude should consider when generating responses. **Bottom:** Amazon's AI-powered "Interests" tool asks users to provide personalized shopping prompts & uses those preferences to continuously search for & surface matching products.

Source: *https://www.aboutamazon.com/news/retail/artificial-intelligence-amazon-features-interest*

Some companies are integrating system prompt-like personalization directly into existing product offerings, inviting users to submit information about their interests in natural language to power product recommendations.



Has a 2 year old daughter named Lina

Daughter, Lina, loves jellyfish

Prefers meeting summaries to have headlines with bullets and action items summarized at the end.

Prefers assistance with writing blog posts to be more concise, straightforward, and less emotive.

Loves to travel.

Is interested in traveling to Mexico for April vacation.

Figure 4. OpenAI summarizes and stores details about users based on chat history, as well as explicit requests by users for the company to remember certain information, and offers an interface for users to see and delete such memories.

Source: *https://openai.com/index/memory-and-new-controls-for-chatgpt*

Long-term Memory

Storing semi-structured information in a way that can be easily retrieved in future interactions

Long-term, cross-session memory transforms AI assistant interactions from transient exchanges to more personalized experiences by preserving information across multiple conversations. Akin to a filing cabinet, this approach allows details to be summarized, stored, and drawn upon as needed. While this approach does not involve foundation models updating themselves directly to retain information, developers can enable long-term memory by storing chat logs or summarizing conversations and making those summaries available to the base model to reference in later interactions. Long-term memory can also involve extracting specific details from interactions, such as personal information like names, locations, or family members, capturing structured or semi-structured information like user preferences, or modeling interaction patterns, and persistently storing those details to call on later.

Implementation of long-term memory can range from explicit user-guided storage, where individuals directly specify what should or should not be remembered, to system-driven approaches that dynamically recognize and store key information without explicit instruction as a user interacts with their AI assistant. Some products including Gemini and ChatGPT offer settings for users to turn off retention of chat history, delete related logs, or delete specific memories.

Knowledge Bases and Retrieval-Augmented Generation

Connecting generative AI interactions with more structured databases of specific details or summaries of pertinent information

Common methods to customize AI systems involve relying on retrieval engines, or methods to make additional data available to AI systems to reference when fulfilling requests, such as a user's or organization's documents, transaction histories, or account information. These sorts of repositories are often referred to as *knowledge bases*, and the process of a generative AI system calling on them is called *retrieval-augmented generation* (RAG). An AI assistant embedded into an e-commerce platform might, for example, reference a user's purchase history to recommend complementary items, while a workplace-oriented AI assistant could leverage organizational documents to understand industry-specific concepts and communication norms. Data in knowledge bases can be combined behind the scenes with prompts, or the AI system can look up information in such databases as needed to produce a more personally relevant response. Importantly, knowledge bases need not be simple databases: Summaries of earlier conversation segments can also be stored in knowledge bases (in this case, *vector stores* or *vector databases*) so the model can reference them as it crafts subsequent responses through semantic comparison — that is, finding similarities in the meaning of words, phrases, and content — rather than keyword-based searches, which can facilitate longer-term memory.

User Modeling

Analyzing patterns of user behavior to predict future behavior and explicitly or implicitly predict user's personal details or preferences, and considering both these profiles and user prompts when generating outputs

Similar to technologies used for content recommendation or targeted advertising systems, user modeling involves summarizing user behaviors or preferences into formats that AI models can process. With user modeling, AI assistants can adapt communication styles, refine task assistance, and provide contextually relevant responses across sessions without needing to frequently refresh the underlying model or rely on information provided in prompts. For instance, an AI assistant might adjust its response complexity based on a user's professional background, learn preferred problem-solving approaches by tracking the corrections or clarifications a user makes, and proactively suggest resources aligned with a user's demonstrated or implied interests. User modeling tends to happen behind the scenes by using other AI models to summarize user behavior, map similarities and differences across users, and store those learnings in accessible formats — determining without instruction the most important features to compress into a complex string of numbers. The nature of this approach means it can be difficult to determine what features and characteristics end up being captured or emphasized, let alone make those learnings transparent to or controllable by users. A system relying on user modeling does not mean that the underlying foundation model itself is being directly updated to include personal details about users; rather, user profiles can be provided alongside prompts to the foundation models to inform the model’s output.

Static Settings

> Collecting structured and semi-structured information or system preferences from users at the account level to constrain, direct, or nudge system outputs

User-level account information and user profiles often capture broad demographic information or general settings such as preferred language and location. In many cases, users are asked to provide structured information such as sex or gender, date of birth, location, and predefined topical interests; in others, users can provide unstructured information through bios, free-form fields, or text interactions. Personalization using these user-provided details could look like taking a user’s provided age into account to block inappropriate content or shape age-appropriate responses, adjusting language or tone of responses, or influencing other content or product behavior based on user preferences.

System settings can also be used to manage data permissions, such as whether a user in a certain category of accounts can access particular information. For example, a member on an organization’s HR team may have permission to access employee information while people on other teams would not. They can also be used to define whether a user’s interactions with an AI assistant can be used to update its training set.

These settings or preferences may be appended to user prompts behind the scenes, or be used to configure more structured behavior (e.g. applying filters to generated responses) or to inform data use within an AI system’s architecture (e.g. granting access to certain fields within a database that the AI system can call).

Model Modification

> Directly changing foundation models to reflect personal preferences through methods like fine-tuning

While the foundation models that power AI assistants are trained on large datasets, and initially evaluated on their performance at general tasks, portions of these models can be further trained to reflect specific communication styles, domain expertise, and organizational or even personal needs. With this approach, developers curate datasets representing desired interaction patterns, communication norms, or specific domain vocabularies. The resulting experience more strongly mirrors the examples provided during fine-tuning, allowing interactions to be more relevant to particular contexts. Organizations can fine-tune assistants to reflect their unique brand guidelines, technical language, workflow requirements, or areas of expertise, which may be particularly attractive to deployers in fields like healthcare, finance, and law that require support for highly specialized tasks. Fine-tuning is most commonly used to customize large language models to particular segments of users or domains, rather than granular, individual preferences, but it is conceivable that fine-tuning a model on a single account's transactions, behavior, or communication history, or locally on edge devices like mobile phones or wearables, could enable personalized experiences.

Implications of Personalization in Advanced AI

Personalization can be an appealing feature of technical systems, seeming to offer more utility to users looking for easier or more customized experiences. But personalization can quickly lead to undesirable externalities: for instance, dynamically pricing goods and services purchased online based on assumed willingness or ability to pay could result in discrimination and price-gouging. Targeted ad delivery, in particular, has faced scrutiny for leaning on protected characteristics such as race, gender, sexuality, and health or disability status, either directly or by proxy, due to the potential for stereotyping and inequality. Serving users only preferred information has led many to worry about the cementing of a fragmented information landscape. When considering the implications of personalized AI systems, we encourage stakeholders to reflect on important questions that will emerge around what data is processed and why; how these systems are monetized; how to avoid the pitfalls of business models that exploit personalized experiences; potential discriminatory impacts if personalized systems lean on stereotypes; and how to prevent models from learning things they shouldn't.

1. As AI developers look to further personalize their products, they will seek to access increasingly sensitive user information.

Even if a model is not trained with personal data, private details and preferences can still be used later on once an AI system is deployed in order to power personalized experiences. And, aside from personalization based solely on data entered by users into context windows, all of the techniques described above require some degree of storing, accessing, and analyzing data about a user or their behavior. As developers of advanced AI assistants seek more comprehensive access to users' digital lives — monitoring screen activities, communication patterns, and personal interactions across devices — stakeholders should be attentive to what data is collected, stored, and used separate and apart from model training. Moreover, interactive language-based AI systems only deepen concerns that AI systems may implicitly infer, or appear to infer, sensitive information like users' emotional state, sexual orientation, or health or disability status — not least because users may expressly seek support or resources from chatbots around these topics.

As AI products mature, some have introduced more robust options for users to control whether their data is stored (and for how long), as well as how it is used, and some developers are prioritizing ways to process data on-device or in more secure cloud environments. Natural language interfaces offer an opportunity for people to more easily change privacy preferences or settings, but additional engineering effort is generally needed to ensure these instructions are followed with fidelity. Choices around data architecture and other technical infrastructure can determine whether personal data will leak into contexts it shouldn't, and whether explanations around how data is used will remain accurate.

Short of more affirmative safeguards around data processing for AI established in law or regulation, features that help users at least understand how their data is being used—and take advantage of options to control or limit such data practices—will be particularly important as AI tools are adapted to support increasingly personalized tasks. For instance, OpenAI has introduced a feature that lets users ask what ChatGPT remembers about them, and to request that certain information be modified or forgotten. But, as with many other digital contexts, awareness and control don't solve for all relevant harms that personalization can present, such as the normalization of continuous collection, bias amplification, and exploitative practices that take advantage of personal details or circumstances even if those details are not explicitly collected or disclosed. As users turn to personalized AI systems for a growing variety of tasks, controls that support specific purpose limitations will be needed to manage appropriate uses of different types of data. For example, users may disclose a health condition to a chatbot so that the system can provide answers to questions about symptoms, treatments, and risk factors, but may want to prevent the system from using that information to power targeted advertising or to personalize responses to prompts unrelated to that health condition. Importantly, placing the burden on users to manage their data in this way also wrongly assumes they have sufficient capacity and understanding to make informed choices across dozens or hundreds of digital services they may use, when in reality this approach is increasingly seen as overwhelming and ineffective.

2. Technical approaches to preventing AI systems from learning things they are not supposed to remain underdeveloped.

As AI developers exhaust existing sources of raw data to train foundation models, they are leaning on other sources of data, including user data, to further develop and improve their models. In some cases, companies are declining to let users opt out of having their data used for this purpose. Training on personal data introduces substantial privacy and ethical risks, particularly as user experiences become more individually customized: Data from personalized interaction becomes a potential vector for sensitive details to be absorbed by models and improperly exposed, creates the opportunity for stereotypes to be reinforced and cemented within foundation models that are trained or fine-tuned on that data, and makes it possible for models to improperly rely on actual or perceived immutable characteristics to influence their interactions with or recommendations to users. In addition to the difficulties of removing information from underlying foundation models, those designing privacy-preserving interventions in personalized, AI-powered products must contend with the fact that assistants and agents are designed to ingest and process unstructured data. When data is clearly structured and categorized, it's relatively straightforward to define when certain data can and can't be used. With unstructured information, though, developers must engineer and design deliberate interventions to detect and prevent more sensitive categories of data such as health conditions from being improperly used or stored by the model or related products, and to distinguish between meaningful user preference signals and automated assumptions or stereotypes.

3. Monetization efforts will incentivize increased data collection.

As personalization and monetization of AI products become more important, related choices will have strong implications for what data developers opt to collect and store by default, who has access to key services, and how aggressively AI service providers will pursue users to engage with their experiences. Many consumer AI platforms offer freemium tiers of service, but companies are actively exploring business models including subscription plans (that offer more or unlimited use, additional features, or speedier service), inserting ads into AI-powered interactions and search results, making proactive product recommendations to drive consumer purchasing, and offering enterprise and specialized services.

While subscription-based or enterprise services might appear to reduce incentives to collect and leverage personal data as compared to advertising-driven products, this assumption may be overly optimistic: Companies will still be incentivized to make their tools as "useful" as possible in order to attract and retain users and customers, which in many cases involves customizing product experiences to people's interests and needs in a manner that relies on deep knowledge of users' contexts and preferences. In the more extreme scenarios, products that aim to attract repeat users or retain subscribers month over month by making their experiences hyper-personalized can take on addictive characteristics and lead to a variety of downstream harms.

Meanwhile, companies are likely to explore approaches to advertising that don't overtly interrupt a user's experience, so we would not be surprised to see AI assistants and other tools that include opportunities for product placement (e.g. including branded content in generated images that did not explicitly ask for it), influencer-style content (e.g. systems providing in-depth knowledge of particular products or experiences), paid recommendations (e.g. brand pay for increased likelihood of being recommended over competitors), and referral schemes (e.g. AI agents getting kickbacks for making purchases on certain retail platforms).

In addition to raising familiar concerns around transparency, discrimination, and exposure to unwanted content, these potential advertising paradigms also complicate differentiation between contextual and behavioral advertising, since personalized AI systems directly shape their content — and therefore the relevant context — based on behavioral signals.

4. Personalization will almost certainly reinforce stereotypes and lead to discrimination.

Offering more personalized experiences with AI assistants to consumers has the potential to meaningfully help users by providing more contextually-relevant information, recommendations, and services — particularly for people who may not be able to access such support at all or as easily through other means. But over-personalizing or making assumptions about people’s interests and needs based on details about them can also be fraught. For instance, a personalized system could adapt explanations to accommodate different levels of fluency in English, but this sort of adaptation could also reinforce harmful assumptions about users’ interests and skills that might limit what sort of professions a system recommends to users with lower fluency, unnecessarily limiting their economic opportunity. Personalized educational experiences could support students at different levels of proficiency, particularly those who might otherwise fall behind, but models could also pick up on stereotypes and nudge students from low-income neighborhoods toward subjects and skills less likely to result in sustainable careers. A similar scenario could occur for users with disabilities who request that explanations be provided in plain language – the personalized system could adapt and provide this, which would increase the system’s accessibility, but could make similar harmful assumptions about a user’s ability or skill that could have analogous detrimental impacts on the disabled user’s economic or employment opportunity.

More work is needed from practitioners and researchers to determine what degrees of personalization are appropriate for AI assistants across the variety of scenarios and contexts in which they are being deployed, and how to prevent these foreseeable consequences. Moreover, the design, development, and deployment of fair and useful personalized AI assistants must remain grounded in the understanding that these tools are not only technical systems, but social systems requiring the expertise of user experience researchers, AI auditors, social scientists, and users ultimately impacted by personalized experiences.

Conclusion

AI assistants are poised to shift quickly from general tools or systems that may be customized for particular domains, to highly personalized products that not only rely on user-provided information but also on implicit signals — a pattern that mirrors the evolution of other online platforms from information retrieval tools to personalized recommendation systems. Such shifts have led to questionable transparency practices, unexplainable and troubling system behavior, discriminatory effects, and loss of privacy and users’ agency over how their data was collected and used. Developers should take care to avoid falling into these familiar patterns and instead use the opportunity that novel AI technologies may provide to reimagine how to balance the development of helpful personalized experiences with protections for the people they are intended to benefit.

Find more from CDT's AI Governance Lab at *cdt.org*

The ***Center for Democracy & Technology*** *(CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.*