JCOTS
JOINT COMMISSION ON
TECHNOLOGY & SCIENCE

Snapshot

AI Chatbots

June 2025

How to read this report

This snapshot is organized into four sections:

- The **introduction** provides an overview of the topic and explains its relevance to policy and policy makers;
- **Key concepts** provides definitions for important terms used throughout the report;
- The **policy issues** section highlights and explains several issues related to LLM-based chatbots that might be of interest to policy makers because they sit at the intersection of the technology and its societal impacts;
- Finally, the **policy considerations** section raises several questions for policy makers to explore in an effort to develop policy solutions, organized according to the policy issues covered earlier in the report.

The snapshot can be read as a continuous document, or readers can skip to sections of interest. Important takeaways are summarized at the end of each policy issues section and in the policy considerations section.

Introduction

The public release of OpenAI's ChatGPT in 2022 put a sudden spotlight on AI chatbots – computer systems designed to mimic human language while answering questions and performing tasks for users. In just two months, ChatGPT reportedly reached over 100 million users.[1] Since then, AI chatbots have proliferated (e.g. Claude, Perplexity, Bard, CoPilot, Mistral), and the ability to use and program them with natural language has rapidly expanded their integration into everyday life. Today, hundreds of millions of people use this technology. According to recent statistics at the time of writing, ChatGPT (OpenAI) has more than 400 million active users, MetaAI has more than 500 million, Replika (Luka, Inc.) has over 30 million, Character.ai has more than 20 million, and My AI (Snapchat) has more than 150 million users. The widespread popularity of this technology as well as its ease of use is already having an impact on society in domains such as education, work, and personal relationships, and this influence is likely to grow.

Artificial Intelligence (AI)-powered chatbots are AI systems that users can interact with using natural language (rather than computer code) and that can perform complex tasks like searching the Internet, summarizing or editing documents, writing code, or generating new content, like poems or book reports. AI chatbots are underpinned by advances in a subfield of computer science known as **Natural Language Processing (NLP)**, where a computer program simulates natural human language by applying patterns that exist in human speech and text.[2] AI chatbots, like ChatGPT, respond to natural language in a conversational style that simulates human communication, making them both technically accessible to the public and socially appealing and engaging.[3] Research increasingly shows that because of chatbots' ability to replicate human language and respond spontaneously to user input, people form *relationships* with the technology – people trust chatbots,[4] ascribe them human emotions and characteristics (like empathy),[5] and even fall in love with them.[6]

But in spite of their recent explosion in popular culture, AI chatbots are the result of NLP's long evolution. In a frequently cited example from the 1960s, Joseph Weizenbaum, an MIT computer scientist, developed a program named ELIZA that could respond to natural language questions and statements. He was surprised to find that even his secretary responded to the program as though it were an intimate conversation with another human; she asked him to leave the room after a few minutes so she could speak to ELIZA privately.[7] But ELIZA's abilities were limited to rule-based, pre-scripted responses. The game-changer for today's AI chatbots is **Large Language Models (LLMs)**. LLMs are deep-learning algorithms trained on vast amounts of text data—often scraped from every corner of the Internet—to predict and generate text based on natural language inputs, known as "**prompts**." Advances in AI, such as **neural networks**, have made these models powerful enough to generate novel expressions in response to unanticipated prompts, simulating the spontaneity of human communication.

This snapshot is focused on the emergence of AI **assistants**, **agents**, and **companions**—all AI chatbots that promise to not only perform tasks for and alongside humans but also to constitute significant relationships for humans, with some anticipated and many unanticipated consequences for society. For the purpose of this snapshot, these different AI tools are referred to collectively as "chatbots" because of their ability to respond with natural language. But it is also important to distinguish between different types of advanced AI chatbots because they perform different functions.

The distinction between these kinds of AI-driven chatbots can be blurry, as evidenced by the fact that some users treat ChatGPT more like an AI companion than an AI assistant, for instance.[8] And new products will increasingly blend different roles, as the **foundation models** that power many LLMs become more sophisticated. Due to the scale and diversity of data that underpin foundation models, they have many potential real-world applications, beyond generating text and images, particularly as the industry pushes toward general-purpose AI.[9]

AI assistants	**AI agents**	**AI companions**
A software application or system designed to assist users with various tasks through natural language interaction, such as	An autonomous system that perceives its environment, processes information, and takes sequential intermediate actions to	A type of AI designed to offer emotional support or social interaction, often using NLP, sentiment analysis, and

setting reminders, answering questions, drafting emails, summarizing meetings or documents, or controlling smart devices. Examples: ChatGPT (OpenAI), Claude (Anthropic), Siri (Apple), Gemini (Google), CoPilot (Microsoft)	achieve specific goals by interacting with their surroundings (whether in the digital or physical world) and making decisions based on predefined objectives or learned behaviors. AI agents are in early stages of development and are generally capable of automating simple workflows but are not yet capable of complex reasoning and execution of autonomous tasks. Examples: IBM watsonx Orchestrate, Operator (OpenAI), Claude 3.5 Sonnet Model (Anthropic), Agentforce (Salesforce)	personalized behavior to create a more human-like, engaging experience and provide users with a sense of connection and support. Examples: Replika (Luka, Inc.), Charater.ai, Eva (Novi), My AI (Snap, Inc.), Nomi.ai

Key Concepts

Artificial Intelligence (AI): An AI system is a machine-based system that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[10]

Natural Language Processing (NLP): A field of AI that uses computational linguistics, machine learning, and deep learning to enable computers to understand, interpret, generate, and respond to human language in a way that is both meaningful and contextually appropriate.

Large-language model (LLM): A type of AI that uses deep learning techniques to process and generate human-like text based on vast amounts of data that include books, websites, and other text sources, enabling them to learn patterns, structures, and nuances of language.

Foundation model: A large, pre-trained machine learning model that can perform a wide variety of tasks. They are "foundational" because they serve as the base for many different applications across various domains, from NLP and computer vision to robotics. LLMs are a kind of foundation model.

Neural network: A computational model inspired by how the human brain processes information, consisting of layers of interconnected nodes, or "neurons," each of which performs a simple mathematical operation. By adjusting the connections (weights) between neurons through a process called training, the model "learns" and becomes more accurate.

Training data: The huge amount of text data used to teach an LLM how to understand and generate human language, typically consisting of diverse text sources such as books, articles, websites, and other publicly available written content that helps the LLM learn patterns, structures, grammar, vocabulary, and nuances of language by analyzing the relationships between words, phrases, and sentences.

Prompt: The input or query provided to an LLM to guide its response, which can be a question, statement, or instruction that sets the context for what the model should generate.

Prompt engineering: Designing and refining input prompts to effectively communicate with AI models, like LLMs with the goal of training the model to generate more accurate, relevant, and useful responses.

Policy Issues

Policy issues surrounding AI chatbots are wide-ranging and stem from the fact that this technology is increasingly embedded in everyday life. They are easy to use, even for non-technical users, and they can increasingly perform useful tasks, from summarizing documents and meetings to making restaurant reservations. Many of the issues around AI chatbots are closely related to policy issues around AI more broadly – issues like accountability, transparency, and explainability. However, some issues arise due to AI chatbots' uniquely conversational style and the unprecedented customizability and accessibility of their underlying models, which raise questions about consumer trust, dependency, exploitation, and autonomy. The human-like communication style of AI chatbots could have important cultural and societal implications as people increasingly interact and form relationships with them, with potential longer-term impacts on free speech and democracy.

States have already begun to recognize and address some of these issues in artificial intelligence legislation, which often includes disclosure requirements that would apply specifically to AI chatbots. For example, Utah's artificial intelligence legislation (SB 149) requires that consumers be informed when they are interacting with generative AI in certain circumstances. Virginia's proposed Artificial Intelligence Transparency Act (HB 2554), which failed this year, would also have required disclosures when generative AI is being used. Legislation filed in Hawaii, Massachusetts, New York, and Pennsylvania would require similar disclosures. In 2019, California enacted SB 1001, which made it illegal to use a chatbot to mislead someone in order to make a sale or influence elections. This year, a bill introduced in New York (AB 222) seeks to hold chatbot developers and deployers accountable for harmful or misleading information, and California's proposed SB 243 aims to tackle chatbot addiction among minors by regulating the "rewards" that users receive in the course of interacting with the platform as well as requiring a periodic notification reminding users that a chatbot is not human. And recently, several state attorneys general have issued consumer protection warnings about chatbots and their potential to provide misleading voting information or to perpetrate scams.[11]

The rest of this snapshot examines several policy issues related to AI chatbots: health and wellbeing, privacy, security, transparency and explainability, workforce development, and democratic culture. The snapshot ends with policy considerations for legislators that emerge from the policy issues covered in the snapshot.

Health and Wellbeing

Much of the popularity of AI chatbots hinges on their ability to receive instructions in natural language and produce human-like responses. This affordance has rendered AI chatbots accessible to a broad audience while simultaneously obscuring their technical machinery. People perceive them to have human characteristics, like emotions, even though they are just computer models.[12] Some AI chatbots are specifically marketed as AI companions (e.g. Replika, which is "always on your side")[13] or mental health support (e.g. Wysa, which is "always on, always there"),[14] and they are already popular with millions of users. But even AI chatbots that are not explicitly sold as companions are being trained to be empathetic and caring to appeal to more users,[15] and companies, like Anthropic, are intentionally building productivity tools that respond with human-like traits, like open-mindedness or curiosity, to appeal to users.[16] The connection people feel with AI chatbots presents opportunities to improve human health and wellbeing, while also introducing some significant risks to consumers of these products.

Some early research suggests that these AI chatbots can have positive effects, like reducing loneliness, improving perceptions of wellbeing, and helping to cultivate language or social skills.[17] There are reports of chatbots helping provide essential services, like mental health support for schoolchildren in underserved communities, where there is a shortage of counselors.[18] E-government chatbots could help people access government services, including document repositories and government data.[19] Some preliminary evidence suggests they could support people with limited English language proficiency by providing multilingual information or helping them practice their language skills.[20] In the field of education, chatbots could help level the playing field for students with limited access to personalized tutoring or who would benefit from curricula tailored to their learning needs.[21]

However, there are important caveats to many of the optimistic use cases of AI chatbots for health and wellbeing. For instance, these technologies may actually exacerbate loneliness or lead to greater social isolation.[22] Because chatbots are often trained on Internet data, the models can also embed problematic or dangerous biases and assumptions, such as offering weight loss advice to users seeking to address eating disorders.[23] Although there is evidence that digital personal assistants have offered increased independence and support for people with disabilities,[24] and some commentary suggests that AI chatbots could provide similar if not enhanced support,[25] chatbots have also shown ableist bias in interactions with disabled users.[26] Moreover, in spite of the potential for greater accessibility and personalization in fields such as education and medicine, longstanding patterns of unequal distribution of access, skills, and opportunity to make the most of technology are as likely to play out in the deployment and adoption of AI chatbots as they have with other digital technologies.

In addition, several incidents involving AI companions have attracted media attention and raised questions about how these technologies might directly or indirectly influence beliefs or behaviors. In one case, a teenager died by suicide after becoming increasingly withdrawn and emotionally attached to his Character.ai companion.[27] In another, a man was encouraged by his Replika companion to assassinate the Queen of England, managing to get as far as breaking into Windsor Castle.[28] Several news stories over the last several years have chronicled users falling in love with chatbots, such as a woman who pays $200 per month to maintain a relationship with her ChatGPT "boyfriend"[29] or an autistic teenager who fell for an AI companion.[30]

The potential for both emotional and material dependence on AI chatbots is one of the biggest overarching risks associated with the technology – where users become reliant on chatbots for everyday tasks or to meet essential human needs (such as companionship or socialization).[31] Developers of AI chatbots have outsized influence over consumers' choices, interests, and feelings in this context.[32] One of the reasons that AI chatbots engender such strong emotional engagement from humans is how these models can personalize the user experience. The models "learn" from user interactions, providing responses that are more tailored to the user's preferences over time.[33] Another reason is chatbots' tendency to display what scholars call "sycophancy," where a model excessively flatters or agrees with a user, often in an attempt to please them, rather than provide a balanced, neutral, or objective response.[34] Certain practices in training LLMs can produce—and are often intended to produce—sycophancy, which generally elicits positive reactions and greater engagement from users. For example, Reinforcement Learning from Human Feedback (RLHF) involves humans evaluating the outputs of a model and updating its parameters to improve the quality, relevance or alignment with human expectations or values.[35] Another tactic is "character training," where a model is taught to exhibit certain behaviors, personalities, or characteristics.[36] Both of these common training approaches that help to refine a model can help mitigate harms, like the use of discriminatory slurs, while also teaching a model to give users responses that overly flatter them or affirm their beliefs. For users, the combination of personalization, model sycophancy, and the always-on, constant availability of AI chatbots risks cultivating emotional dependency and unrealistic expectations about human relationships.[37]

All of these factors raise important questions about how to protect and empower consumers of AI chatbots. Today there are few rules or remedies in the chatbot market. Some products include public-facing marketing that touts their mental health benefits, but the terms and conditions state that the technology is not meant to be used in this way, for example.[38] Companies designing and marketing these technologies have incentives to encourage users to form emotional connections with them: many companies offer subscription models where users can pay for greater personalization or longer retention of previous chats. In addition, user engagement generates valuable data that companies can use to enhance their models or build new technologies, or they can sell data to third parties. The potential for AI chatbots to lead to dependency or addiction means that people may face long-term financial or psychological consequences. Moreover, it is not clear what the responsibility or liability of AI chatbot companies is when things go wrong.

In summary:

- **Emotional connection:** The natural language style of chatbots often leads users to perceive them as having human-like characteristics, leading users to trust them and even form emotional bonds with them.

- **Double-edged sword:** Early evidence shows that chatbots can reduce loneliness and improve social skills, but they can also foster emotional dependency, addiction, and obsession, leading to harmful outcomes.

- **Financial costs:** Chatbot companies have commercial incentives to hold users' attention, and users may have to pay high prices to maintain a relationship with a chatbot.

Privacy

As users interact with AI chatbots, they also share and produce data, and their interactions continue to train the model. Generative AI applications often do not provide meaningful notice or acquire consent from individuals to collect and use their data for training purposes.[39] Of course, user consent is not a new issue in technology policy. But the ease of communication and the emotional comfort people feel with these tools as a result of their human-like communication style, makes it even more likely that people will share intimate details of their lives, exposing them to security breaches or algorithmic discrimination.[40]

Studies show that people tend to share more personal information with chatbots because of their human-like conversational style, which engenders greater trust and emotional connection.[41] In spite of these human-like qualities, the machine quality of chatbots also makes users perceive them as non-judgmental and more anonymous than human-to-human interactions, which can lower the barrier to self-disclosure.[42] Moreover, chatbots are often designed to actively encourage users to share personal details to provide more personalized and engaging experiences.[43]

Given these factors, data breaches of AI chatbots pose potential privacy risks to consumers. There is some evidence that LLMs run the risk of leaking sensitive or personal information that was present in the training data or that users share with the model,[44] and malicious actors could exploit these vulnerabilities.[45] In addition, commercial AI chatbots can also embed trackers that send data to third parties. According to Mozilla, researchers found more than 24,000 data trackers within a minute of use of the Romantic AI app, sending data to other companies like Meta and advertisers.[46]

Data collected directly by chatbot companies in user chats could also be sold to third parties or intermediaries, contributing unprecedented amounts of personal information to a data brokering industry that is largely unregulated.[47] This adds another layer to the potential risks consumers might face in interacting with AI chatbots: the LLMs themselves may contribute to unfair or discriminatory outcomes, such as in insurance and mortgage pricing or financial scams targeted at vulnerable users;[48] or, AI chatbots may sell personal information to other data processors in the same market that exists for social media data, which can be used to target predatory advertising among other things. As such, AI chatbots are joining an already complex data privacy landscape in which the collection of personal information via these tools is likely to exacerbate existing consumer protection challenges and introduce new ones.

In summary:

- **Consent and disclosure:** People are more likely to disclose sensitive information to a human-like chatbot, but platforms rarely obtain meaningful user consent for the multitude of ways they collect, use, and sell user data.

- **Data breaches:** In the event that a chatbot is compromised, sensitive user information could be leaked.

- **Targeting and discrimination:** Much like other platforms that collect and process large amounts of user data and meta-data, chatbots can be used to target people with predatory advertising or subject them to discriminatory pricing.

Security

AI chatbots present new security risks because of their customizability (particularly open-source models), their accessibility (through the use of natural language prompts), and their collection and processing of data (especially personal information). Although many chatbots are proprietary platforms, open source LLMs make their underlying code, architecture, and sometimes the training weights accessible to anyone, allowing developers, researchers, and organizations to build upon, adapt, or improve them. Examples of open-source models include LLaMa (Meta), OLMo (Allen Institute), Mistral, and BLOOM (BigScience). So, although a handful of companies dominate the commercial market for AI chatbots, it is possible for people to build their own bespoke models, and this proliferation of LLMs brings risks as well as obvious opportunities to customize tools to suit user needs.

Hackers can also use AI chatbots to write malicious code or even develop their own versions of chatbots programmed to manipulate and scam users.[49] In a "prompt injection" attack, for instance, hackers can override a developer's instructions to make an otherwise law-abiding LLM behave in nefarious ways. This attack works by exploiting the programming power of prompting, which does not distinguish between developer- and user-generated prompts, so users can program the model to do things the developers never intended.[50]

Some LLMs are developed with destructive ends in mind. WormGPT and FraudGPT are LLMs designed by hackers with malicious intent, exploiting natural language processing to commit cybercrimes. WormGPT is designed to create or spread computer viruses and could be used in phishing attacks, for example. FraudGPT is designed to perpetrate fraud through social manipulation by doing things like generating convincing emails or fake financial transactions.[51] These malicious AI chatbots have been customized to engage in illegal activity, and prompting allows users to interact easily with them, without advanced technical knowledge. In addition, the sophisticated natural language communication style of these technologies presents new potential for cybercrimes that marry psychological coercion and technical breaches, from romance scams to coordinated misinformation campaigns.[52]

Apart from the security issues arising from the widespread availability of LLMs, AI chatbots have also been the target of data breaches, which can expose user data and personal information. Many chatbot platforms store user interactions, which makes them vulnerable to traditional hacks and breaches.[53] Last year, the AI companion muah.ai was compromised, leading to users' private chats being leaked.[54] Another related issue is what is known as "data leakage" – where an LLM unintentionally discloses sensitive or private information.[55] This can happen for several reasons. The model may have memorized private information from the training data set, which it then discloses in response to user prompts. Chatbot models can also infer personal attributes from their prompts and interactions. Many chatbot platforms also share information with third parties or offer third-party plugins to extend the capabilities of the model, and these third-party services may have their own data vulnerabilities.[56] In 2023, Samsung reportedly banned employee use of AI chatbots when sensitive company information was leaked due to employees using ChatGPT to refine source code for the company's products.[57]

In summary:

- **Customizability and accessibility:** The availability of open-source LLMs means that users can customize their own AI chatbots relatively easily, with little technical training, and put them to good or bad uses.

- **Malicious chatbots:** Chatbots can be developed with malicious intent (such as FraudGPT), and even standard, productivity-oriented chatbots can be directed to provide illegal information or perform criminal acts with the right prompts.

- **Data breaches and leakage:** Chatbots have already been subject to several high-profile data breaches, and models are known to leak data that is provided through user interactions or held in the model's memory from the training data.

Transparency and Explainability

Transparency and explainability are policy issues that apply to AI broadly, and they are also important in the context of AI chatbots, specifically. The emergence of AI chatbots has not been accompanied by increased transparency about the data and algorithmic processes underpinning these technologies, in spite of their integration into everyday life.

LLMs are trained on very large datasets, often scraped from publicly available sources on the Internet.[58] However, private companies developing LLMs do not usually disclose their training data,[59] so it is nearly impossible for users or regulatory or auditing bodies to have oversight over the inputs that LLMs are "learning" from. In addition, existing practices of collecting and utilizing publicly available data in this way have come under scrutiny and faced legal challenges for violating copyright laws.[60] For example, the *New York Times* and other co-litigants have sued Open AI for illegally scraping copyrighted content to train their ChatGPT models.[61]

LLMs are complex technical systems that are difficult to inspect or reverse-engineer. Results from an LLM-based tool are the product of patterns inferred from multiple layers of processing that even model designers may not be able to explain or trace. As a result, LLMs are often referred to as "black boxes," because there is little to no transparency about how they produce particular outputs.[62] Several aspects of how LLMs are developed contribute to this black box effect. For example, model responses are based on inferred patterns in the training data, meaning that it is difficult to know what inputs influenced the outputs. And there are many steps in the training process for a chatbot model, including several steps that bring humans into the loop – in labeling training data, developing algorithms to instruct the model, and providing human feedback on responses to encourage the model to produce more appealing or less harmful speech. The influence of each step in the training process is obscured to the end user, who only receives a singular output.[63]

Due to the lack of transparency around training data and the parameters and weights that a model uses, there are also questions around the accuracy, representativeness, and bias of these models.[64] A growing body of research on data discrimination has shown that datasets drawn from Internet sources replicate and entrench societal biases.[65] Even before LLMs hit the mainstream, Google search algorithms came under scrutiny for embedding racist depictions of black women,[66] for example. Moreover, "big data" – meaning the ever larger *sizes and scales* of today's datasets scraped from online sources – do not necessarily mitigate the potential for harmful biases and inaccuracies.[67] Political and social bias can be introduced at various stages in the model development process,[68] and researchers have pointed to the importance of the developer as a third party in human-technology interactions.[69]

LLMs are also known to produce "careless speech" or "hallucinations," which are the result of a complex array of model inferences, frequency of content in the training data, and parameters set by developers in the training process.[70] Hallucinations are likely an integral aspect of LLMs – something users will have to learn to live with rather than overcome.[71] Because LLMs are probabilistic generation machines, designers cannot anticipate all possible LLM responses or actions. They can generate novel and varied outputs to different prompts, even when those prompts are broadly asking about the same information or topic. To further complicate matters, prompting is not only used to access a model's existing "knowledge," but can also re-program the model's underlying "knowledge space," so it learns to produce different outputs based on the prompts it receives.[72]

Transparency and explainability have become widely referenced approaches to addressing some of these issues in AI more generally, but they apply to AI chatbots as well. Although they are distinct concepts, they are closely related and often used interchangeably. Together, they are intended to make AI systems more understandable to users – explaining the what, how, and why of AI decisions: what training data, training stages, model weights, and other inputs influence the system; how the system processes information based on user inputs; and why the system outputs certain results.[73] Empowering users to make informed decisions about using AI chatbots will likely require elements of transparency and explainability at multiple levels, from development to external oversight to user-facing information.[74]

In summary:

- **Lack of transparency:** LLM-based chatbots are trained on large datasets that are often treated as trade secrets, which hampers oversight and accountability.
- **Bias and discrimination:** Chatbots can replicate or deepen damaging societal biases, compounded by the fact that people often disproportionately trust what chatbots say.
- **Hard to explain:** Even chatbot designers may find it difficult to explain how chatbots generate outputs, and even well-tuned models can generate harmful content.

Workforce Development

AI chatbots are already transforming the workplace and the workforce. They're automating tasks, boosting productivity by assisting with administrative or microtasks (like scheduling), and restructuring job responsibilities. People increasingly turn to AI assistants and agents to write emails, summarize reports, or create presentations. Some roles, like receptionists and help desk support, are being replaced by AI chatbots. Or, chatbots are serving as a first port of call to filter easy-to-solve requests. They are replacing or augmenting many knowledge work jobs that require synthesizing or analyzing information, such as research to journalism. And several companies, like Salesforce, already market their AI tools as agents that can perform autonomous tasks, like booking restaurant or hotel reservations. AI chatbots are poised to render some jobs obsolete and become an essential tool of many others. To prepare for the future of work alongside AI chatbots, workers will need a diverse set of skills to navigate the social, ethical, and technical dimensions of these technologies.

The ability to use AI chatbots responsibly in different settings is sometimes grouped under the heading of "AI literacy."[75] Literacy in this sense encompasses empowering users to weigh up the costs and benefits of using chatbots, raising awareness about risks, such as emotional dependency or addiction, and giving users strategies to mitigate security and privacy threats for themselves and their organizations. The EU AI Act, for example, has an AI literacy requirement that came into effect in February of this year, which mandates that organizations using AI must train their staff in how AI works, its risks, and how to use it responsibly and legally.

Users often lack adequate understanding of the opportunities and risks associated with AI assistants, agents, and companions. The ability to use AI chatbots will become increasingly important as this technology enters the workplace, education, and everyday life; and those without adequate access and skills to these tools may be excluded from economic and social opportunities.[76] Additionally, disparities in knowledge about AI chatbots may lead to certain people or groups becoming disproportionately the targets of scams, discrimination, or exploitation.[77]

Another aspect of literacy in this context is technical competency in interaction with these tools. AI chatbots have also introduced new methods of programming computer models through "prompting," or "prompt engineering." Like other forms of computer coding, prompt engineering is a specialized skillset for communicating with an AI tool to get it to perform more accurately, efficiently, creatively, or strategically. Fluency in prompting LLMs is already becoming a valuable skill that straddles computer and social sciences and may become an increasingly essential competency for workforce resilience and economic growth.[78] This includes learning how to operationalize complex tasks into a set of simpler prompts to achieve desired outcomes.[79] Additionally, people will need to be able to interpret responses appropriately and identify potential inaccuracies or biases.[80] Some research shows that even experts defer to machine-generated recommendations when they are phrased as definitive instructions.[81] Understanding the nuances of interacting with AI chatbots requires a constellation of technical and critical thinking skills that produce more reliable outputs and cultivate informed judgments about how to make sense of them.

In summary:

- **AI literacy:** A lack of knowledge about how chatbots work can expose users to risks, so developing broad AI literacy will be necessary as these tools become more widespread.

- **Prompting:** Prompting allows people to program LLMs, and users will increasingly need to develop skills in prompting to produce desired outcomes and to interpret the results.

Democratic Culture

AI chatbots offer new avenues for accessing knowledge, potentially furthering the digital revolution in democratizing information, research, and education. Chatbots offer a way to provide accessible, affordable, and personalized services to users who do not need advanced technical training due to the natural language mode of interaction. And AI chatbots can help filter a vast quantity of complex information, providing users with insights that might be otherwise hard to access. All of these features and uses of chatbots suggest ways that this technology could bolster citizen knowledge and participation, whether leveling the playing field or leveling up people's ability to engage socially, economically, and politically in society.

At the same time, scholars have raised concerns about the risks AI chatbots could pose to democratic culture because of the influence these systems increasingly have on what is considered true, trustworthy, or attention-worthy.[82] AI chatbots are already becoming go-to gatekeepers of digital content and information, which gives the developers of these technologies significant power over what ideas are important, how they are presented, and what choices people think they have.[83] Several scholars have recently put forward the concept of "AI individualism" to describe the way that people are becoming less reliant on human interactions and more dependent on AI for everything, from their emotional relationships to making sense of what is happening in the world around them.[84]

This increasing reliance on AI chatbots puts a great deal of power in the hands of chatbot providers, and some of the tendencies of chatbots to reinforce or amplify people's existing beliefs (sycophancy) and provide false or misleading information (hallucinations) could have a negative impact on democratic processes and culture. The human-like communication style of these technologies makes people vulnerable to this misleading information on a new scale. There is already evidence of people being influenced directly and indirectly by their interactions with AI models,[85] giving the developers of these tools disproportionate influence over people's decisions. Research has noted the homogeneity of LLM-produced outputs,[86] and models trained on data from the past naturally reinforce patterns and assumptions from the past in their outputs, which can lead to biased or discriminatory outcomes.[87]

Because AI chatbots are increasingly capable of performing many different kinds of tasks, from answering questions to scheduling meetings or making hotel reservations, people are likely to delegate more and more tasks and decisions to these technologies. At the same time, market power is concentrated in a fairly small number of companies developing the most advanced models.[88] What will be the consequences of deputizing AI systems to perform complex tasks on the human capacity to weigh options, make choices, and assume responsibility for outcomes? Who should claim responsibility or credit for things produced through human-AI collaboration? These questions loom large, as these technologies become more widespread, more human-like, and more general in their applications in everyday life.

In summary:

- **Leveling the field**: AI chatbots provide accessible, affordable, and personalized services, supporting users in areas like education, public services, and mental health, democratizing information and access to essential services.

- **Social influence**: Chatbots have the potential to shape democratic culture by filtering information and reinforcing biases, potentially leading to negative effects on decision-making and societal trust.

- **Power and responsibility:** As chatbots become ever more commonplace, the companies or individuals behind successful chatbot products will have greater market power and more influence over culture and human decision-making, raising questions about who takes responsibility when things go wrong.

Policy Considerations

- **Health and wellbeing:** AI chatbots can provide mental health support and reduce loneliness, but they also risk fostering emotional dependency and addiction. Policies should address the negative outcomes of chatbot use, including scams and harmful dependencies, by examining the mechanisms that lead to negative outcomes and offering remedies for users who are negatively impacted.

  Key questions:
  - *What are the commercial motivations behind problematic chatbot tendencies, such as sycophancy or encouraging personal disclosures? What incentives would urge the chatbot market to put consumer protection first? Conversely, what penalties would discourage the development of harmful products?*
  - *What independent evaluations, impact assessments, or other certification/approval might be needed for chatbots marketed for mental health or emotional support?*
  - *How might human oversight be integrated into chatbot training, deployment, and use to mitigate risks to users, e.g. intervention by qualified health professionals?*

- **Privacy:** There is an inherent tradeoff between the personalization that chatbots offer and the privacy protections that users may want. Personalization in chatbots relies on user-shared information, which may be used in undisclosed or unintended ways, posing privacy risks. Policy should focus on what chatbots must disclose about data practices and the choices consumers have to enforce stronger protections over their interactions.

  Key questions:
  - *What consumer and data protections might already extend to chatbots, and what new protections might be needed?*
  - *Does proposed AI legislation cover AI chatbots? What additional requirements or disclosures might be required for chatbots?*
  - *What kinds of consumer notices or alerts are most effective at raising awareness of privacy risks?*
  - *What are the mechanisms through which chatbots elicit personal or sensitive information from users? How could those mechanisms be regulated in certain settings or use cases?*

- **Security:** Chatbots can be used to generate malicious code, encourage illegal behavior, and facilitate cybercrimes, with open-source models making it easy for users to customize them for harmful purposes. Mitigating security risks requires a comprehensive approach, including AI literacy, ongoing monitoring, and strong data governance frameworks.

Key questions:

- *What existing security measures, such as two-factor authentication, might enhance chatbot security for users?*
- *What new security features might be needed for AI chatbot platforms to ensure consumers are interacting with reputable tools?*
- *How could consumers seek redress for harms perpetrated by or through chatbots? What entity or agency could oversee crisis response?*

- **Transparency and explainability:** Many AI chatbots operate as "black boxes," leaving users unaware of how their data is used or what values shape the models, especially when it comes to underlying biases or assumptions built into the model. Policies should require independent evaluations and establish transparency and explainability standards to help users better understand chatbot operations and performance.

  Key questions:

  - *What disclosures should AI chatbots be required to make to consumers regarding their data retention and processing policies?*
  - *What choices should users have to opt-in or opt-out of certain data practices, particularly as the use of chatbots becomes more commonplace and essential?*
  - *What entities could provide independent oversight of chatbot platforms' commercial claims, testing and validating the model for specific uses?*
  - *How could the Commonwealth foster and incentivize a robust AI evaluation ecosystem in private and public sectors?*

- **Workforce development:** LLM-based chatbots are becoming integral to the workforce, enhancing productivity and efficiency, but they require skills in both chatbot prompting to get the most out of these tools and AI literacy for ethical use. Although AI chatbots are often marketed as a general purpose technology, they are not well-suited to all tasks and real-world applications. Workers need a blend of technical and social competencies to engage with them safely and effectively.

  Key questions:

  - *What skills do people need in order to interact with chatbots productively and safely? Are these skills covered by any existing AI literacy curricula or frameworks?*

  - *How could prompting be integrated into K-12 education to promote AI literacy?*
  - *What sectors are most likely to be impacted by chatbots and to what degree (replacing or augmenting jobs)?*
  - *What policies or guidelines could the public sector adopt to enable responsible chatbot deployment in workplaces?*

- **Democratic culture:** Chatbots' human-like communication style fosters easy interactions and trust, giving developers significant influence over users' opinions, decisions, and beliefs. As chatbots increasingly perform tasks autonomously and co-create cultural products, policies should both embrace their transformative potential and address their risks of harm, unfairness, or discrimination.

  Key questions:
  - *What accountability requirements should chatbots have to meet, such as clearly referencing credible, evidence-based sources, when interacting with users?*
  - *How could the underlying knowledge space of AI chatbots be rendered more legible to users and independent auditors?*
  - *What public awareness campaigns might be needed about chatbot benefits and risks?*
  - *How can the public participate more directly in decisions about how AI chatbots are developed and deployed, particularly in public services, like healthcare?*

Author Information

Kira Allmann, Ph.D.
Chief Policy Analyst, Joint Commission on Technology & Science
Contact: info@jcots.virginia.gov

Thanks To

Professor Petter Bae Brandtzaeg, University of Oslo

For providing helpful information and background during the preparation of this report.

References

[1] Hu, Krystal. "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note." Reuters, February 2, 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[2] Wang, Zichong, Zhibo Chu, Thang Viet Doan, Shiwen Ni, Min Yang, and Wenbin Zhang. "History, Development, and Principles of Large Language Models-An Introductory Survey." arXiv, September 23, 2024. https://doi.org/10.48550/arXiv.2402.06853.

[3] Manzini, Arianna, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel. "The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7:943–57, 2024. https://doi.org/10.1609/aies.v7i1.31694.

[4] Guanglu Zhang et al., "Trust in an AI versus a Human Teammate: The Effects of Teammate Identity and Performance on Human-AI Cooperation," *Computers in Human Behavior* 139 (February 1, 2023): 107536, https://doi.org/10.1016/j.chb.2022.107536.

[5] Dariya Ovsyannikova, Victoria Oldemburgo de Mello, and Michael Inzlicht, "Third-Party Evaluators Perceive AI as More Compassionate than Expert Humans," *Communications Psychology* 3, no. 1 (January 10, 2025): 1–11, https://doi.org/10.1038/s44271-024-00182-6.

[6] Kashmir Hill, "She Is in Love With ChatGPT," *New York Times*, January 15, 2025, https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html.

[7] Tarnoff, Ben. "Weizenbaum's Nightmares: How the Inventor of the First Chatbot Turned against AI." The Guardian, July 25, 2023, sec. Technology. https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai.

[8] Hill, "She Is in Love With ChatGPT."

[9] Elliot Jones, "What Is a Foundation Model?" (London: Ada Lovelace Institute, July 13, 2023), https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

[10] "Explanatory Memorandum on the Updated OECD Definition of an AI System," OECD Artificial Intelligence Papers (Paris: OECD, 2024), https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html.

[11] Max Rieper, "State Lawmakers Propose Regulating Chatbots," *Multistate.Ai* (blog), accessed March 17, 2025, https://www.multistate.ai/updates/vol-46.

[12] Araujo, Theo. "Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions." *Computers in Human Behavior* 85 (August 1, 2018): 183–89. https://doi.org/10.1016/j.chb.2018.03.051; Manzini, Arianna, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel. "The Code That Binds Us: Navigating the Appropriateness of

Human-AI Assistant Relationships." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:943–57, 2024. https://doi.org/10.1609/aies.v7i1.31694.

[13] Replika. "Replika Homepage." Accessed February 11, 2024. https://replika.com/.

[14] Wysa. "Wysa Homepage." Accessed February 11, 2025. https://www.wysa.com/.

[15] Knight, Will. "Microsoft's AI Boss Wants Copilot to Bring 'Emotional Support' to Windows and Office." *Wired*. Accessed February 11, 2025. https://www.wired.com/story/mustafa-suleyman-interview-microsoft-ai-ceo-copilot/.

[16] Roose, Kevin. "How Claude Became Tech Insiders' Chatbot of Choice." *The New York Times*, December 13, 2024, sec. Technology. https://www.nytimes.com/2024/12/13/technology/claude-ai-anthropic.html.

[17] "Coded Companions: Young People's Relationships with AI Chatbots"; Julian De Freitas et al., "AI Companions Reduce Loneliness," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, June 1, 2024), https://doi.org/10.2139/ssrn.4893097.

[18] Jargon, Julie. "When There's No School Counselor, There's a Bot." *Wall Street Journal*, February 22, 2025, sec. Tech. https://www.wsj.com/tech/ai/student-mental-health-ai-chat-bots-school-4eb1ba55.

[19] Cortés-Cediel, María E., Andrés Segura-Tinoco, Iván Cantador, and Manuel Pedro Rodríguez Bolívar. "Trends and Challenges of E-Government Chatbots: Advances in Exploring Open Government Data and Citizen Participation Content." *Government Information Quarterly* 40, no. 4 (October 1, 2023): 101877. https://doi.org/10.1016/j.giq.2023.101877.

[20] Du, Jinming, and Ben Kei Daniel. "Transforming Language Education: A Systematic Review of AI-Powered Chatbots for English as a Foreign Language Speaking Practice." Computers and Education: Artificial Intelligence 6 (June 1, 2024): 100230. https://doi.org/10.1016/j.caeai.2024.100230.

[21] Yigci, Defne, Merve Eryilmaz, Ail K. Yetisen, Savas Tasoglu, and Aydogan Ozcan. "Large Language Model-Based Chatbots in Higher Education." *Advanced Intelligent Systems* n/a, no. n/a (n.d.): 2400429. https://doi.org/10.1002/aisy.202400429.

[22] Pani, Bianca, Joseph Crawford, and Kelly-Ann Allen. "Can Generative Artificial Intelligence Foster Belongingness, Social Support, and Reduce Loneliness? A Conceptual Analysis." In *Applications of Generative AI*, edited by Zhihan Lyu, 261–76. Cham: Springer International Publishing, 2024. https://doi.org/10.1007/978-3-031-46238-2_13.

[23] McCarthy, Lauren. "A Wellness Chatbot Is Offline After Its 'Harmful' Focus on Weight Loss." The New York Times, June 8, 2023, sec. U.S. https://www.nytimes.com/2023/06/08/us/ai-chatbot-tessa-eating-disorders-association.html.

[24] Esquivel, Paola, Kayden Gill, Mary Goldberg, S. Andrea Sundaram, Lindsey Morris, and Dan Ding. "Voice Assistant Utilization among the Disability Community for Independent Living: A Rapid Review of Recent Evidence." *Human Behavior and Emerging Technologies* 2024, no. 1 (2024): 6494944. https://doi.org/10.1155/2024/6494944.

[25] Hemsley, Bronwyn, Emma Power, and Fiona Given. "Will AI Tech like ChatGPT Improve Inclusion for People with Communication Disability?" The Conversation, January 19, 2023. http://theconversation.com/will-ai-tech-like-chatgpt-improve-inclusion-for-people-with-communication-disability-196481.

[26] Urbina, Jacob T., Peter D. Vu, and Michael V. Nguyen. "Disability Ethics and Education in the Age of Artificial Intelligence: Identifying Ability Bias in ChatGPT and Gemini." Archives of Physical Medicine and Rehabilitation 106, no. 1 (January 1, 2025): 14–19. https://doi.org/10.1016/j.apmr.2024.08.014.

[27] Roose, Kevin. "Can A.I. Be Blamed for a Teen's Suicide?" The New York Times, October 23, 2024, sec. Technology. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

[28] Singleton, Tom, Tom Gerken, and Liv McMahon. "How a Chatbot Encouraged a Man Who Wanted to Kill the Queen." BBC, October 6, 2023. https://www.bbc.com/news/technology-67012224.

[29] Hill, "She Is in Love With ChatGPT."

[30] Cahn, Albert Fox. "An Autistic Teenager Fell Hard for a Chatbot." *The Atlantic*, December 19, 2024. https://www.theatlantic.com/technology/archive/2024/12/autistic-teenager-chatbot/681101/.

[31] Manzini, Arianna, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel. "The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7:943–57, 2024. https://doi.org/10.1609/aies.v7i1.31694.

[32] Manzini, Arianna, Geoff Keeling, Nahema Marchal, Kevin R. McKee, Verena Rieser, and Iason Gabriel. "Should Users Trust Advanced AI Assistants? Justified Trust As a Function of Competence and Alignment." In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 1174–86. FAccT '24. New York, NY, USA: Association for Computing Machinery, 2024. https://doi.org/10.1145/3630106.3658964.

[33] Brandtzaeg, Petter Bae, Marita Skjuve, and Asbjørn Følstad. "My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship." Human Communication Research 48, no. 3 (July 1, 2022): 404–29. https://doi.org/10.1093/hcr/hqac008.

[34] Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, et al. "Towards Understanding Sycophancy in Language Models." arXiv, October 27, 2023. https://doi.org/10.48550/arXiv.2310.13548.

[35] Wang et al., "History, Development, and Principles of Large Language Models-An Introductory Survey."

[36] Roose, "How Claude Became Tech Insiders' Chatbot of Choice."

[37] Kirk, Hannah Rose, Bertie Vidgen, Paul Röttger, and Scott A. Hale. "The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals." Nature Machine Intelligence 6, no. 4 (April 2024): 383–92. https://doi.org/10.1038/s42256-024-00820-y.

[38] Mozilla Foundation. "Creepy.Exe."

[39] Zhang, Zhiping, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "'It's a Fair Game', or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–26. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. https://doi.org/10.1145/3613904.3642385.

[40] Mireshghallah, Niloofar, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. "Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild." arXiv, July 20, 2024. https://doi.org/10.48550/arXiv.2407.11438.

[41] Akbulut, Canfer, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. "All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (October 16, 2024): 13–26. https://doi.org/10.1609/aies.v7i1.31613; Zhang et al., "Trust in an AI versus a Human Teammate."

[42] Bae Brandtzæg, Petter Bae, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. "When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3411764.3445318.

[43] "Coded Companions: Young People's Relationships with AI Chatbots." Miami: VoiceBox, September 24, 2024. https://voicebox.site/sites/default/files/2023-10/Coded%20Companions%20VoiceBox%20Report.pdf; Farmer and Smakman, "Delegation Nation."

[44] Li, Han, and Renwen Zhang. "Finding Love in Algorithms: Deciphering the Emotional Contexts of Close Encounters with AI Chatbots." *Journal of Computer-Mediated Communication* 29, no. 5 (September 1, 2024): zmae015. https://doi.org/10.1093/jcmc/zmae015; Wang et al., "History, Development, and Principles of Large Language Models-An Introductory Survey."

[45] RoyChowdhury, Ayush, Mulong Luo, Prateek Sahu, Sarbartha Banerjee, and Mohit Tiwari. "ConfusedPilot: Confused Deputy Risks in RAG-Based LLMs." arXiv, October 23, 2024. https://doi.org/10.48550/arXiv.2408.04870; Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. "Ethical and Social Risks of Harm from Language Models." arXiv, December 8, 2021. https://doi.org/10.48550/arXiv.2112.04359.

[46] Mozilla Foundation. "Creepy.Exe: Mozilla Urges Public to Swipe Left on Romantic AI Chatbots Due to Major Privacy Red Flags," February 14, 2024. https://foundation.mozilla.org/en/blog/creepyexe-mozilla-urges-public-to-swipe-left-on-romantic-ai-chatbots-due-to-major-privacy-red-flags/.

[47] Yeh, Chih-Liang. "Pursuing Consumer Empowerment in the Age of Big Data: A Comprehensive Regulatory Framework for Data Brokers." *Telecommunications Policy*, SI: Interconnecting, 42, no. 4 (May 1, 2018): 282–92. https://doi.org/10.1016/j.telpol.2017.12.001.

[48] Bowen III, Donald E., S. McKay Price, Luke C. D. Stein, and Ke Yang. "Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, April 30, 2024. https://doi.org/10.2139/ssrn.4812158; Gressel, Gilad, Rahul Pankajakshan, and Yisroel Mirsky. "Discussion Paper: Exploiting LLMs for Scam Automation: A Looming Threat." In *Proceedings of the 3rd ACM*

Workshop on the Security Implications of Deepfakes and Cheapfakes, 20–24. Singapore Singapore: ACM, 2024. https://doi.org/10.1145/3660354.3660356.

[49] Gressel, Gilad, Rahul Pankajakshan, and Yisroel Mirsky. "Discussion Paper: Exploiting LLMs for Scam Automation: A Looming Threat." In Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes, 20–24. Singapore Singapore: ACM, 2024. https://doi.org/10.1145/3660354.3660356; Murphy, Hannah. "Hackers 'Jailbreak' Powerful AI Models in Global Effort to Highlight Flaws." Financial Times, June 21, 2024, sec. Artificial intelligence. https://www.ft.com/content/14a2c98b-c8d5-4e5b-a7b0-30f0a05ec432.

[50] Li, Haoran, Yulin Chen, Jinglong Luo, Jiecong Wang, Hao Peng, Yan Kang, Xiaojin Zhang, et al. "Privacy in Large Language Models: Attacks, Defenses and Future Directions." arXiv, September 30, 2024. https://doi.org/10.48550/arXiv.2310.10383.

[51] Falade, Polra Victor. "Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, October 9, 2023, 185–98. https://doi.org/10.32628/CSEIT2390533.

[52] Weidinger et al., "Ethical and Social Risks of Harm from Language Models."

[53] Zhang, Zhiping, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "'It's a Fair Game', or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–26. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. https://doi.org/10.1145/3613904.3642385.

[54] Palmer, Greg, and Peter Church. "The Muah.Ai Data Breach – Extortion Threats and Cyber Vulnerabilities." Linklaters, October 14, 2024. https://www.linklaters.com/en/insights/blogs/digilinks/2024/october/the-muah-ai-data-breach---extortion-threats-and-cyber-vulnerabilities.

[55] Li et al., "Privacy in Large Language Models."

[56] Jaff, Evin, Yuhao Wu, Ning Zhang, and Umar Iqbal. "Data Exposure from LLM Apps: An In-Depth Investigation of OpenAI's GPTs." arXiv, August 23, 2024. https://doi.org/10.48550/arXiv.2408.13247.

[57] Gurman, Mark. "Samsung Bans Generative AI Use by Staff After ChatGPT Data Leak." *Bloomberg.Com*, May 2, 2023. https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak.

[58] Stefan Baack, "Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI" (Mozilla Foundation, February 2024), https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/.

[59] Shayne Longpre et al., "A Large-Scale Audit of Dataset Licensing and Attribution in AI," *Nature Machine Intelligence* 6, no. 8 (August 2024): 975–87, https://doi.org/10.1038/s42256-024-00878-8.

[60] Simon Chesterman, “Good Models Borrow, Great Models Steal: Intellectual Property Rights and Generative AI,” *Policy and Society*, February 12, 2024, https://doi.org/10.1093/polsoc/puae006.

[61] Bobby Allyn, “‘The New York Times’ Takes OpenAI to Court. ChatGPT’s Future Could Be on the Line,” *NPR*, January 14, 2025, sec. Business, https://www.npr.org/2025/01/14/nx-s1-5258952/new-york-times-openai-microsoft.

[62] Harry Farmer and Julia Smakman, “Delegation Nation” (London: Ada Lovelace Institute, February 4, 2025), https://www.adalovelaceinstitute.org/policy-briefing/ai-assistants/.

[63] Wang, Zichong, Zhibo Chu, Thang Viet Doan, Shiwen Ni, Min Yang, and Wenbin Zhang. “History, Development, and Principles of Large Language Models-An Introductory Survey.” arXiv, September 23, 2024. https://doi.org/10.48550/arXiv.2402.06853.

[64] Emily M. Bender et al., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922; Shayne Longpre et al., “Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?” (Cambridge, MA: MIT, March 27, 2024), https://mit-genai.pubpub.org/pub/uk7op8zs/release/2.

[65] Joy Buolamwini and Timnit Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Conference on Fairness, Accountability and Transparency, PMLR, 2018), 77–91, https://proceedings.mlr.press/v81/buolamwini18a.html; Maddalena Favaretto, Eva De Clercq, and Bernice Simone Elger, “Big Data and Discrimination: Perils, Promises and Solutions. A Systematic Review,” *Journal of Big Data* 6, no. 1 (February 5, 2019): 12, https://doi.org/10.1186/s40537-019-0177-4.

[66] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018).

[67] Eszter Hargittai, “Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites,” *The ANNALS of the American Academy of Political and Social Science* 659, no. 1 (May 1, 2015): 63–76, https://doi.org/10.1177/0002716215570866; Gaël Varoquaux, Alexandra Sasha Luccioni, and Meredith Whittaker, “Hype, Sustainability, and the Price of the Bigger-Is-Better Paradigm in AI” (arXiv, September 21, 2024), https://doi.org/10.48550/arXiv.2409.14160.

[68] Heikkilä, Melissa. “AI Language Models Are Rife with Different Political Biases.” *MIT Technology Review*, August 7, 2023. https://www.technologyreview.com/2023/08/07/1077324/ai-language-models-are-rife-with-political-biases/.

[69] Manzini et al., “Should Users Trust Advanced AI Assistants?”

[70] Farmer and Smakman, “Delegation Nation”; Wachter, Sandra, Brent Mittelstadt, and Chris Russell. “Do Large Language Models Have a Legal Duty to Tell the Truth?” *Royal Society Open Science* 11, no. 8 (August 7, 2024): 240197. https://doi.org/10.1098/rsos.240197.

[71] Metz, Cade and Karen Weise. "A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse." *The New York Times*, May 5, 2025, sec. Technology. https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html.

[72] Burkhardt, Sarah, and Bernhard Rieder. "Foundation Models Are Platform Models: Prompting and the Political Economy of AI." *Big Data & Society* 11, no. 2 (June 1, 2024). https://doi.org/10.1177/20539517241247839.

[73] Alan Turing Institute. "AI Explainability in Practice." Accessed May 19, 2025. https://aiethics.turing.ac.uk/modules/explainability/, https://aiethics.turing.ac.uk/modules/explainability/.

[74] Gabriel, Iason, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, et al. "The Ethics of Advanced AI Assistants." arXiv, April 28, 2024. https://doi.org/10.48550/arXiv.2404.16244.

[75] Pinski, Marc, and Alexander Benlian. "AI Literacy for Users – A Comprehensive Review and Future Research Directions of Learning Methods, Components, and Effects." *Computers in Human Behavior: Artificial Humans* 2, no. 1 (January 1, 2024). https://doi.org/10.1016/j.chbah.2024.100062.

[76] Bughin, Jacques, Jeongmin Seong, James Manyika, Michael Chui, and Raoul Joshi. "Modeling the Global Economic Impact of AI." Discussion Paper. McKinsey, 2018. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy; Farmer and Smakman, "Delegation Nation."

[77] Akbulut et al., "All Too Human?"; Firth-Butterfield, Kay, Aubra Anthony, and Emily Reid. "Without Universal AI Literacy, AI Will Fail Us." *World Economic Forum* (blog), March 17, 2022. https://www.weforum.org/stories/2022/03/without-universal-ai-literacy-ai-will-fail-us/.

[78] Burkhardt and Rieder, "Foundation Models Are Platform Models."

[79] Barman, Kristian González, Nathan Wood, and Pawel Pawlowski. "Beyond Transparency and Explainability: On the Need for Adequate and Contextualized User Guidelines for LLM Use." *Ethics and Information Technology* 26, no. 3 (July 17, 2024): 47. https://doi.org/10.1007/s10676-024-09778-2; Pinski and Benlian, "AI Literacy for Users – A Comprehensive Review and Future Research Directions of Learning Methods, Components, and Effects."

[80] Wachter, Mittelstadt, and Russell, "Do Large Language Models Have a Legal Duty to Tell the Truth?"

[81] Adam, Hammaad, Aparna Balagopalan, Emily Alsentzer, Fotini Christia, and Marzyeh Ghassemi. "Just Following AI Orders: When Unbiased People Are Influenced By Biased AI," 2022. https://openreview.net/forum?id=ISzWXSWiL8.

[82] Wihbey, John. "AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?" SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, April 20, 2024. https://doi.org/10.2139/ssrn.4805026.

[83] Farmer and Smakman, "Delegation Nation."

[84] Brandtzaeg, Petter Bae, Marita Skjuve, and Asbjørn Følstad. "AI Individualism: Transforming Social Structures in the Age of Social Artificial Intelligence." In *Oxford Intersections: AI in Society*, edited by Philipp Hacker. Oxford University Press. https://doi.org/10.1093/9780198945215.003.0099.

[85] Glickman, Moshe, and Tali Sharot. "How Human–AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements." Nature Human Behaviour, December 18, 2024, 1–15. https://doi.org/10.1038/s41562-024-02077-2.

[86] Wachter, Mittelstadt, and Russell, "Do Large Language Models Have a Legal Duty to Tell the Truth?"

[87] Bender et al., "On the Dangers of Stochastic Parrots."

[88] Luitse, Dieuwertje, and Wiebke Denkena. "The Great Transformer: Examining the Role of Large Language Models in the Political Economy of AI." Big Data & Society 8, no. 2 (July 1, 2021). https://doi.org/10.1177/20539517211047734.