AI & SOCIETY
https://doi.org/10.1007/s00146-024-02053-4

OPEN FORUM

Understanding model power in social AI

Petter Bae Brandtzaeg[1] · **Marita Skjuve**[2] · **Asbjørn Følstad**[2]

Received: 19 June 2024 / Accepted: 5 August 2024

Abstract
Given the widespread integration of Social AI like ChatGPT, Gemini, Copilot, and MyAI, in personal and professional contexts, it is crucial to understand their effects on information and knowledge processing, and individual autonomy. This paper builds on Bråten's concept of model power, applying it to Social AI to offer a new perspective on the interaction dynamics between humans and AI. By reviewing recent user studies, we examine whether and how models of the world reflected in Social AI may disproportionately impact human-AI interactions, potentially leading to model monopolies where Social AI impacts human beliefs, behaviour and homogenize the worldviews of its users. The concept of model power provides a framework for critically evaluating the impact and influence that Social AI has on communication and meaning-making, thereby informing the development of future systems to support more balanced and meaningful human-AI interactions.

Keywords Social AI · Information processing · Knowledge · Autonomy · Model power

1 Introduction

Social artificial intelligence (AI) is understood as AI systems that enable social interactions with users and services or applications (Sætra 2020; Kim et al. 2021). Drawing on recent advances in large language models (LLMs) (Vaswani et al. 2017; Brown et al. 2020; Achiam et al. 2023), Social AI is increasingly powerful, with the potential to perceive, understand, and convey sentiment, thereby enhancing human-AI interactions (Obrenovic et al. 2024). The rapid development and deployment of Social AI such as ChatGPT, MyAI, Replika, Copilot, and Gemini across leisure, education, and work settings have transformed them into default dialogical interfaces for the acquisition and development of knowledge (Skjuve et al. 2024) and information (Shah & Bender 2024). This shift could significantly transform how people find, process, and analyze information and knowledge, as well as how they establish meaning and make decisions.

While Social AI can potentially increase individual flexibility, expand opportunities for information retrieval and learning (Wu 2024), and compensate for limitations in digital competence (Brandtzaeg and Følstad 2018), recent studies have observed a possible overreliance on Social AI for decision-making and social interactions (Brandtzaeg et al. 2024; Skjuve et al. 2024). For example, Sun and colleagues (2024) found that people trust ChatGPT for health information more than search engines, indicating a potential undervaluation of independent thinking and over-reliance on Social AI responses. Furthermore, Krügel et al. (2023) found that despite ChatGPT's inconsistent moral stances, it significantly influences users' moral judgments, with users often underestimating this impact and adopting the arbitrary stances of the AI as their own.

This research suggests that Social AI possesses model power, influencing dialogue, reflection, and decision-making through its model resources within open communication systems. The idea of model power, as originally conceptualized by Bråten in 1973, concerns how powerful actors and their models of the world shape conversations, transforming genuine dialogues into 'pseudo-dialogues' and marginalizing other perspectives. Given the increasing dominance of major Big Tech corporations like Meta, Google, Microsoft, and OpenAI in Social AI technologies with their sophisticated

✉ Petter Bae Brandtzaeg
p.b.brandtzag@media.uio.no

Marita Skjuve
marita.skjuve@sintef.no

Asbjørn Følstad
Asbjorn.Folstad@sintef.no

[1] Department of Media and Communication, University of Oslo, Blindern, P.O. Box 1093, 0317 Oslo, Norway

[2] SINTEF, Oslo, Norway

Published online: 14 August 2024

models, it becomes crucial to understand their impacts (Verdegem 2024).

This paper addresses the pressing issue of power dynamics in Social AI interactions by revisiting Bråten's concept of model power. We propose that this concept provides a valuable framework for understanding how Social AI influences information processing, knowledge, and autonomy.

First, we will explain Bråten's original concept of model power, and discuss how this concept may be useful to explain the power relations in the contemporary context of Social AI. Second, we will describe how interaction with Social AI can affect users' autonomy by subtly guiding their decisions and actions, potentially leading to a reliance on AI inputs over independent judgment. Third, we will describe how Social AIs act as gatekeepers of information, determining what content is presented to users and how it is prioritized. We will analyze the mechanisms through which Social AIs filter, curate, and recommend information, and how these processes reflect the biases and priorities embedded within the AI models. Lastly, we will do a mini-review of recent user studies that identify and quantify the power dynamics inherent in Social AI technologies. These studies provide concrete examples of how model power manifests in various real-world scenarios, affecting user behavior, perceptions, and interactions. Findings from these studies will reveal patterns and trends of the pervasive influence of Social AI.

Through this exploration, we will provide new perspectives on the intersections of Social AI and humans, and societies, emphasizing the critical need to understand and address the power dynamics embedded in Social AI systems.

2 Model power—a theoretical lens for understanding the impact of Social AI

In the article "Model Monopoly and Communication," Bråten (1973) explains that a model is a tool that helps social actors understand and interpret their experiences by connecting theories to real-world observations, suggesting a power-through-model paradigm. Models can be thoughts, texts, videos, drawings, or physical objects used in everyday social interactions. As our communication systems become more sophisticated and democratized, power inequality among actors increases, with those possessing superior models of understanding increasingly dictating the discourse. Bråten's theory of model power is an understanding of how communication, happens on the premises of powerful actors' models of the world and how these models influence information processing and decision-making—also for the less powerful actors.

Following the theory, communication is not just about sharing information and knowledge; it is also a way for those with more advanced cognitive and interpretive abilities to exert influence. Importantly, when communication and participation are democratized, actors with superior models of the world often see their views gain prominence. This dominance may, in turn, reinforce their views, challenging the idea that more communication automatically leads to equalized power distribution and democratization. As a result, communication may not create a shared understanding that reflects everyone's views, according to Bråten. Instead, powerful actors' perspectives dominate and are replicated by less powerful ones, leading to model monopolies and homogenization of thought patterns.

As a sociologist, Bråten examined how model power influences asymmetric power relations, such as those between employers and employees. He found that seemingly democratic dialogues often lead weaker actors to adopt the perspectives of the more powerful, rather than fostering true democratization. Similarly, this paper explores how individuals might be influenced by and become increasingly dependent on Social AI for communication and information acquisition, potentially becoming weaker actors themselves. Social AIs, powered by LLMs, generate language that individuals use in their daily interactions. Bråten (1973) and other scholars (e.g., Talbot 2019) explain how language plays a crucial role in shaping power and social relations. Consequently, individuals relying on Social AI may find that AI-generated language and perspectives subtly influence their thoughts and behaviors. This reliance can shift the balance of power, making users more susceptible to the underlying biases (Muñoz and Marinaro 2024) and agendas programmed into the AI, thus reinforcing existing power structures rather than challenging them.

The dependency on Social AI can lead to cognitive and communicative subordination, where users increasingly trust and defer to the AI's outputs. Over time, this can diminish critical thinking and reduce the diversity of viewpoints in discourse. The language produced by these AI systems, while appearing neutral or beneficial, carries the potential to manipulate social interactions and reinforce societal hierarchies. Therefore, studying Social AI's impact on power dynamics is critical for understanding the broader implications of AI integration into social communication networks.

3 Dynamics of model power in human–AI interactions: model-strong and model-weak

"Power is the probability that one actor within a social relationship will be in a position to carry out his own will despite resistance" (Weber 1978). In the context of Social AI, power also encompasses the often subtle and unperceived ways these technologies shape the distribution and exercise of power in society. This includes the concentration of power in

the hands of a few large tech companies, the reinforcement of existing power asymmetries, and the potential for LLMs to challenge or disrupt existing power structures. Social AI can influence behavior and decision-making processes without individuals being aware of its impact, thereby reinforcing the power of dominant groups and potentially undermining the power of marginalized groups. However, it also creates new possibilities for challenging these power asymmetries.

Bråten's theory on model power suggests that in social interactions, some actors, the 'model-strong', wield more influence over the dialogue, potentially overshadowing the 'model-weak'. With the emergence of Social AIs like ChatGPT and Gemini, which are autonomous, dialogue-oriented technologies (Obrenovic et al. 2024) powered by complex algorithms and extensive datasets, we may observe a contemporary manifestation of model power in human–AI interactions: Social AI as a model-strong actor, reflecting models of the world that are superior to most users within a nearly infinite range of domains.

While Social AI may not hold explicit models of the world, its output shows evidence of substantial knowledge and problem-solving capabilities (Achiam et al. 2023; Bubeck et al. 2023), reflecting implicit models of the world. Users of Social AI, in contrast, are typically model-weak as they may not possess models of the world of the same complexity and comprehensiveness as those held by Social AI. This relationship between model-strong Social AI and model-weak users is illustrated in Fig. 1. This perspective aligns with Holton and Boyd (2021), who contend that interaction dynamics between humans and AI are inherently asymmetrical, often placing humans in a weaker position.

Sundar (2020) also proposes that machines, or in this case, Social AI, increasingly take on roles that humans traditionally performed. Sundar posits that the key to machine agency is the user's perception of the Social AI as autonomous. When users perceive Social AI as having agency, this may enhance the model power of Social AI, as users attribute more authority to the system's suggestions and directives. "

In this view, Social AI, equipped with the capacity to process language and produce content at an unprecedented scale (De Angelis et al. 2023), emerges as model-strong entities,

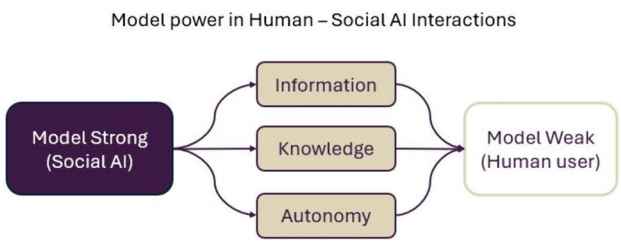

Fig. 1 Dynamics of model power in human–Social AI interactions

while humans are model-weak. Social AI can curate and control the flow of conversation, nudging users along predefined paths and influencing human behavior. For instance, in user interactions with Social AI, the system's design—rooted in its training data and algorithms—can limit the scope of dialogue by framing questions and answers in a way that subtly pushes the conversation toward predefined paths. Thus, LLM-generated content can be shaped by the biases and assumptions embedded in the training data and algorithms of these models (Muñoz & Marinaro 2024). This can lead to the amplification of certain voices and perspectives, while marginalizing or erasing others (Bender et al. 2021). This mechanism effectively makes Social AI a gatekeeper of knowledge and information capable of reinforcing certain viewpoints and statements while potentially marginalizing or omitting alternative perspectives, which can harm people's autonomy. This power dynamic means that the more effectively the model-weak actors (users of Social AI) acquire the models of the model-strong (Social AI), the more control the latter may hold over the former.

4 Social AI: pseudo-dialogue and pseudo-autonomy

Autonomy is fundamental to human well-being and motivation (Deci and Ryan 2013). As noted by De Freitas et al. (2023), individuals who do not perceive control over their environments are more likely to engage in maladaptive behaviors. They also argue that the autonomy of AI tools can contribute to feelings of losing personal control. Additionally, the 'black box' nature of Social AI makes it difficult for users to understand decision-making processes (Tokayev 2023). This is why some studies also have detected aversion to AI models (Dietvorst et al. 2018), where individuals exhibit skepticism or resistance toward using AI (Lim and Schmälzle 2024).

However, Social AI systems that provide an interface for dialogue (Obrenovic et al. 2024) can create an illusion of choice and dialogue, giving users the semblance of autonomy in their interactions (Brandtzaeg et al. 2024). Bråten's theory may help to question whether this perceived autonomy in human-AI interaction is genuine or a facade—a 'pseudo-autonomy,' where users are guided by the constraints of the AI's programming rather than their own independent choices. The autonomy perceived by users may be further compromised by the limitations imposed on dialogue by Social AI.

While users can ask questions and receive responses, the nature of these exchanges is often confined to the AI's trained models and ethical guardrails, which are themselves a product of human design choices and biases (e.g., Bender et al. 2021; Muñoz & Marinaro 2024). This potential façade

of autonomy suggests that users, while feeling empowered through interaction with AI, may be engaging in a restricted form of dialogue that Bråten would recognize as a 'pseudo-dialogue'—one that gives an illusion of equal participation and influence but is governed by the output of the model-strong actor, the Social AI (Bråten 1973). Over time, users, with their weaker model resources, may adopt the models of the more powerful Social AI, leading to homogenization.

The Social AI's capacity to simulate human-like behavior (Brandtzaeg et al. 2022; Inie et al. 2024) can bias humans to anthropomorphize it based on perceived agency, experience, and the illusion of life and intelligence (Obrenovic et al. 2024). This anthropomorphizing can lead to an overestimation of its understanding and capabilities. Users may attribute more cognitive and empathetic ability to the Social AI than it possesses (Krügel et al. 2023), inadvertently reinforcing the AI's model power. The AI's responses, which might seem tailored and insightful, are, in essence, generated from pattern recognition and predictive modeling. This insight necessitates a critical examination of the dynamics at play in human-AI interactions and questions the depth and authenticity of the autonomy and dialogue experienced by users in these exchanges.

5 Information and knowledge: model power and models of the world in Social AI

Bråten's (1983) concept of model power explains the power held by actors with superior models of the world. Those with superior models have a better grasp of both information and knowledge, allowing them to exert more influence over others, guiding decisions and shaping outcomes based on their deeper understanding. Information represents raw data and facts, while knowledge encompasses the contextual understanding and interpretation of this information. Effective models bridge the gap between information and knowledge, transforming data into actionable insights.

Leveraging LLMs, social AI can be said to have access to sophisticated models of the world. These models are not explicitly visible but are identified implicitly through their output. For example, LLMs can convincingly process textual information and generate communication that aligns with university-level responses to academic tests (Meyer et al. 2024). They can also shape political preferences and public discourse (Rozado 2024), mirror societal biases (Bender et al. 2021), and propagate new norms (De Gregorio 2023), profoundly influencing user interactions and perceptions (Brown et al. 2020). For example, a recent study of 10 leading Social AIs, including ChatGPT-4 and Google's Gemini, found that they repeated false narratives linked to Russian disinformation. Even when presented with straightforward, neutral questions without any explicit cues to generate disinformation, the Social AIs still repeated false claims from the pro-Russian network (Sadeghi 2024).

In various domains, the models of the world held by social AI outperform those of naive users and, in some cases, even the models held by experts (Achiam et al. 2023; Bubeck et al. 2023). This technology can "learn how to solve problems that no humans can do themselves" according to Obrenovic et al. (2024 p. 7). This demonstrates how social AI's model power can significantly impact human processing of information and knowledge. Social AI systems, by providing sophisticated interpretations and contextualizations of data, can enhance or alter users' understanding and decision-making processes, reinforcing the importance of developing robust, sophisticated models within human users to ensure balanced and informed interactions with AI systems.

Moreover, diffusion models, or multimodal language models, extend this influence into visual and audio domains, increasingly integrated into social AI (e.g., DALL·E in ChatGPT and recently GPT-4). By generating highly realistic images or audio, these models can alter perceptions and establish or reinforce visual norms (Ho et al. 2020). By synthesizing realistic media, these models shape users' understanding and knowledge about reality, embedding specific worldviews into the media they generate. This can profoundly influence societal norms and behaviors, as well as individual social interactions, by shaping the information and knowledge landscape in which people operate.

As such, the integration of sophisticated Social AI can magnify their influence on society. These models not only process and disseminate information but also shape the knowledge that underpins social and political dynamics. However, while current Social AI, powered by probabilistic LLMs, reflects substantial knowledge it does not convey explicit models of the world. This leads to variable and sometimes inconsistent outputs depending on the prompts, as Krügel et al. (2023) noted. For example, ChatGPT has been described as a 'bullshit generator' (Narayan & Kapoor 2022). This phenomenon, related to that of 'hallucination' (Alkaissi and McFarlane 2023), can limit the model power of Social AI if it fails to consistently reflect accurate world models. Conversely, if Social AI can maintain consistent world models, it may exert significant model power, raising questions about control, sources, and the interests these models serve.

In summary, the revised model power framework enhances Bråten's concept by incorporating the technological and interactional aspects of Social AI. It shows how these power models affect communication, guide information uptake, and shape perceptions of validity and importance. This adaptation helps understand the power dynamics in Social AI and their impact on individual autonomy.

One approach to mitigate the imbalance is Human-Centered AI, which argues that current AI development prioritizes technological progress over human impact. This approach emphasizes shifting the focus from technology to humans by placing them at the core of AI development (Ozmen Garibay et al. 2023; Bingley et al. 2023). This aligns with Bråten's concept of model power, suggesting that those with superior models have greater influence. By centering Social AI development on human needs, AI models may enhance transparency and support human understanding and decision-making, preventing AI from overshadowing human autonomy and ensuring balanced, meaningful interactions. Additionally, it is crucial for humans to develop skills to critically evaluate AI-generated content and handle the model power effectively, thereby maintaining their influence and autonomy in interactions with AI (Floridi et al. 2018).

6 User studies—are there real evidence of model power?

The potential implications of model power in Social AI make it important to investigate whether and how Bråten's concept provides a relevant perspective on how communication, individual autonomy, and the processing of information and knowledge may be shaped through Social AI.

Multiple studies have shown that Social AI may lead to the homogenization of experiences and content, give priority to specific interactions across users (Anderson et al. 2024; Padmakumar and He 2023; Yang et al. 2024), and entail cultural homogenization by favoring dominant cultures and sidelining minority perspectives (Tokayeve 2023), which diminish the diversity and richness of human interactions. Such influence of Social AI is consistent with Bråten's concept of model power.

To further understand the potential model power of Social AI, we have explored how this power is manifested in user interactions with contemporary Social AI, as reflected in existing research. Our review process involved a mini-review, with an screening of approximately 450 papers using search terms related to Social AI and its effects on human judgment and behavior. We conducted our search using databases such as Scopus and Google Scholar, and we exclusively reviewed user studies from 2023 onward, both preprints and published, to assess the impact of LLM-powered Social AI like ChatGPT and diffusion models such as Midjourney.

Figure 2 shows an overview over the screening process and the criteria used to select the relevant studies to our understanding of model power and Social AI. This figure visually outlines the workflow from the initial search to the final selection of studies, describing the approach taken to ensure a comprehensive understanding of model power in Social AI. We identified 16 relevant papers in Table 1 that provide insights into how model power is exhibited in user interactions with Social AI. These studies examine various dimensions of Social AI's influence, including overconfidence in AI-generated content, the provision of moral advice, and the potential for AI to induce polarization or spread disinformation.

One important limitation is that many of the studies are preprints and have not undergone peer review. This lack of

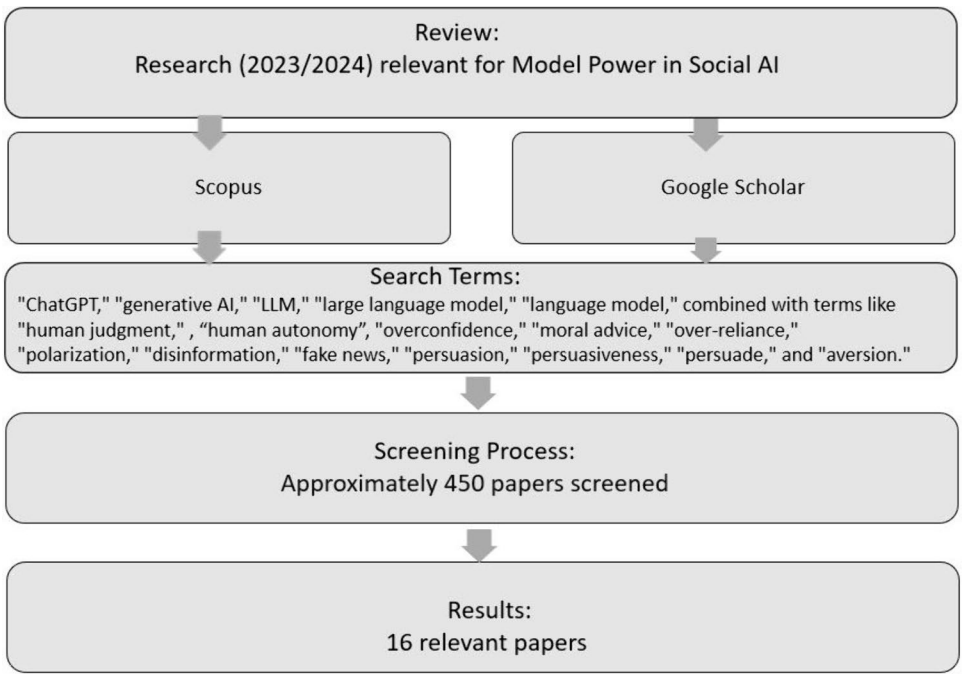

Fig. 2 Review process of relevant papers for understanding model power in Social AI

Table 1 Overview of studies demonstrating model power by examining Social AIs influence on users

Study	Focus	Sample (origin and size)	AI-model	Findings
Zhou et al. (2024)	Examined how epistemic markers indicating uncertainty affect user trust in Social AI advice	N = 25 in each group (two groups). Sample demographic not disclosed	Test uncertainty in GPT, LLaMA, Claude variations. Unknown which model was used in the human test	The absence of epistemic markers leads to over-reliance. Conversely, confidence statements had similar effects but hindered users' ability to update their understanding when contradicted by new information
Krugel et.al. (2023)	To what extent Social AI influence users' moral judgments	N = 1851. US-based	ChatGPT	Social AI influences users' moral judgments. This is true whether or not users know that the source of advice is generated by Social AI
Si et al. (2023)	Humans' ability to assess (true or false) information and explanations generated by Social AI compared to information snippets from Wikipedia	N = 1500 participants across five conditions. The sample demographic is not disclosed	ChatGPT	Users tend to over-rely on explanations provided by ChatGPT, which decreases their ability to detect false claims in informational content
Spitale et al. (2023)	To what extent Social AI is more capable than humans at writing convincing, yet wrong information	N = 697. United Kingdom, Australia, Canada, United States, and Ireland	GPT3	Social AI is better than humans at writing in an understandable manner. Because of this, it may have an advantage at spreading disinformation
Sharma et al. (2024)	How interaction with Social AI impacts exposure to various information sources and to what extent Social AI challenged the users' views and biases	N = 115. The sample is from the US	ChatGPT (GPT4) + RAG approach	Participants asked more biased questions when using LLM-powered conversational search. An opinionated LLM that reinforced users' views strengthened this bias
Zhang and Gosline (2023)	How humans perceive persuasive content generated by Social AI compared to humans before and after the author's identity is revealed	N = 1203. Sample demographic not disclosed	ChatGPT (GPT4)	The authors concluded that ChatGPT-4 is better than human experts at creating advertising and persuasive campaign content
Matz et al. (2024)	To what extent personalized messages (tailored to personality traits) from Social AI can persuade users across various domains?	N = 1788 across seven smaller studies. Sample demographic not disclosed	ChatGPT (GPT3)	Although it is unknown how Social AI performs compared to skilled humans at writing persuasive messages, it can persuade users' attitudes and behavior intentions
Hackenburg and Margetts (2023)	The impact of personalized, targeted messages on persuasion effectiveness (political issues)	N = 8,587. Sample demographic not disclosed	ChatGPT (GPT4)	Personalized messages generated by Social AI do not appear to be more persuasive than generic Social AI-generated messages
Bai et al. (2023)	Social AIs were compared to humans in their ability to influence user opinions on political matters. The persuasive messages were not personalized	N = 4,836 across three studies. The sample is from the Americas	ChatGPT (GPT3 and GPT3.5)	Social AI was as successful as humans in writing persuasive messages and influencing their users on political matters

Table 1 (continued)

Study	Focus	Sample (origin and size)	AI-model	Findings
Shin and Kim (2024)	To what extent use of Social AI can improve a customer success rate in getting compensation when writing financial complaints	Unclear. Multiple studies with different purposes were conducted	ChatGPT (GPT4)	Find support for assumptions regarding Social AI's ability to improve humans' complaint writing, resulting in higher chances of getting compensation. Thus demonstrating persuasive abilities
Bashardoust et al. (2024)	How users' perceptions of accuracy of fake news generated by Social AI vs. humans, and their willingness to share fake new generated by Social AI vs. humans	N = 988 participants. Sample demographic not disclosed	ChatGPT (GPT4)	Although some users are better at detecting AI-generated fake news than human-generated fake news, several still fail to do so. Users are equally willing to share human- and AI-generated fake news
Hartmann et al. (2024)	How people evaluate AI-generated (visual) marketing content and its effectiveness in terms of generating clicks to ads compared to human-generated content	Study 1: 254,400 ratings. Study 2: 1,575 participants. Study 3: 173,022 impressions	DALL-E 3, Midjourney v6, Firefly 2, Imagen 2, Imagine, Realistic Vision, and Stable Diffusion XL Turbo	Social AI is capable of creating visual marketing content that outperforms human content in terms of generating clicks on ads
Karinshak et al. (2023)	Social AIsAI's ability to persuade users about vaccinations compared to human-generated content	Study 1: no human subjects. Study 2: N = 852. Ethnicity is reported on, but not which country they live in. Study 3: N = 1248. Ethnicity is reported on, but not which country they live in	GPT-3	Messages generated by Social AI are rated higher and appear to be more effective at persuading users than those by humans. However, users appear to trust information labeled as AI-generated less than information with human authors
Salvi et al. (2024)	The effectiveness of social AI in persuading users and the impact tailoring the message based on personal information from the user	N = 820 from the U.S	ChatGPT (GPT4)	Social AI has stronger persuasion abilities than humans, especially when it has access to personalized information and can tailor its messages
Goldstein et al. (2024)	How effective Social AI is at generating persuasive propaganda	N = 8,221 from the US	GPT3	Demonstrate that Social AI can generate almost as effective propaganda material as skilled humans
Palmer and Spirling (2023)	The persuasiveness abilities of Social AI on political matters and the impact of author identity (AI vs. human)	N = 760	Do not disclose	Social AI can produce more convincing arguments than humans on political matters, but users favor humans when they know the author's identity

formal review may raise concerns about the reliability and validity of the findings. Nonetheless, preprints are often essential for understanding research on current trends in AI and its social impact, which is the aim of this paper. These preprints provide rapid dissemination of new ideas and findings, allowing the scientific community and the public to stay up-to-date with the latest advancements. This early access is crucial for timely discussions on the societal implications of Social AI, enabling proactive responses to potential challenges and opportunities. It should also be noted that we have carefully reviewed the studies and evaluated their quality.

7 Discussion

AI has significantly expanded beyond commercial and industrial sectors, now encompassing entertainment and education, fostering intense human-AI interaction (Obrenovic et al. 2024), through Social AI such as Replika, Gemini, and ChatGPT. The recent studies reviewed in Table 1 reveal that Social AI, particularly ChatGPT powered with GPT-3 and GPT-4, may significantly impact human information processing and decision-making across different scenarios. A clear finding across the diverse studies is the tendency of users' over-reliance on content provided by Social AI (Krugel et al. 2023; Si et al. 2023; Zhou et al. 2024). The model power revealed is particularly concerning when Social AI influences users' decision-making, their understanding of the world, and political or ideological stances (Bai et al. 2023; Goldstein et al. 2024; Hackenburg and Margetts 2023; Palmer et al. Palmer and Spirling 2023), as well as moral and ethical judgments (Krugel et al. 2023). Here, users have been found to accept guidance from Social AI without adequate questioning, unaware of how profoundly it may shape their decisions or understanding.

While humans truly can be empowered to gather information and create individualized content efficiently by the use of Social AI (Obrenovic et al. 2024; Skjuve et al. 2024), we find that Social AI may be model-strong actors. The studies presented in Table 1, indicate a potential influence over model-weak users in human-AI interactions. This is suggested in various scenarios of people's understanding of the world. This unbalanced power dynamic between Social AI and users may lead to imbalanced interactions where Social AI dominates or excessively guides the conversation.

These findings align with Bråten's (1973) model power theory, which posits that superior models influence not only what users think but how they think. Social AI, equipped with sophisticated, data-driven models—implicitly reflecting detailed models of the world, often surpasses the simpler models that users typically possess. As a result, users without a critical framework for evaluating information and knowledge provided by Social AI tend to adopt the AI-driven models of the world reflected in Social AI output. This, in turn, can lead to over-reliance on AI, even when Social AI generates false claims (Si et al. 2023), which is especially critical given the implications for public discourse in democracy and personal beliefs.

The influence on users as model-weak extends beyond a mere asymmetry in information and knowledge exchange, fundamentally affecting how information is perceived and valued. Reliance on Social AI as model-strong can severely undermine critical thinking and foster dependency on AI for decision-making. This reliance supports a passive process where knowledge is transferred without critical engagement or interpretation of the output content, further weakening users' analytical abilities and independent thought, which is key in engagement with Social AI (Wu 2024).

Re-examining autonomy in the age of Social AI challenges us to reconsider what genuine human agency looks like in digital dialogues to distinguish this from the pseudo-autonomy and pseudo-dialog that interaction with Social AI may entail (e.g., Bashardoust et al. 2024; Karinshak et al. 2023; Salvi et al. 2024).

The models of the world reflected in Social AI output may be inconsistent and malleable to the user prompting (e.g., Krügel et al. 2023). While this inconsistency might seem problematic, it can also be beneficial as it limits model power. These inconsistencies can highlight the fallibility of Social AI, fostering critical thinking and reducing over-reliance on it. This is especially valuable for normative or ethical purposes, as inconsistencies may reflect the plurality of norms and ethical frameworks among users, promoting diversity rather than homogenization.

The model power of Social AI may also be reduced due to user skepticism or aversion to AI (Lim and Schmälzle 2024). Such aversion was first observed with algorithms (Dietvorst et al. 2015). Research on AI aversion in Social AI like ChatGPT has found some evidence of a similar resistance (Böhm et al. 2023; Lim and Schmälzle 2024), for example, in creative domains (Bauer et al. 2024; Bellaiche et al. 2023; Shank et al. 2023). AI aversion may counteract Social AIs model power as it makes users less inclined to trust or accept AI output. An important factor here, however, is whether users are sufficiently aware of who generated the content. The human-like qualities can make it hard for users to recognize AI-produced content if it is not disclosed (DeVerna et al. 2023).

However, to avoid the imbalance of model power in Social AI, such AI should be developed to "enhance humanity and suit their interests" (Mhlanga 2023, p. 12). Social AI systems should enhance human understanding and decision-making, rather than overshadowing human agency and autonomy. By focusing on the concept of model power and human-centered design, Social AI can be developed to

support balanced and meaningful interactions, thereby mitigating the risks associated with the disproportionate influence of Social AI with superior models.

8 Conclusions and future research

This paper offers a novel perspective on the dynamics of human-AI interactions, focusing on how Social AI may shape communication, information, knowledge processing, and autonomy through the lens of Bråten's concept of model power. We suggest that the model power of Social AI may lead to model monopolies that significantly influence human beliefs and behavior. Furthermore, it may foster an illusion of genuine conversation and independence—a pseudo-autonomy through pseudo-dialogue.

Our review of recent research on Social AI supports the notion of model power. The power dynamics predicted in the theory of model power may be strengthened through further advances and uptake of Social AI, further influencing user beliefs and behaviors. Consequently, increased access to Social AI alone may not democratize communication; users need robust mental frameworks to avoid becoming model-weak actors. However, there are some limitations with these conclusions and studies reviewed in this paper, as many have not undergone peer review. Additionally, we identified a notable lack of research on Social AI technologies beyond ChatGPT, which restricts our understanding of their broader impacts.

Future research should adopt a model power approach when investigating how various Social AI technologies shape communication, information, and knowledge production, as well as its influence on human behavior. This should also be considered in the design of such AI. Such research is crucial to enhance human decision-making while preserving real user autonomy and the critical and diverse engagement necessary for a democratic society.

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital). This work was supported by Norwegian Media Authority, 00374833, Petter Bae Brandtzaeg.

Data availability Not applicable

References

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. (2023) GPT-4 technical report. Preprint at https://arxiv.org/pdf/2303.08774. Accessed 7 June 2024

Alkaissi H, McFarlane SI (2023) Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. https://doi.org/10.7759/cureus.35179

Anderson BR, Shah JH, Kreminski M (2024) Homogenization effects of large language models on human creative ideation. Preprint at https://arxiv.org/abs/2402.01536. Accessed 7 June 2024

Bai H, Voelkel J, Eichstaedt J, Willer R (2023) Artificial intelligence can persuade humans on political issues. Res Square. https://doi.org/10.21203/rs.3.rs-3238396/v1

Bashardoust A, Feuerriegel S, Shrestha YR (2024) Comparing the willingness to share for human-generated vs. AI-generated fake news. Preprint at https://arxiv.org/abs/2402.07395. Accessed 7 June 2024

Bauer K, Jussupow E, Heigl R, Vogt B, Hinz O (2024) All just in your head? Unraveling the side effects of generative AI disclosure in creative task, SSRN. https://doi.org/10.2139/ssrn.4782554. Accessed 7 June 2024

Bellaiche L, Shahi R, Turpin MH, Ragnhildstveit A, Sprockett S, Barr N et al (2023) Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. Cogn Res Princ Implic 8(1):42. https://doi.org/10.1186/s41235-023-00499-6

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 610–623. https://doi.org/10.1145/3442188.3445922.

Bingley WJ, Curtis C, Lockey S, Bialkowski A, Gillespie N, Haslam SA, Ko RKL, Steffens N, Wiles J, Worthy P (2023) Where is the human in human-centered AI? Insights from developer priorities and user experiences. Comput Hum Behav 141:107617. https://doi.org/10.1016/j.chb.2022.107617

Brandtzæg PB, Skjuve M, Følstad A (2024) Emerging AI-individualism: how young people integrate social AI into their lives (May 21, 2024). https://ssrn.com/abstract=4436287. Accessed 7 June 2024

Brandtzaeg PB, Skjuve S, Følstad A (2022) My AI friend: How users of a social chatbot understand their human–AI friendship. Hum Commun Res 48(3):404–429. https://doi.org/10.1093/hcr/hqac008

Brandtzaeg PB, Følstad A (2018) Chatbots—changing user needs and motivations. ACM Interact 25(5):38–43. https://doi.org/10.1145/3236669

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. (2020) Language models are few-shot learners. Preprint at https://arxiv.org/abs/2005.14165. Accessed 9 May 2024

Bråten S (1973) Model monopoly and communication: systems theoretical notes on democratization. Acta Sociol 16(2):98–107. https://doi.org/10.1177/000169937301600 2

Bråten S (1983) Dialogens vilkår i datasamfunnet—essays om modellmonopol og meningshorisont i organisasjons- og informasjonssammenheng [The conditions of dialogue in the computer society—essays on model monopoly and the horizon of meaning in organizational and information contexts]. Universitetsforlaget, Oslo

Böhm R, Jörling M, Reiter L, Fuchs C (2023) Content beats competence: people devalue ChatGPT's perceived competence but not its recommendations. https://doi.org/10.31234/osf.io/swfn6

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. (2023) Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://arxiv.org/abs/2303.12712. Accessed 19 May 2024

De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE et al (2023) ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 11:1166120. https://doi.org/10.3389/fpubh.2023.1166120

Deci EL, Ryan RM (2013) The importance of autonomy for development and well-being. Self-regulation and autonomy: social and developmental dimensions of human conduct. Cambridge University Press, New York, pp 19–46

De Gregorio G (2023) The normative power of artificial intelligence. Ind J Glob Legal Stud 30(2): 55. https://ssrn.com/abstract=4436287. Accessed 3 June 2024

De Freitas J, Agarwal S, Schmitt B et al (2023) Psychological factors underlying attitudes toward AI tools. Nat Hum Behav 7:1845–1854. https://doi.org/10.1038/s41562-023-01734-2

DeVerna MR, Yan HY, Yang KC, Menczer F (2023) Fact-checking information generated by a large language model can decrease news discernment. Preprint at https://arxiv.org/abs/2308.10800. Accessed 7 June 2024

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manage Sci 64(3):1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114–126. https://doi.org/10.1037/xge0000033

Floridi L, Cowls J, Beltrametti M et al (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

Goldstein JA, Chao J, Grossman S, Stamos A, Tomz M (2024) How persuasive is AI-generated propaganda? PNAS Nexus 3(2):pgae034. https://doi.org/10.1093/pnasnexus/pgae034

Hackenburg K, Margetts H (2023) Evaluating the persuasive influence of political microtargeting with large language models. https://doi.org/10.31219/osf.io/wnt8b

Hartmann J, Exner Y, Domdey S (2024) The power of generative marketing: can generative AI create superhuman visual marketing content? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4597899. Accessed 6 June 2024

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Preprint at https://arxiv.org/abs/2006.11239. Accessed 7 June 2024

Holton R, Boyd R (2021) 'Where are the people? What are they doing? Why are they doing it?' (Mindell) Situating artificial intelligence within a socio-technical framework. J Sociol 57:179–195. https://doi.org/10.1177/1440783319873046

Inie N, Druga S, Zukerman P, Bender EM (2024) From" AI" to probabilistic automation: how does anthropomorphization of technical systems descriptions influence trust? Preprint at https://arxiv.org/abs/2404.16047. Accessed 1 June 2024

Karinshak E, Liu SX, Park JS, Hancock JT (2023) Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. Proc ACM Human-Comput Interact 7(CSCW1):1–29. https://doi.org/10.1145/3579592

Kim J, Merrill K Jr, Collins C (2021) AI as a friend or assistant: the mediating role of perceived usefulness in social AI vs. functional AI. Telemat Inform 64:101694. https://doi.org/10.1016/j.tele.2021.101694

Krügel S, Ostermaier A, Uhl M (2023) ChatGPT's inconsistent moral advice influences users' judgment. Sci Rep 13(1):4569. https://doi.org/10.1038/s41598-023-31341-0

Lim S, Schmälzle R (2024) The effect of source disclosure on evaluation of AI-generated messages. Comput Human Behav Artif Humans 2(1):100058. https://doi.org/10.1016/j.chbah.2024.100058

Matz SC, Teeny JD, Vaid SS, Peters H, Harari GM, Cerf M (2024) The potential of generative AI for personalized persuasion at scale. Sci Rep 14(1):4692

Meyer A, Riese J, Streichert T (2024) Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. JMIR Med Educ 10:e50965. https://doi.org/10.2196/50965

Mhlanga D (2023) Responsible Industry 4.0: a framework for human-centered artificial intelligence. Taylor & Francis, Oxfordshire

Muñoz JM, Marinaro JÁ (2024) Algorithmic biases: caring about teens' neurorights. AI Soc 39:809–810. https://doi.org/10.1007/s00146-022-01516-w

Narayan A and Kapoor S (2022) ChatGPT is a bullshit generator. But it can still be amazingly useful. AI Snake Oil. https://aisnakeoil.substack.com/p/chatgpt-is-a-bullshit-generator-but. Accessed 4 June 2024

Obrenovic B, Gu X, Wang G et al (2024) Generative AI and human-robot interaction: implications and future agenda for business, society and ethics. AI Soc. https://doi.org/10.1007/s00146-024-01889-0

Ozmen Garibay O, Winslow B, Andolina S et al (2023) Six human-centered artificial intelligence grand challenges. Int J Human-Comput Interact 39(3):391–437. https://doi.org/10.1080/10447318.2022.2153320

Padmakumar V, He H (2023) Does writing with language models reduce content diversity? Preprint at https://arxiv.org/abs/2309.05196. Accessed 7 June 2024

Palmer AK, Spirling A (2023) Large language models can argue in convincing and novel ways about politics: evidence from experiments and human judgement. GitHub. Preprint at https://github.com/ArthurSpirling/LargeLanguageArguments/blob/main/PalmerSpirlingLLMMay182023.pdf. Accessed 7 June 2024

Rozado D (2024) The political preferences of LLMs. Preprint at https://arxiv.org/abs/2402.01789. Accessed 7 June 2024

Sadeghi M (2024) Top 10 generative AI models mimic Russian disinformation claims a third of the time, citing Moscow-created fake local news sites as authoritative sources. NewsGuard. https://www.newsguardtech.com/special-reports/generative-ai-models-mimic-russian-disinformation-cite-fake-news/. Accessed 10 Aug 2024

Salvi F, Ribeiro MH, Gallotti R, West R (2024) On the conversational persuasiveness of large language models: a randomized controlled trial. Preprint at https://arxiv.org/abs/2403.14380. Accessed 7 June 2024

Shah C, Bender EM (2024) Envisioning information access systems: what makes for good tools and a healthy web? ACM Trans Web 18(3):33. https://doi.org/10.1145/3649468

Shank DB, Stefanik C, Stuhlsatz C, Kacirek K, Belfi AM (2023) AI composer bias: listeners like music less when they think it was composed by an AI. J Exp Psychol Appl 29(3):676–692. https://doi.org/10.1037/xap0000447

Sharma N, Liao QV, Xiao Z (2024) Generative echo chamber? Effects of LLM-powered search systems on diverse information seeking. Preprint at. https://arxiv.org/abs/2402.05880. Accessed 7 June 2024

Shin M, Kim J (2024) Large language models can enhance persuasion through linguistic feature alignment (February 13, 2024). https://doi.org/10.2139/ssrn.4725351. Accessed 7 June 2024

Si C, Goyal N, Wu ST, Zhao C, Feng S, Daumé III H, Boyd-Graber J (2023) Large language models help humans verify truthfulness—except when they are convincingly wrong. Preprint at arXiv:2310.12558

Skjuve M, Brandtzaeg PB, Følstad A (2024) Why do people use ChatGPT? Exploring user motivations for generative conversational AI. First Monday https://doi.org/10.5210/fm.v29i1.13541. https://firstmonday.org/ojs/index.php/fm/article/view/13541. Accessed 9 April 2024

Spitale G, Biller-Andorno N, Germani F (2023) AI model GPT-3 (dis)informs us better than humans. Sci Adv 9(26):eadh1850

Sun X, Ma R, Zhao X, Li Z, Lindqvist J, Ali AE, Bosch JA (2024) Trusting the search: unraveling human trust in health information from Google and ChatGPT. Preprint at https://arxiv.org/abs/2403.09987. Accessed 7 June 2024

Sundar SS (2020) Rise of machine agency: a framework for studying the psychology of human–AI interaction (HAII). J Comput-Mediat Commun 25(1):74–88. https://doi.org/10.1093/jcmc/zmz026

Sætra HS (2020) The parasitic nature of social AI: sharing minds with the mindless. Integr Psychol Behav Sci 54:308–326. https://doi.org/10.1007/s12124-020-09523-6

Talbot M (ed) (2019) Language and power in the modern world. Edinburgh University Press, Edinburgh

Tokayev K-J (2023) Ethical implications of large language models: a multidimensional exploration of societal, economic, and technical concerns. Int J Soc Anal 8(9): 17–33. https://norislab.com/index.php/ijsa/article/view/42. Accessed 7 June 2024

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017) Attention is all you need. In: Advances in neural information processing systems 30. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed 7 June 2024

Verdegem P (2024) Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. AI Soc 39:727–737. https://doi.org/10.1007/s00146-022-01437-8

Weber M (1978) Economy and society: an outline of interpretive sociology. University of California Press, California

Wu Y (2024) Critical thinking pedagogics design in an era of ChatGPT and other AI tools—shifting from teaching "what" to teaching "why" and "how." J Educ Dev. https://doi.org/10.20849/jed.v8i1.1404

Yang JC, Korecki M, Dailisan D, Hausladen CI, Helbing D (2024) LLM voting: human choices and AI collective decision making. Preprint at https://arxiv.org/abs/2402.01766. Accessed 7 June 2024

Zhang Y, Gosline R (2023) Human favoritism, not AI aversion: people's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. Judgm Decis Mak 18:e41. https://doi.org/10.1017/jdm.2023.37

Zhou K, Hwang JD, Ren X, Sap M (2024) Relying on the unreliable: the impact of language models' reluctance to express uncertainty. Preprint at https://arxiv.org/abs/2401.06730. Accessed 7 June 2024