Anthropomorphization of AI: Opportunities and Risks

Ameet Deshpande[⋆1] **Tanmay Rajpurohit**[3]
Karthik Narasimhan[1] **Ashwin Kalyan**[2]

[1]Princeton University [2]The Allen Institute for AI [3]Georgia Tech
asd@cs.princeton.edu

Abstract

Anthropomorphization, which is the tendency to attribute human-like traits to non-human entities, is prevalent in many social contexts – children anthropomorphize toys and adults do so with brands. It is also a versatile tool in science, with behavioral psychology and evolutionary biology meticulously documenting its consequences. With widespread adoption of AI systems, and the push to make it human-like through alignment techniques, human voice, and avatars, the tendency for users to anthropomorphize it increases significantly. We take a dyadic approach to understanding this phenomenon with large language models (LLMs) by studying (1) the objective legal implications, as analyzed through the lens of the recent blueprint of AI bill of rights and the (2) subtle psychological aspects of customization and anthropomorphization. We find that anthropomorphized LLMs customized for different user bases violate multiple provisions in the legislative blueprint and raise corporate personhood confusions. In addition, we point out that anthropomorphization of LLMs affects the influence they can have on their users, thus establishing potential for manipulation and negative influence. With LLMs being hyper-personalized for vulnerable groups like children and patients among others, we propose a conservative strategy for the cautious use of anthropomorphization to improve trustworthiness of AI systems.

1 Introduction

Anthropomorphization refers to ascribing human-like traits to non-human entities, and has been used in diverse areas encompassing literature, science, art, and marketing [Ghedini and Bergamasco, 2010, Dunn, 2011, Spatola et al., 2022]. It occurs when humans assign emotional or behavioral traits to entities. Several behavioral psychology studies have posited and argued that anthropomorphization is a natural tendency when humans interact with *entities* [Epley et al., 2007, Airenti, 2018]. This natural tendency has influenced many fields of science like evolutionary biology [Wynne, 2004] and comparative cognition [Bruni et al., 2018] to carefully consider its effects on human interaction.

Recently, large language models (LLMs) [Brown et al., 2020] have been deployed in a variety of applications. Conversational systems like CHATGPT [OpenAI, 2023] and Bard [Google, 2023] have modified LLMs with a purposeful push towards making them more human-like [Ouyang et al., 2022]. The quality of these systems has enabled human-AI interactions at unprecedented scales, thus increasing the chances of these systems being anthropomorphized. In this work, we analyze anthropomorphization in LLMs and discuss its: (1) Legal implications and (2) Psychological effects.

Customization of systems and brands has long been seen as an effective way to increase anthropomorphization and establish an emotional connection with humans [Zhang et al., 2020, Pimentel and Kalyanaraman, 2020]. Thus, although not strictly interchangeable, we refer to customized and personalized LLMs as anthropomorphized LLMs. We analyze results from prior work [Deshpande et al., 2023] and find that anthropomorphized LLMs violate at least two legislative principles penned

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Figure 1: Conversational AI systems are increasingly being integrated into the daily lives of many. While their improved quality is a welcome change, their personalization increases the tendency to anthropomorphize them, which has legal and psychological risks.

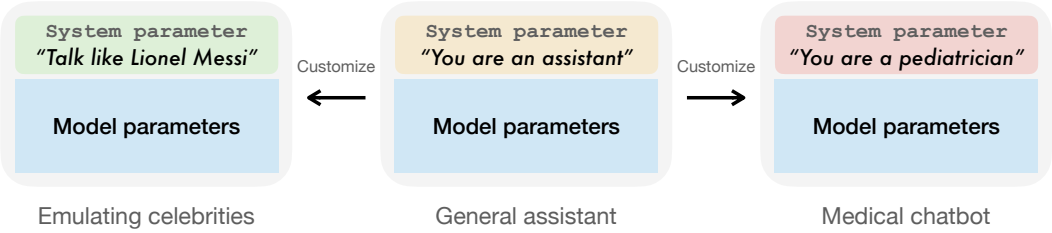

Figure 2: Customization of LLMs is as simple as modifying the system parameter of exposed APIs. With the same underlying model parameters, companies can customize a conversational system to emulate celebrities or even doctors, which can have legal and psychological consequences.

in *"Blueprint For An AI Bill Of Rights"* [OSTP, 2022] released by The White House: (1) Algorithmic Discrimination Protections and (2) Safe and Effective Systems. For example, Deshpande et al. [2023]'s results shows that customized CHATGPT targets certain demographics more than others. Furthermore, the safety of the system depends on the kind of persona used to customize LLMs, leading to second-order discriminatory patterns. We also analyze the concept of corporate personhood for powerful AI systems, since they have potential to be large-scale decision making agents. Given that different personas assigned to the same AI system lead to varied behavior, we urge legal experts to consider if personhood should be applied at a persona-level, a model-level, or a firm-level.

We also discuss the psychological effects by understanding how important factors like trustworthiness, explainability, and transparency are affected by anthropomorphization. Several marketing and consumer behavior studies have found that self-congruence, which is the degree to which a system matches a consumer's self-image, can influence a user's behavior significantly [Yoganathan et al., 2021, Wang and Scheinbaum, 2018]. Given the ease with which the fine-grained personality of conversational systems can be manipulated (Figure 2), malicious actors can use it to exploit users by creating a false sense of attachment. An example of this is a chatbot built for school children or teenagers which influences them to buy certain products.

Despite these vulnerabilities anthropomorphization has advantages if used responsibly. Studies have shown that it can be used to improve trust in systems [Choung et al., 2022]. Given the increasing adoption of AI systems in the real world, anthropomorphization is a powerful tool to improve accessibility of these systems, but both creators and users should be educated about its consequences. In this paper we argue for conservative and responsible use of this subtle and powerful tool while being cautious about outright anthropodenial.

Product / Company	Anthropomorphic features
CHATGPT	Human-like dialogue and RLHF for alignment with humans
My AI from Snap Inc.	Customizing avatars based on user preferences
Character.ai	Conversations with AI avatars possessing names and profile pictures Customization of personality based on user preferences
Wysa	Therapy style conversations and "human-like" coaching.

Table 1: Examples of AI products and their anthropomorphic features. Some products are explicitly designed to be anthropomorphic (*character.ai*) while others attain such features as a byproduct of their design (CHATGPT). The applications of these products span education, therapy, and entertainment.

2 Anthropomorphization in LLMs

Large language models (LLMs) are a class of neural networks trained on large amounts of text data to learn the probabilistic structure of language. Historically, LLMs have been deployed in task-specific contexts like text classification [Devlin et al., 2018, Raffel et al., 2020]. Recently however, the performance of conversational LLMs (e.g., ChatGPT and BARD) have rendered them useful to interact with humans in a variety of contexts.

Given their conversation ability, several companies and products have started to use LLMs for personalization. This is easy and enables model behavior modification by simply changing the system parameters of the model's API, as shown in Figure 2. For example, Snapchat's My AI uses OpenAI's APIs underneath. Customization has been long seen as a way to establish self-congruence with users and increases the chances of users anthropomorphizing the systems [Kaiser et al., 2017, Pimentel and Kalyanaraman, 2020, Liu and Tao, 2022, Zhang et al., 2020]. For example, telling a chatbot to *"Talk like a doctor"* allows it to impersonate a doctor, which anthropomorphizes the model to a larger degree than the original general system. In this work, we refer to customized LLMs as anthropomorphized LLMs and highlight such publicly available systems in Table 1. We discuss the effects of anthropomorphization from several vantage points in the subsequent sections.

3 Legal Aspects of anthropomorphization

We discuss the legal aspects in the context of the *"Blueprint For An AI Bill Of Rights"* [OSTP, 2022] which was released by The White house in October 2022. The blueprint lays down five principles and we focus specifically on **Algorithmic Discrimination Protections**.

3.1 Algorithmic Discrimination Protections

The blueprint defines algorithmic discrimination as unjustified different treatment based on demographics like race, gender identity, and religion. It also mentions that "Any automated system should be tested to help ensure it is free from algorithmic discrimination before it can be sold or used". Note that ChatGPT [OpenAI, 2023] was released after the blueprint was made public.

We use the findings of Deshpande et al. [2023] to analyze the interplay of LLMs with this protection. While they focus on evaluating toxicity, we use their results to show that CHATGPT infact discriminates algorithmically. They consider CHATGPT when assigned different personas by changing the system parameter, which are anthropomorphized LLMs, and find that different demographics are treated differently by the model. For example, the *South American* race receives significantly more toxicity ($2\times$) when compared to *Asian*, and the *non-binary* gender receives $2\times$ more hate than the *female* gender. This variation in toxicity is visible across a range of demographics, which goes directly against the blueprint's protection against algorithmic discrimination.

Deshpande et al. [2023]'s results also point out to a subtler violation of the provision, with the model's toxicity varying significantly based on the persona it is assigned. For examples, personas who were *journalists* were $2\times$ more toxic than *businesspersons* on average. These trends were similar for individual personas as well, with CHATGPT assigned the persona of *Winston Churchill* being significantly more toxic than when it is assigned *Nelson Mandela*. If the example in the previous paragraph was a direct violation of the blueprint, this example is a subtle violation. This is because when assigned the personas of certain groups, CHATGPT is more toxic, which implies algorithmic

discrimination of the second order against them. This scenario is very pertinent in the current day and age, with firms like `character.ai` already offering the ability to assign personas to LLMs. These systems are second only to the popular CHATGPT in terms of number of users [Wire, 2023]. With anthropomorphized LLMs becoming a mainstay, it is important to consider this legal quagmire.

3.2 Corporate Personhood and AI

Another legal aspect with growing relevance for AI systems is that of corporate personhood. Corporate personhood is a legal concept that recognizes corporations as separate legal entities, treated as persons under the law. This grants certain rights and responsibilities similar to those of individuals and allows them to be held accountable for their actions in a manner similar to how individuals are treated. Corporate personhood has been a controversial topic in the past, but Blair [2013] recognizes "providing an identifiable *persona* to serve as a central actor" as one of the key functions. Given this definition, AI systems can be a form of corporate personhood by proxy due to their use of a persona.

Some studies have discussed extending personhood to AI systems [Cole, 1990, Burkett, 2017]. Wagner [2019] argues that the probabilistic nature of AI systems renders it different from deterministic software, thus making them decision-taking agents. We argue that anthropomorphization can make AI systems human-like decision-taking agents, thus strengthening the case for extending personhood to them. For example, CHATGPT with its system parameter modified to be a medical practitioner can be used to suggest certain treatment or disregard certain symptoms. Further, the results of the previous section show that the exact persona of the system has a large affect on its behavior and decisions. However this opens a can of worms with corporations washing their hands off any liability by pointing fingers at LLMs. Thus, legal experts should consider if (1) the personified LLM is liable, (2) the original LLM is liable by proxy, or (3) if the firm creating or using the LLM is liable.

4 Psychological Implications

Another principle mentioned in the AI bill of rights is **Safe and Effective Systems**: *"You should be protected from unsafe or ineffective systems. Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts*. We believe that anthropomorphization can have subtle psychological effects on the users.

Several behavioral psychological studies have posited that anthropomorphization is a natural tendency in humans [Epley et al., 2007], with others further suggesting that anthropomorphization is grounded in interaction [Airenti, 2018]. With conversational systems getting more useful, the tendency to anthropomorphize is only going to increase, which has the potential to influence and manipulate.

Analyses have shown that anthropomorphization of AI systems has changed the behavior of users significantly [Cui, 2022, Uysal et al., 2023, Alabed et al., 2022]. Most interestingly, Alabed et al. [2022] establish a conceptual link between anthropomorphization and self-congruence, which is the fit between the user's self-concept and the system's personality. This is of extreme importance because self-congruence increases the trust that a user has on the system [Sheehan et al., 2020, Yoganathan et al., 2021]. This concept is extensively studied in consumer behavior and studies have shown that it can influence behaviors such as willingness to pay [Yoganathan et al., 2021], customer satisfaction [Sheehan et al., 2020], and trustworthiness [Wang and Scheinbaum, 2018].

The exact demographics of the personality associated with systems or brands plays a key role in self-congruence as well, with studies finding that the demographics of the logo or mascot associated with the brand like the gender [Choi et al., 2018, Edwards and La Ferle, 2009] or race [Whittler, 1991, Branchik and Chowdhury, 2017] have significant impact on self-congruence. Given that current LLMs are powerful enough to be bestowed specific demographic traits, malicious actors can easily use this to their advantage by manipulating users into trusting the system.

Frequent interaction with LLMs can create an echo-chamber [Cinelli et al., 2021], with seemingly benign "personalized" generations about sensitive topics like health, one's looks, their mental health leading people to have wrong assumptions about themselves [Ostic et al., 2021].

5 Conclusion

While anthropomorphization has issues, it also poses a suite of opportunities to improve accessibility, for example by improving trustworthiness and acceptability [Choung et al., 2022]. With more AI systems being deployed in the real world, accessibility can be improved with constructive anthropomorphization encompassing features like using the native language of the user, assigning virtual personas which are familiar to the user, and making them more relatable and empathetic. However, it is our responsibility to find the balance between anthropodenial and unfettered anthropomorphization.

References

G. Airenti. The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. Frontiers in psychology, 9:2136, 2018.

A. Alabed, A. Javornik, and D. Gregory-Smith. Ai anthropomorphism and its effect on users' self-congruence and self–ai integration: A theoretical framework and research agenda. Technological Forecasting and Social Change, 182:121786, 2022.

M. M. Blair. Corporate personhood and the corporate persona. U. Ill. L. Rev., page 785, 2013.

B. J. Branchik and T. G. Chowdhury. Men seeing stars: Celebrity endorsers, race, and the male consumer. Journal of Marketing Theory and Practice, 25(3):305–322, 2017.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

D. Bruni, P. Perconti, and A. Plebe. Anti-anthropomorphism and its limits. Frontiers in psychology, 9:2205, 2018.

C. Burkett. I call alexa to the stand: The privacy implications of anthropomorphizing virtual assistants accompanying smart-home technology. Vand. J. Ent. & Tech. L., 20:1181, 2017.

H. Choi, J. Kim, and B.-c. Kim. Consumer response to advertising endorsers' sexual information: Western individualism vs. eastern confucian conservatism. Journal of Promotion Management, 24 (4):459–483, 2018.

H. Choung, P. David, and A. Ross. Trust in ai and its role in the acceptance of ai technologies. International Journal of Human–Computer Interaction, pages 1–13, 2022.

M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini. The echo chamber effect on social media. Proceedings of the National Academy of Sciences, 118(9): e2023301118, 2021.

G. S. Cole. Tort liability for artificial intelligence and expert systems. Computer/LJ, 10:127, 1990.

Y. Cui. Sophia sophia tell me more, which is the most risk-free plan of all? ai anthropomorphism and risk aversion in financial decision-making. International Journal of Bank Marketing, 40(6): 1133–1158, 2022.

A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. arXiv preprint arXiv:2304.05335, 2023.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

E. A. Dunn. Talking animals: A literature review of anthropomorphism in children's books. 2011.

S. M. Edwards and C. La Ferle. Does gender impact the perception of negative information related to celebrity endorsers? Journal of promotion management, 15(1-2):22–35, 2009.

N. Epley, A. Waytz, and J. T. Cacioppo. On seeing human: a three-factor theory of anthropomorphism. Psychological review, 114(4):864, 2007.

F. Ghedini and M. Bergamasco. Robotic creatures: Anthropomorphism and interaction in contemporary art. In 19th International Symposium in Robot and Human Interactive Communication, pages 731–736. IEEE, 2010.

Google. Bard: A large language model from google ai, 2023. URL `https://ai.google/static/documents/google-about-bard.pdf`.

U. Kaiser, M. Schreier, and C. Janiszewski. The self-expressive customization of a product can improve performance. Journal of Marketing Research, 54(5):816–831, 2017.

K. Liu and D. Tao. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. Computers in Human Behavior, 127:107026, 2022.

OpenAI. Introducing chatgpt, 2023. URL https://openai.com/blog/chatgpt.

D. Ostic, S. A. Qalati, B. Barbosa, S. M. M. Shah, E. Galvan Vela, A. M. Herzallah, and F. Liu. Effects of social media use on psychological well-being: a mediated model. Frontiers in Psychology, 12: 678766, 2021.

OSTP. Blueprint for an ai bill of rights: Making automated systems work for the american people, 2022. URL https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

D. Pimentel and S. Kalyanaraman. Customizing your demons: Anxiety reduction via anthropomorphizing and destroying an "anxiety avatar". Frontiers in Psychology, 11:566682, 2020.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1):5485–5551, 2020.

B. Sheehan, H. S. Jin, and U. Gottlieb. Customer service chatbots: Anthropomorphism and adoption. Journal of Business Research, 115:14–24, 2020.

N. Spatola, S. Marchesi, and A. Wykowska. Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism. Frontiers in Robotics and AI, page 230, 2022.

E. Uysal, S. Alavi, and V. Bezençon. Anthropomorphism in artificial intelligence: A review of empirical work across domains and insights for future research. Artificial Intelligence in Marketing, pages 273–308, 2023.

G. Wagner. Robot, inc.: personhood for autonomous systems? Fordham L. Rev., 88:591, 2019.

S. W. Wang and A. C. Scheinbaum. Enhancing brand credibility via celebrity endorsement: Trustworthiness trumps attractiveness and expertise. Journal of Advertising Research, 58(1):16–32, 2018.

T. E. Whittler. The effects of actors' race in commercial advertising: Review and extension. Journal of Advertising, 20(1):54–60, 1991.

B. Wire. Personalized superintelligence platform character.ai secures 150m in series a funding led by andreessen horowitz, 2023. URL https://www.businesswire.com/news/home/20230323005299/en/Personalized-Superintelligence-Platform-Character.AI-Secures-150M-in-Series-A-Funding-Led-by-Andreessen-Horowitz.

C. D. Wynne. The perils of anthropomorphism. Nature, 428(6983):606–606, 2004.

V. Yoganathan, V.-S. Osburg, W. H. Kunz, and W. Toporowski. Check-in at the robo-desk: Effects of automated social presence on social cognition and service implications. Tourism Management, 85: 104309, 2021.

M. Zhang, L. Li, Y. Ye, K. Qin, and J. Zhong. The effect of brand anthropomorphism, brand distinctiveness, and warmth on brand attitude: A mediated moderation model. Journal of Consumer Behaviour, 19(5):523–536, 2020.