Posthumanist AI

Rethinking ‘the human’ as a model for generative AI

MATT RATTO Faculty of Information, University of Toronto, Toronto, Canada, matt.ratto@utoronto.ca

SARAH GRAM Faculty of Information, University of Toronto, Toronto, Canada, sara.gram@mail.utoronto.ca

OLIVIA DOGGETT Faculty of Information, University of Toronto, Toronto, Canada, olivia.doggett@mail.utoronto.ca

PETER SELBY Center for Mental Health and Addiction, Toronto, Canada, peter.selby@camh.ca

NADIA MINIAN Center for Mental Health and Addiction, Toronto, Canada, Nadia.Minian2@camh.ca

MARTA MASLEJ Center for Mental Health and Addiction, Toronto, Canada, Marta.Maslej@camh.ca

OSNAT MELAMED Center for Mental Health and Addiction, Toronto, Canada, osnat.melamed@camh.ca

JONATHAN ROSE, Faculty of Engineering, Toronto, Canada, jonathan.rose@utoronto.ca

The power of generative AI systems is often described as their ability to produce ‘human-like’ creative outputs, including images and text. But what is considered ‘human-like’? What humans and behaviors serve as the models for generative AI systems? And how might unexamined concepts of ‘humans’ impact the design and operation of AI systems? Scholars have described how reductive concepts of humanness have been the source of many inequities in society including apartheid [1], colonialism)[2, 3, 4], and gender inequality)[5, 6, 7, 8]. To advance our understanding of humanness in the context of generative AI, three empirical questions are important: what kinds of humans serve as the models for AI; what human traits are considered appropriate and appropriable; and how are these traits operationalized within specific AI systems? We use these questions to support an initial engagement with the concept of posthumanist AI, building on existing and growing posthumanist scholarship. If we are ourselves ‘posthuman’ as Katherine Hayles has defended, [9] what are the issues with designing and building generative AI systems based on a reductive vision of humanness? What possibilities unfold when we open our minds to a more expansive vision?

CCS CONCEPTS •Human-centered computing~Collaborative and social computing~Collaborative and social computing theory, concepts and paradigms

Additional Keywords and Phrases: generative AI, humanness, design, posthumanism

ACM Reference Format:

Matt Ratto, Sarah Gram, Olivia Doggett, Peter Selby, Nadia Minian, Osnat Melamed, Jonathan Rose. 2024. Posthumanist AI: Rethinking ‘the human’ as a model for generative AI. In GenAICHI: CHI 2024 Workshop on Generative AI and HCI.

GenAICHI: CHI 2024 Workshop on Generative AI and HCI

1 INTRODUCTION

Most of us have been amazed at what new transformer-based NLP systems can do. Whether it is gas-lighting a reporter on Valentine's Day to convince him to leave his wife, [10], acing standardized tests (except English), [11] or telling jokes [12], enthusiasm regarding LLM-powered conversation agents have centered around their apparent capacity to generate creative outputs, learn by example, carry out introspective reflections, engage in chain-of-thought reasoning, and even deceive and manipulate users. These technologies are troubling established (though constantly renegotiated) boundaries between humans and computational systems, primarily by generating forms of social interaction that were previously only considered possible between humans -- and sometimes extended to other biological entities such as pets and other companion species. [13] Moving beyond ontological questions regarding AI and humans, we see a need to more closely examine what work is being done, deliberately or accidentally, by using humanness as a criterion for designing and evaluating AI systems. Here, we follow Suchman's guidance to move from: "…categorical debates to empirical investigations of the concrete practices through which categories of human and nonhuman are mobilized" [14]. To begin this work, we briefly focus on three main questions, which we use to explore why defining humanness remains a key concern. First, how did we come to correlate humanness and AI? Second, how are humans/humanness interpolated within AI systems? Finally, how is humanness defined within technical literature associated with the evaluation of generative AI systems? Noting its underdetermined nature within the development of generative AI, we highlight the need to better define what counts as 'human' both for generative AI and for ourselves and end with a suggested starting place of posthumanism.

2 BACKGROUND

2.1 How did we come to correlate humanness and ai?

We have a long history in building things that represent human behaviors and capabilities with linguistic ability often serving to define what remains purely human. Both Descartes and Turing used the capability of language to draw a clear line between humans and other beings including animals and machines. Descartes notes that even if we could build a machine that perfectly mimicked human actions and appearance "…they could never use words or other signs arranged in such a manner as is competent to us in order to declare our thoughts to others" [15] Similarly, Turing posited his famous test which leveraged language use and its evaluation as a way to produce an empirically determinant way to address machine cognition:

> "We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'" [16]

Correlating language ability with language to intelligence continues to shape our understanding of machine cognition. But here is where we run into the first issue. Attempts to deploy Turing's test in applied settings have often resulted in confusions about what counts as human. An attempt to administer an annual Turing Test was created in 1991. Called the Loebner Prize, it pitted humans recruited via a newspaper advertisement against multiple computational agents. Stuart Shieber, writing about the results of the first contest, emphasized some of the challenges. He noted:

> "Ms. Cynthia Clay, the Shakespeare aficionado, was thrice misclassified as a computer. At least one of the judges made her classification on the premise that (no) human would have that amount of knowledge about Shakespeare… Ms. Lisette Gozo was honored as the most human of the agents for her discussion of women's clothing, although one judge rated two computer programs above her." [17]

GenAICHI: CHI 2024 Workshop on Generative AI and HCI

In the quotes above, we see two problems; first that a human showing too much knowledge is mischaracterized as a computer; second, that knowledge of a certain sort – women's clothing – is considered 'more human.' Both examples show how unexamined ideas of 'humanness' can result in forms of bias. The judges' assumptions about what female human knowledge is, e.g clothes and not Shakespeare, ends up impacting their comparison and ultimately what counts as human vs. machine.

2.2 How are humans/humanness interpolated within AI systems?

2.2.1 Choices of datasets - what language being used to train?

In their now canonical paper on "Stochastic Parrots", Bender, Gebru, McMillan-Major, and Mitchell note that "it is easy to imagine that very large datasets, such as Common Crawl "…petabytes of data collected over 8 years of web crawling" a filtered version of which is included in the GPT-3 training data, must therefore be broadly representative of the ways in which different people view the world" [18]. The authors argue that this fantasy is, in fact, untrue.

> …white supremacist and misogynistic, ageist, etc. views are overrepresented in the training data, not only exceeding their prevalence in the general population but also setting up models trained on these datasets to further amplify biases and harms.[18]

They attribute this to a number of factors, including the primary use of Reddit data in training data sets, systematic exclusionary tactics that reduce the participation by marginalized populations in online user-generated content more broadly, and even techniques used to 'clean' and 'filter' this data that, while intended to remove pornography and hate speech, also work to delete texts from online sites built for and by LGBTQ people. Monolinguistic data, mainly English, and primarily western and represented by Latin characters [19] further encodes systematic bias into AI systems.

2.2.2 What counts as human - how is humanness evaluated?

In addition to the data used to train models, another way that humanness is brought into LLMs is through its use as a criterion for evaluating models. This is a less examined aspect of LLMs – what conceptual tools are used to determine the quality of their outputs? A search on the ACM Digital Archive reveals that 'humanness' is highly considered in ACM papers that also reference LLMs. A basic search on the terms llm OR LLM AND 'human-like' OR 'human-mimicking' results in 326,979 hits. However, in most of these papers what is meant by 'humanness' or 'human-like'-ness is never defined. Instead, it is assumed that readers and authors share a common understanding. Other articles rely on use surveys [20, 21, 22] following interactions with generative AI systems to evaluate how 'human-like' users found the bot. While 'humanness' serves as a category of evaluation, what it means remains underdetermined, leaving decisions over what counts as human-like behavior to the evaluator. This would be fine if there was consistent agreement as to what attributes are included, but it seems clear from the work described above that there is a need to unpack and denaturalize this category.

3 DEFINING 'HUMANNESS'

The brief answers to the three questions posed above highlight the need to better define humanness in relation to AI as we build, train, and evaluate these systems.

3.1 Posthumanism

Posthumanism, as distinct from transhumanism, [23] offers a radical denaturalization of the human that seems necessary, given the ways generative AI systems 'work the boundary' between human and machines. Posthumanism is exemplified in concepts such as Donna Haraway's "cyborg" [6], which is, at its core, a critique and rejection of the boundaries separating human, machine, and animal and an insistence on the plurality and multiplicity of the subject. Braidotti [7, 8]) has outlined the ramifications and potentialities of this, including the need to move beyond simplistic understandings of individualistic subjectivities, a recognition of the systematic or 'networked' nature of identity [24], and the importance of 'situated knowledge' [25]. Further, Braidotti draws upon the anti-humanism of post-structuralist and anti-colonial scholars [2, 3, 4, 5] to examine the privileging of certain aspects of 'humanness', including rationality, reason, and a related set of "mental, discursive and spiritual values." Unexamined ideas of humanness highlight the ways in which normalizing humans within a white and unexamined ideal supports political and cultural acts of colonization.

The goal of posthumanist scholarship is to reopen social conventions to critique, allowing for more diversity of human subject positions including those of non-white, non-European, and non-heterosexual identities, as well as humans with non-neurotypical and divergent bodies. Mellamphy has called for a similar engagement with AI, noting:

> "While most discussions of AI are still anchored in humanistic/humancentric narratives, there are reasons for rejecting this model and for turning to alternative worldviews." [26]

Just as Roe and Marathe have called for better definitions of race in HCI research [27] we see a need to engage similarly with definitions of humanness. The posthumanist literature is key, not just to unpacking and reopening our normative and potentially biased concepts of humanness, but also in positing alternatives. Concepts drawn from this literature can provide starting points for the design of generative AI systems in ways that do not reify existing cultural biases through normative and assumed concepts of humanness. We are currently exploring such terms as assemblage, cyborg, embodiment, emergence, co-individuation, and others in relation to an applied project building a conversational agent [28] and will publish the results in longer work.

CONCLUSION

In Vandana Shiva's seminal work she describes the portrayal of science as universal, as a 'monoculture of the mind', which works to reduce and denaturalize indigenous ways of knowing [29]. In naturalizing certain assumptions about humans and human capacities within AI, we risk creating another kind of 'monoculture of the mind' where only certain types of interactions and forms of behavior are considered appropriately human. Could LLMs end up producing a kind of 'cognitive imperialism'? More optimistically, what forms of cognition and interaction might we develop if we work from a more complex and equitable concept of humanness?

ACKNOWLEDGMENTS

The authors would like to thank the CAMH and U of Toronto MIBot team for their continued conversations and reflections on these themes.

REFERENCES

[1] George C. Bowker and Susan Leigh Star. 2000. *Sorting Things out: Classification and Its Consequences*. MIT Press.

[2] Franz Fanon. 1967. *Black Skin, White Masks*. Grove Press, New York.

[3] Luce Irigaray. 1977. *This Sex Which Is Not One.* Cornell University Press.

[4] Edward Said. 1978. *Orientalism*. (25th Anniversay Edition). Vintage Books.

[5] Gayatri Chakravorty Spivak. 1998. Can the subaltern speak? In Cary Nelson and Lawrence Grossberg (eds). *Marxism and the Interpretation of Culture*. Macmillan, London.

[6] Donna Haraway. 1991. *Simians, Cyborgs and Women: The Reinvention of Nature.* Routledge, New York.

[7] Rosi Braidotti. 2013. *The Posthuman*. Polity Press, New York.

[8] Rosi Braidotti. 2019. *Posthuman Knowledge*. Polity Press:,New York

[9] N. Katherine Hayles. 1999. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics.* University of Chicago Press, Chicago.

[10] Kevin Roose. Feb 7, 2023. A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *The New York Times*. Retrieved from https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

[11] GPT-4. Retrieved February 26, 2024 from https://openai.com/research/gpt-4

[12] GPT-4 is bigger and better than ChatGPT—but OpenAI won't say why. MIT Technology Review. Retrieved February 25, 2024 from https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai/

[13] Donna Haraway. 2003. *The Companion Species Manifesto*. University of Chicago Press, Chicago.

[14] Lucy Suchman. 2006. *Human-Machine Reconfigurations: Plans and Situated Actions* (2 edition ed.). Cambridge University Press, Cambridge; New York.

[15] Descartes, R. (1975). A Discourse on Method: Meditations on the First Philosophy; Principles of Philosophy. Dent.

[16] Turing, A.M. (1950) Computing machinery and Intelligence, Mind, A Quarterly Review (VOL. LIX. NO. 236. October, 1950)

[17] Shieber, S. (1994) Lessons from a Restricted Turing Test, Communications of the Association for Computing Machinery, volume 37, number 6, pages 70-78.

[18] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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 (FAccT '21), March 01, 2021, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 610–623. . https://doi.org/10.1145/3442188.3445922

[19] D. Pargman and J. Palme, "ASCII imperialism," in Standards and Their Stories : How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life, Lampland, Martha och Leigh Star, Susan Ed., Ithaca : Cornell University Press, 2009, pp. 177-199.

[20] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. Int J of Soc Robotics 1, 1 (January 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[21] N. Haslam, S. Loughnan, Y. Kashima, and P. Bain. 2008. Attributing and denying humanness to others. European Review of Social Psychology, Vol. 19, 1 (2008), 55--85. https://doi.org/10.1080/10463280801981645

[22] Haslam, N., Loughnan, S., & Holland, E. (2013). The psychology of humanness. In S. J. Gervais (Ed.), Objectification and (de)humanization: 60th Nebraska symposium on motivation (pp. 25–51). Springer Science + Business Media. https://doi.org/10.1007/978-1-4614-6959-9_2

[23] Ron Cole-Turner. 2022. Posthumanism and Transhumanism. In *Encyclopedia of Religious Ethics*. John Wiley & Sons, Ltd, 1098–1105. https://doi.org/10.1002/9781118499528.ch122

[24] Bruno Latour. 2007. *Reassembling the Social: An Introduction to Actor-Network-Theory.* Oxford University Press, Oxford.

[25] Donna Haraway. 1988. Situated Knowledges: the Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14(3), 575-599. https://doi.org/10.2307/3178066

[26] N. Mellamphy. 2021. Re-thinking "Human-centric" AI: An Introduction to Posthumanist Critique. Europe Now 45 (2021).

[27] Devon Roe and Megh Marathe. 2023. Examining Race in Healthcare-Focused HCI Research. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (*CSCW '23 Companion*), October 14, 2023, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 22–26. . https://doi.org/10.1145/3584931.3607021

[28] Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, Vidya Sujaya, Jodi Wolff, Olivia Doggett, Mathew Iantorno, Matt Ratto, Peter Selby, and Jonathan Rose. 2023. A Motivational Interviewing Chatbot With Generative Reflections for Increasing Readiness to Quit Smoking: Iterative Development Study. *JMIR Mental Health* 10, 1 (October 2023), e49132. https://doi.org/10.2196/49132

[29] Vandana Shiva. 1993. *Monocultures of the Mind.* Zed Books Ltd, London