Which Humans?

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, Joseph Henrich

Department of Human Evolutionary Biology, Harvard University

Author Notes

We thank Thomas Talhelm for sharing data, Ali Akbari for feedback on analyses, and Frank Kassanits for suggesting we explore this topic. This work was partly funded by the John Templeton Foundation (#62161).

Author for correspondence: Mohammad Atari, Department of Psychological and Brain Sciences, University of Massachusetts Amherst, 135 Hicks Way, Amherst, MA 01003. Email: matari@umass.edu

Abstract

Large language models (LLMs) have recently made vast advances in both generating and analyzing textual data. Technical reports often compare LLMs' outputs with "human" performance on various tests. Here, we ask, "Which humans?" Much of the existing literature largely ignores the fact that humans are a cultural species with substantial psychological diversity around the globe that is not fully captured by the textual data on which current LLMs have been trained. We show that LLMs' responses to psychological measures are an outlier compared with large-scale cross-cultural data, and that their performance on cognitive psychological tasks most resembles that of people from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies but declines rapidly as we move away from these populations (r = -.70). Ignoring cross-cultural diversity in both human and machine psychology raises numerous scientific and ethical issues. We close by discussing ways to mitigate the WEIRD bias in future generations of generative language models.

Keywords: Culture, Human Psychology, Machine Psychology, Artificial Intelligence, Large Language Models.

Introduction

The notion that AI systems can possess human-like traits is hardly a new observation. Given the increasing societal role played by Large Language Models (LLMs), researchers have begun to investigate the underlying psychology of these generative models. For example, several works have investigated whether LLMs can truly understand language and perform reasoning (Chowdhery et al., 2022), understand distinctions between different moralities and personalities (Miotto et al., 2022; Simmons, 2022), and learn ethical dilemmas (Jiang et al., 2021). Hagendorff et al. (2022), for instance, demonstrated that LLMs are intuitive decision makers, just like humans, arguing that investigating LLMs with methods from psychology has the potential to uncover their emergent traits and behavior. Miotto et al. (2022) found that Generative Pre-trained Transformer-3 (GPT-3) contains an "average personality" compared with "human" samples, has values to which it assigns varying degrees of importance, and falls in a relatively young adult demographic. Horton (2023) argued that LLMs are implicit computational models of humans—a *Homo silicus*. He then suggests that these models can be used in the same manner that economists use *Homo economicus*: LLMs can be given information and preferences, and then their behavior can be explored via simulations. LLMs have also been found to be able to attribute beliefs to others, an ability known as Theory of Mind (Trott et al., 2023; Kosinski, 2023). Finally, Bubeck et al. (2023) recently made the case that there are sparks of general intelligence (i.e., general mental capability including the ability to reason coherently, comprehend complex ideas, plan for the future, solve problems, think abstractly, learn quickly, and learn from experience; Gottfredson, 1997) in GPT-4 (OpenAI, 2023) which was trained using an unprecedented scale of both computational power and data. Some social scientists have gone even further, arguing for LLMs as potential replacements for human participants in psychological research (Dillion et al., 2023; Grossmann et al., 2023).

In the growing literature on probing the psychology of LLMs (see Shiffrin & Mitchell, 2023), researchers have repeatedly argued that these systems respond in ways that are cognitively and attitudinally similar to "humans." For example, the GPT-4 technical report (OpenAI, 2023) introduces GPT-4 (the latest version of the LLM that powers OpenAI's popular chatbot, ChatGPT) as "a large

multimodal model with human-level performance on certain difficult professional and academic benchmarks." Bubeck et al. (2023) mention that "GPT-4's performance is strikingly close to *human*-level performance" and that "GPT-4 attains *human*-level performance on many tasks [...] it is natural to ask how well GPT-4 understands *humans* themselves." Scholars from social sciences (e.g., psychology, economics) have used the same terminology to compare LLMs and "humans." For instance, to showcase the economic decision-making capabilities of LLMs, Horton (2023) argues that they "can be thought of as implicit computational models of *humans*." In quantifying LLMs' personality and moral values, Miotto et al. (2022) argue that GPT-3 "scores similarly to *human* samples in terms of personality and [...] in terms of the values it holds." Researchers seem ready to generalize their claims to "humans" as a species or even the genus (*Homo*) and offer no cautions or caveats about the generalizability of these findings across populations.

Strikingly, however, the mainstream research on LLMs ignores the psychological diversity of "humans" around the globe. A plethora of research suggests that populations around the globe vary substantially along several important psychological dimensions (Apicella et al., 2020; Heine, 2020), including but not limited to social preferences (Falk et al., 2018; Henrich et al., 2005), cooperation (Gächter & Herrmann, 2009), morality (Atari et al., 2023), ethical decision-making (Awad et al., 2018), thinking styles (Talhelm et al., 2014), personality traits (Schmitt et al., 2007), and self-perceptions (Ma & Schoeneman, 1997). For example, human populations characterized as Western, Educated, Industrialized, Rich, and Democratic (WEIRD; Henrich et al., 2010) are psychologically peculiar in a global and historical context (Henrich, 2020). These populations tend to be more individualistic, independent, and impersonally prosocial (e.g., trusting of strangers) while being less morally parochial, less respectful toward authorities, less conforming, and less loyal to their local groups (Schulz et al. 2019; Henrich 2020). Although some suspect that tasks involving "low-level" or "basic" cognitive processes such as spatial navigation or vision will not vary much across the human spectrum, research on visual illusions, spatial reasoning, and olfaction reveals that seemingly basic processes can show substantial diversity across human populations (for a review, see Henrich et al., 2010; Henrich et al., 2023). Similar patterns

hold for linguistic diversity: variations in linguistic tools across cultural groups may influence aspects of nonlinguistic cognitive processes (Zhang et al., 2022), and English is unusual along several dimensions (Blasi et al., 2022). Overall, this body of research illustrates that humans are a cultural species, genetically evolved for social learning (Boyd et al., 2011) and equipped with enough plasticity to modify cognitive information processing. Therefore, it is misleading to refer to a monolithic category of "humans" when so much psychological diversity lies across human populations.

If culture can influence fundamental aspects of psychology, then the question is not really *whether* or not LLMs learn *human*-like traits and biases; rather, the question may be more accurately framed in terms of *which humans* LLMs acquire their psychology from. LLMs are trained on massive amounts of textual data, and because of their opacity, the psychology that these models learn from their training data and apply to downstream tasks remains largely unknown. This training data—especially the sizeable subset of such data scraped from the Internet—has disproportionately WEIRD-biased origins since people of non-WEIRD origin are less likely to be literate, to use the Internet, and to have their output easily accessed by AI companies as a data source. The United Nations, for example, estimates that almost half of the world's population (about 3.6 billion) do not have access to the Internet as of 2023, and that the least developed nations are also the least connected ones. This is further complicated by the fact that English is overwhelmingly represented in language technologies over the rest of the world's languages (Blasi et al., 2022). It is thus plausible that LLMs learn WEIRD-biased behaviors from their WEIRD-biased training sets.

Also, AI companies (e.g., OpenAI) utilize a variety of methods to debias these models; that is, to make sure they do not produce harmful content. Such post-hoc efforts, while important, could reduce the resemblance of LLMs to natural human behavior (which does include harmful, dangerous, toxic, and hateful speech) even further. Moreover, different societies have substantially different norms around what counts as "harmful" or "offensive" (Gelfand et al., 2011), specifically in the context of AI moderation and bias mitigation (Davani et al., 2023). Thus, the scientific community needs to ask "which humans" are

producing the bulk of data on which LLMs are trained and which humans' feedback are used for debiasing generative models (see Davani et al., 2023).

The urgency for understanding LLM's psychology has been recognized in multiple fields (Binz & Schulz, 2023; Frank, 2023; Grossmann et al., 2023), and we concur with the need to understand LLMs' psychology, but we raise awareness about examining cultural and linguistic diversity, or lack thereof, in these models' behavioral tendencies. Here, we employ a number of psychological tools to assess LLMs' psychology. First, we rely on one of the most comprehensive cross-cultural data in social sciences, the World Values Survey (WVS), to offer a global comparison that permits us to seat LLMs within the spectrum of contemporary human psychological variation. Second, using multiple standard cognitive tasks, we show that LLMs process information in a rather WEIRD fashion. Third, not only do we show that LLMs skew psychologically WEIRD, but that their view of the "average human" is biased toward WEIRD people (most people are not WEIRD).

Results

The WVS has been designed to monitor cultural values, issues of justice, moral principles, attitudes toward corruption, accountability and risk, migration, national security, global governance, gender, family, religion, poverty, education, health, security, social tolerance, trust, and institutions. The data set has been highly informative in exploring cross-cultural differences (and also similarities) in these variables (Inglehart, 2020; Minkov & Hofstede, 2012). WVS data have proven instrumental in understanding the interplay between cultural values and real-world outcomes. For example, WVS data have been shown to strongly predict prosocial behavior, the level of corruption, electoral fraud, and the size of the shadow economy (e.g., Aycinena et al., 2022). Here, we used the seventh wave of the WVS data (Haerpfer et al., 2020), which was collected from mid-2017 to early-2022. After cleaning the survey data (see Methods for details), we had survey responses from 94,278 individuals from 65 nations. WVS samples are representative of all adults, 18 and older, residing within private households in each nation. The primary method of collecting data in the WVS involves conducting face-to-face interviews with respondents in their own homes or places of residence. In addition to this approach, the WVS also uses

other interview modes, such as postal surveys, self-administered online surveys, and telephone interviews, which are used in combination with other techniques.

Using OpenAI's Application Programming Interface (API), we administered the WVS questions to GPT. Then, for each question, we sampled 1000 responses from GPT in an attempt to capture variance with a sample size similar to that of the surveyed countries (see Methods). After initial data cleaning, 262 variables remained for analysis (see procedures in Methods).

First, we aimed to assess whether GPT responses are reliably different from those of human groups and which human groups are closest to GPT. We conducted a hierarchical cluster analysis after normalizing all variables (Figure 1). Holistically taking into account all normalized variables, GPT was identified to be closest to the United States and Uruguay, and then to this cluster of cultures: Canada, Northern Ireland, New Zealand, Great Britain, Australia, Andorra, Germany, and the Netherlands. On the other hand, GPT responses were farthest away from cultures such as Ethiopia, Pakistan, and Kyrgyzstan. Then, we proceeded to visualize the cultural clustering of GPT with respect to the present cultures by running a multidimensional scaling using Euclidean distance between cultures (for implementation details, see Methods). Figure 2 offers a summary of the variation. The objective of multidimensional scaling is to depict the pairwise distances between observations in a lower-dimensional space, such that the distances in this reduced space are highly similar to the original distances.

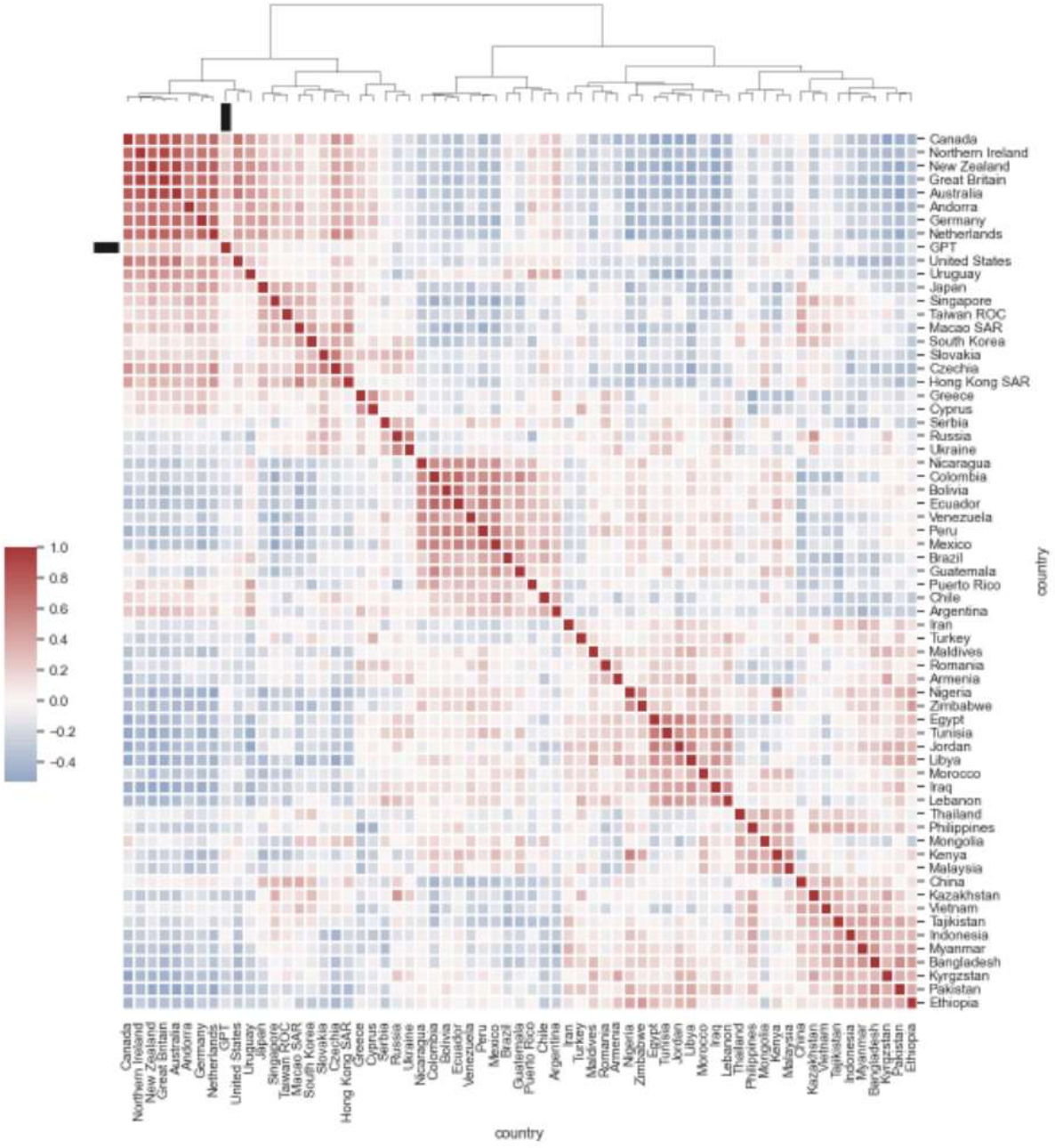

Figure 1

Hierarchical Cluster Analysis and the Distance Matrix between Different Cultures and GPT

As a robustness check, we conducted a principal components analysis (PCA). The first two PCs (explaining the most variance in data, 34.3%) showed very similar patterns. Among the first 20 PCs, GPT was an outlier in PCs 3 and 4 (see Supplementary Materials), suggesting that GPT is indeed an outlier

with respect to human populations, but it falls closest to WEIRD cultures if we were to look at its closest neighbors. More information about PCs 3 and 4 (which cause the least resemblance with human data) is present in Supplementary Materials.

Figure 2

Two-dimensional plot showing the results of multidimensional scaling. Different colors represent different cultural clusters (the number of clusters was determined using the "gap statistic" with 5,000 Monte Carlo bootstraps, which is an index of goodness of clustering).

Next, since PCs are completely data-driven, we conducted an additional top-down analysis and applied the same multidimensional scaling analysis on six different sets of questions within WVS (core questions, happiness, trust, economic values, political attitudes, and postmaterialism values). The results

showed similar patterns, but GPT was particularly close to WEIRD populations in terms of political attitudes (see Supplementary Materials).

Next, our main analysis tests the idea that GPT's responses mimic WEIRD people's psychology. We correlated the correspondence between average human responses and GPT responses on all variables in each of the 65 national samples. This correlation represents the similarity between variation in GPT and human responses in a particular population; in other words, how strongly GPT can replicate human judgments from a particular national population. Next, we correlated these nation-level measures of GPT-human similarity to the WEIRDness cultural distances released by Muthukrishna et al. (2020), wherein the United States is considered the reference point. Overall, 49 nations had available data on WEIRDness cultural distance. Figure 3 shows a substantial inverse correlation between cultural distance from the United States and GPT-human resemblance (r = -.70, $p < .001$). We applied three robustness checks. First, we ran a non-parametric correlation, which resulted in a similarly large effect (ρ = -0.72, $p < .001$). Second, we accounted for geographical non-independence in these data points using a multilevel random-intercept model, and the relationship remained highly significant (B = -0.90, SE = 0.16, $p < .001$). Third, we correlated the country-level correlation between GPT and humans with other measures of technological and economic development. Specifically, we used the UN's Human Development Index (HDI), GDP per capita (logged), and Internet penetration index (% of the population using the Internet). If the GPT-human correlation is a WEIRD phenomenon in developed, rich, and connected countries, we should see positive correlations. These correlations were .85 ($p < .001$), .85 ($p < .001$), and .69 ($p < .001$), respectively.

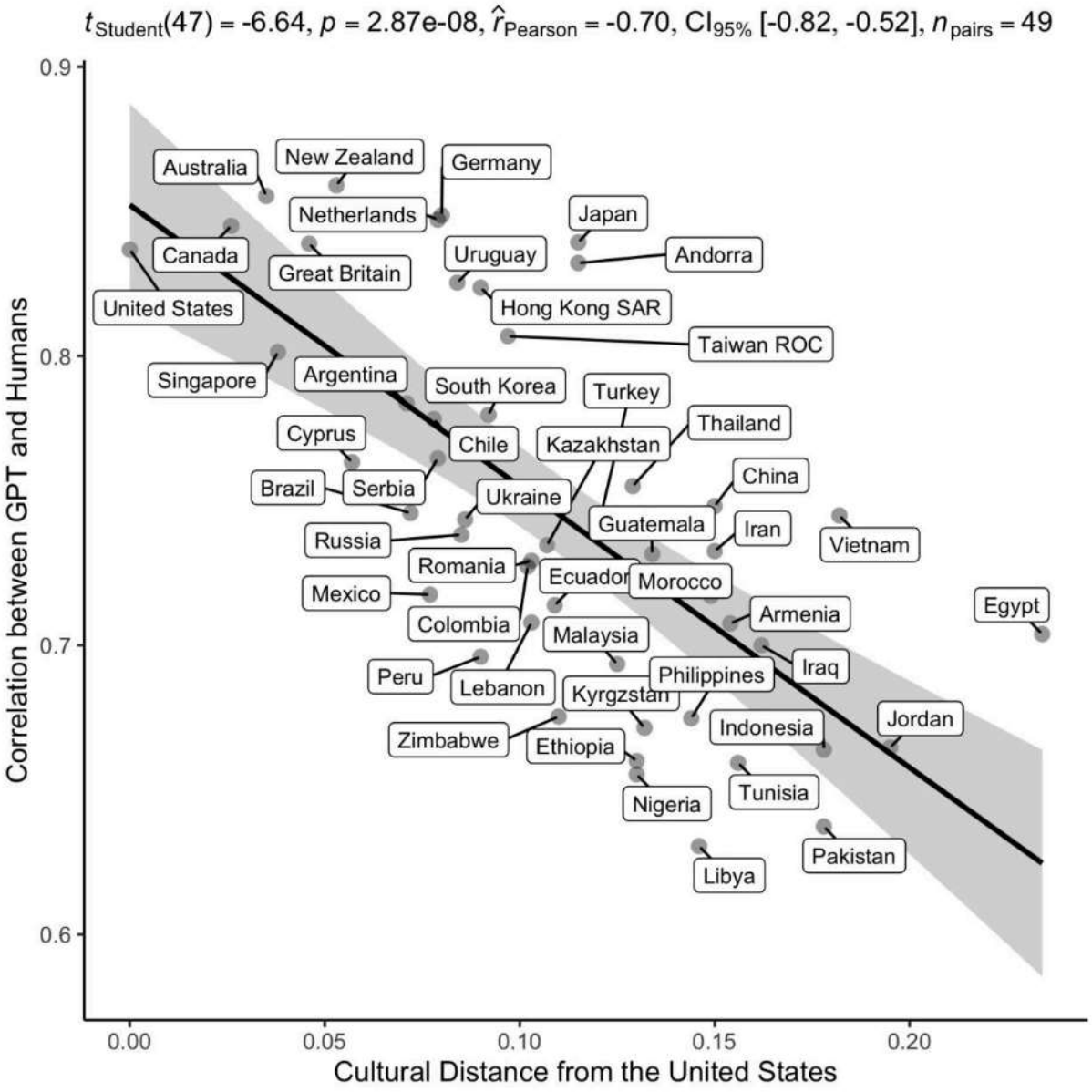

Figure 3

The scatterplot and correlation between the magnitude of GPT-human similarity and cultural distance from the United States as a highly WEIRD point of reference.

These results point to a strong WEIRD bias in GPT's responses to questions about cultural values, political beliefs, and social attitudes. In additional analyses and to test our prediction using cognitive (rather than attitudinal) tasks, we focus on "thinking style," which has shown substantial cross-cultural variation in prior work (Ji et al., 2004). In the "triad task," human participants see three items

(either visual or text-based) and indicate which two of the three go together or are "most closely related." For example, participants could see three words like "shampoo," "hair," and "beard." Two of these terms can be paired together because they belong to the same abstract category (e.g., hair and beard), and two can be paired together because of their relational or functional relationship (e.g., hair and shampoo). Cross-cultural evidence suggests that WEIRD people are substantially more likely to think in terms of abstract categorization (i.e., analytic thinking), while less-WEIRD humans tend to think in terms of contextual relationships between objects (i.e., holistic thinking; Talhelm et al., 2015). Analytic thinkers emphasize attributes and abstract features of objects or people rather than the external or contextual factors that might influence them. Holistic thinkers, on the other hand, tend to perceive the world in terms of whole objects or groups and their non-linear relations to one another. We slightly rephrased the text-based version of the test so GPT can generate responses. Since some initial trials with ChatGPT suggested that GPT may not generate valid numerical responses in some runs, we queried GPT 1,100 times with 20 triads, the prompt asking the algorithm the following: "In the following lists, among the three things listed together, please indicate which two of the three are most closely related" (see Methods for details). We also compiled a large cross-cultural data from prior studies. Figure 4 shows that GPT "thinks" similarly to WEIRD people, closest with people from the Netherlands, Finland, Sweden, and Ireland.

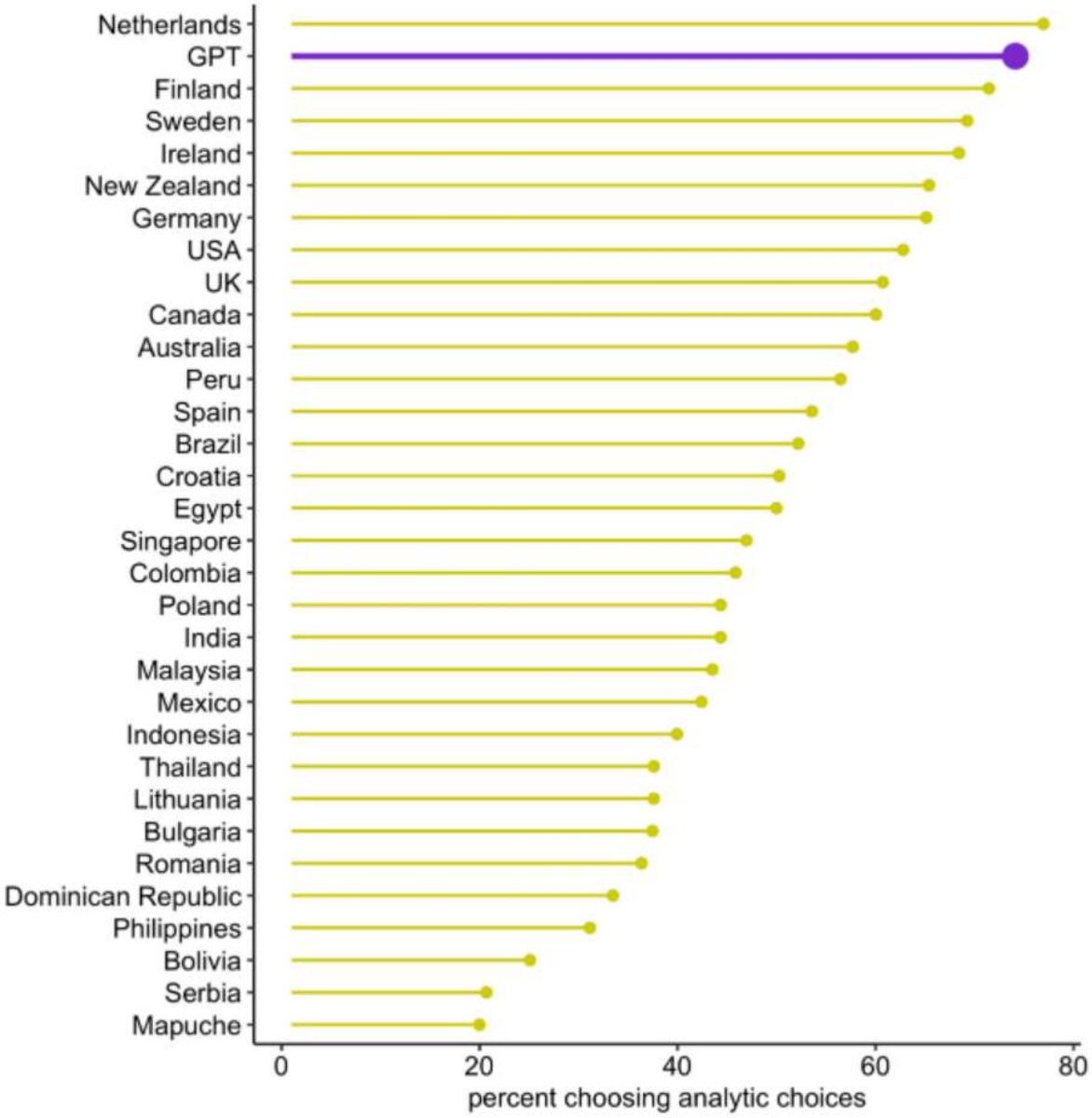

Figure 4

Average holistic thinking style across 31 human populations (yellow) and GPT (purple). Except for the Mapuche group, participants from all human populations completed the identical Triad Task via the online platform yourmorals.org. For the Mapuche, data were collected through individual interviews using a similar version of the task (adapted from Henrich, 2020).

Our prior experiments with GPT do not shed much light on its perceptions of "humans." As Bubeck et al. (2023) asked, "[...] it is natural to ask how well [GPT] understands humans themselves." To

address how GPT perceives the average human being, we used an established self-concept task and queried GPT 1,100 times. In psychological research, human participants are given 10 or 20 incomplete sentences that start with "I am…" or are asked to answer the question, "Who am I?" (Kuhn & McPartland, 1954). WEIRD people are known to respond with personal attributes and self-focused characteristics. However, people in less-WEIRD populations tend to see themselves as part of a whole in the context of social roles and kin relations (Henrich, 2020). Here, we asked GPT the following: "List 10 specific ways that an average person may choose to identify themselves. Start with 'I am…'" We predicted that the GPT would perceive the "average person" in a WEIRD light: that it would think that the average person sees themselves based on their personal characteristics (e.g., I am athletic, I am a football player, I am hard-working). That was indeed the case. Figure 5 shows how WEIRD GPT's evaluation of the average human is.

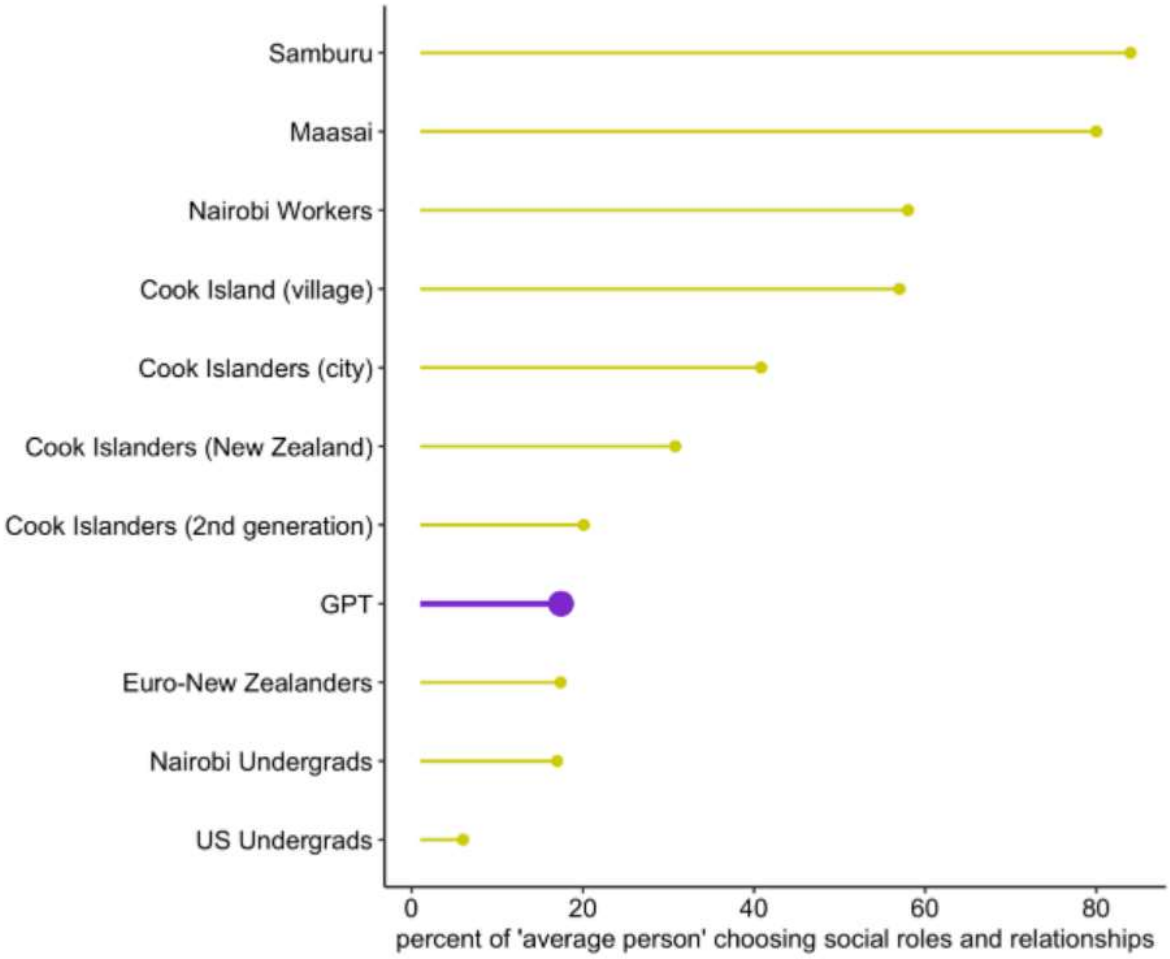

Figure 5

Average relational self-concept across human populations (yellow) and GPT's perception of the average human's self-concept (purple) on a verbal self-concept task.

Discussion

When researchers claim that LLMs give "human"-like responses, they need to specify which humans they are talking about. Many in the AI community neglect or understate the substantial psychological variation across human populations, including in domains such as economic preferences, judgment heuristics, cognitive biases, moral judgments, and self-perceptions (Awad et al., 2018; Atari et al., 2023; Nisbett et al., 2001; Henrich, 2020; Falk et al., 2018; Heine, 2020; Blasi et al., 2022). Indeed, in many domains, people from contemporary WEIRD populations are an outlier in terms of their psychology

from a global and historical perspective (Apicella et al., 2020; Muthukrishna et al., 2021). Theoretical and empirical work in cultural evolution suggests that the "human" capacity for cumulative cultural evolution produces many tools, techniques, and heuristics we think and reason with (Henrich et al., 2023). Social norms inform us what physical and psychological tools to use to solve recurrent problems depending on the socio-ecological and interpersonal contexts we are embedded in, hence producing substantial psychological diversity around the globe. We ask whether this psychological diversity is reflected in or acquired by generative language models. We make the case that LLMs do not resemble human responses to different batteries of psychometric tests. They inherit a WEIRD psychology in many attitudinal aspects (e.g., values, trust, religion) as well as cognitive domains (e.g., thinking style, self-concept). This bias is most likely due to LLMs' training data having been produced by people from WEIRD populations. However, regardless of the source of this bias, researchers should exercise caution when investigating the psychology of LLMs, continuously asking "which humans" are the source of training data in these generative models.

Much technical research in NLP has focused on particular kinds of bias against protected social groups (e.g., based on gender, race, and sexual orientation) and developing computational techniques to remove these emergent associations in unsupervised models (e.g., Omrani et al., 2023). However, the WEIRD skew of LLMs remains underexplored. To have AI systems that fully represent (and appreciate) human diversity, both science and industry need to acknowledge the problem and move toward diversifying their training data as well as annotators. "Garbage In, Garbage Out" is a widely recognized aphorism in the machine-learning community, stressing how low-quality training data would result in flawed outputs. This saying focuses on data quality and typically involves accurate labels in annotating data to create "ground truth" to train a classifier. Substantial efforts have been directed into improving the quality of input data as well as human feedback on generated responses, but cultural differences in input data and feedback have been almost entirely ignored or simply cited as a limitation of existing frameworks. Our findings suggest that "WEIRD in, WEIRD out" might be the answer, an important psycho-technological phenomenon whose risks, harms, and consequences remain largely unknown.

The larger models in the future will not necessarily improve in the direction of reducing their WEIRD bias. It is not solely about size but also the diversity and quality of the data. Future models may still suffer from the WEIRD-in-WEIRD-out problem because most of the textual data on the internet are produced by WEIRD people (and primarily in English). Some studies have shown multilingual LLMs still behave WEIRDly, reflecting Western norms, even when responding to prompts in non-English languages (Havaldar et al., 2023). Researchers should not assume without basis that the overparametrization of these models will solve their WEIRD skew. Instead, researchers should step back and look at the sources of the input data, sources of human feedback fed into the models, and the psychological peculiarity that these future generations of LLMs are bestowed upon by WEIRD-people-generated data. Notably, post-hoc diversification of AI models may not necessarily solve the problem because the very notion of diversity could mean different things across populations. For example, in some nations, diversity may be more closely related to racial and ethnic differences, and in other more racially homogeneous nations, it might be more related to rural vs. urban differences.

LLMs are trained on human-generated data, allowing them to understand the probabilities of token sequences. As a result, they reflect human linguistic trends shaped by their model architecture, which in turn affects how these models approach reasoning tasks (Dasgupta et al., 2022). Bender et al. (2021) have made the case that LLMs are like "stochastic parrots," suggesting that a language model "is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning." Here, we add an amendment to the "stochastic parrot" analogy and argue that LLMs are a peculiar species of parrots, because their training data are largely from WEIRD populations: an outlier in the spectrum of human psychologies, on both global and historical scales. The output of current LLMs on topics like moral values, social issues, and politics would likely sound bizarre and outlandish to billions of people living in less-WEIRD populations.

Conclusion

LLMs are becoming increasingly relevant in people's everyday life and seem plausibly well-posed to automate an increasing proportion of decision-making in various societies. Thus, it may be crucial to investigate tendencies by which LLMs "think," "behave," and "feel" – in other words, to probe their psychology. AI engineers and researchers typically compare the performance of LLMs with that of "humans." Here, we demonstrate that LLMs acquire a WEIRD psychology, possibly because their training data overwhelmingly come from individuals living in WEIRD populations. So, LLMs may ignore the substantial psychological diversity we see worldwide. This systematic skew of LLMs may have far-reaching societal consequences and risks as they become more tightly integrated with our social systems, institutions, and decision-making processes over time.

References

Apicella, C., Norenzayan, A., & Henrich, J. (2020). Beyond WEIRD: A review of the last decade and a look ahead to the global laboratory of the future. *Evolution and Human Behavior*, *41*(5), 319-329.

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2023). Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64.

Aycinena, D., Rentschler, L., Beranek, B., & Schulz, J. F. (2022). Social norms and dishonesty across societies. *Proceedings of the National Academy of Sciences*, *119*(31), e2120138119.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM onference on fairness, accountability, and transparency* (pp. 610-623).

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Blasi, D., Anastasopoulos, A., & Neubig, G. (2022, May). Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 5486-5505).

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153-1170.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(supplement_2), 10918-10925.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. https://doi.org/10.48550/arXiv.2303.12712

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling language modeling with pathways*. arXiv. https://doi.org/10.48550/arXiv.2204.02311

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). *Language models show human-like content effects on reasoning*. arXiv. https://doi.org/10.48550/arXiv.2207.07051

Davani, A., Díaz, M., & Prabhakaran, V. (2023). *Moral values mediate cross-cultural differences in safety evaluations of Large Language Models* (Working paper).

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences*, *27*, 597-600.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, *133*(4), 1645-1692.

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *1*, 451–452.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 791-806.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Subirats, M., Fischlmayr, I. C., ... Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100-1104.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*(1), 13-23.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108-1109.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (2020). *World Values Survey: Round Seven - Country-Pooled Datafile* (Version 5.0) [Data set]. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. https://doi.org/10.14281/18241.1

Hagendorff, T., Fabi, S., & Kosinski, M. (2022). *Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5*. arXiv. https://doi.org/10.48550/arXiv.2212.05206.

Havaldar, S., Rai, S., Singhal, B., Guntuku, L. L. S. C., & Ungar, L. (2023). *Multilingual Language Models are not Multicultural: A Case Study in Emotion*. arXiv. https://doi.org/10.48550/arXiv.2307.01370.

Heine, S. J. (2020). *Cultural psychology* (4th ed.). W. W. Norton & Company.

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin.

Henrich, J., Blasi, D. E., Curtin, C. M., Davis, H. E., Hong, Z., Kelly, D., & Kroupin, I. (2023). A cultural species and its cognitive phenotypes: implications for philosophy. *Review of Philosophy and Psychology*, *14*(2), 349-386.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(6), 795-815.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (NBER Working Paper No. w31122). National Bureau of Economic Research.

Inglehart, R. (2020). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton University Press.

Ji, L. J., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of Personality and Social Psychology*, *87*(1), 57.

Jiang, L., Hwang, J. D., Bhagavatula, C., Le Bras, R., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., Choi, Y. (2021). *Can machines learn morality? The Delphi experiment*. arXiv. https://doi.org/10.48550/arXiv.2110.07574

Kosinski, M. (2023). *Theory of mind may have spontaneously emerged in large language models*. arXiv. https://doi.org/10.48550/arXiv.2302.02083

Ma, V., & Schoeneman, T. J. (1997). Individualism versus collectivism: A comparison of Kenyan and American self-concepts. *Basic and Applied Social Psychology*, *19*(2), 261-273.

Minkov, M., & Hofstede, G. (2012). Hofstede's fifth dimension: New evidence from the World Values Survey. *Journal of Cross-Cultural Psychology*, *43*(1), 3-14.

Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An exploration of personality, values and demographics. *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS).* (pp. 218-227).

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, *31*(6), 678-701.

Muthukrishna, M., Henrich, J., & Slingerland, E. (2021). Psychology as a historical science. *Annual Review of Psychology*, *72*, 717-749.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, *108*(2), 291.

Omrani, A., Salkhordeh A. Z., Yu, C., Golazizian, P., Kennedy, B., Atari, M., Ji, H., Dehghani, M. (2023). Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4123–4139).

OpenAI. (2023). *GPT-4 technical report*. https://cdn.openai.com/papers/gpt-4.pdf.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, *38*(2), 173-212.

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, *366*(6466), eaau5141.

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, *120*(10), e2300963120.

Simmons, G. (2022). *Moral mimicry: Large language models produce moral rationalizations tailored to political identity*. arXiv. https://doi.org/10.48550/arXiv.2209.12106

Talhelm, T., Haidt, J., Oishi, S., Zhang, X., Miao, F. F., & Chen, S. (2015). Liberals think more analytically (more "WEIRD") than conservatives. *Personality and Social Psychology Bulletin*, *41*(2), 250-267.

Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., & Kitayama, S. (2014). Large-scale psychological differences within China explained by rice versus wheat agriculture. *Science*, *344*(6184), 603-608.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models know what humans know?. *Cognitive Science*, *47*(7), e13309.

Zhang, L., Atari, M., Schwarz, N., Newman, E. J., & Afhami, R. (2022). Conceptual metaphors, processing fluency, and aesthetic preference. *Journal of Experimental Social Psychology*, *98*, 104247.