

Quels humains ?

Traduction de l'article *Which Humans ?*, Mohammad Atari, Mona J. Xue, Peter S. Park, Damian E. Blasi, Joseph Henrich, Harvard University¹

Quels humains ?

Mohammad Atari, Mona J. Xue, Peter S. Park, Damian E. Blasi, Joseph Henrich
Département de biologie évolutive humaine, Université Harvard

Notes sur les auteurs

Nous remercions Thomas Talhelm pour le partage de données, Ali Akbari pour ses retours sur les analyses, et Frank Kassanits pour avoir suggéré d'explorer ce sujet. Ce travail a été partiellement financé par la John Templeton Foundation (#62161).

Auteur pour la correspondance : Mohammad Atari, Département de psychologie et des sciences du cerveau, Université du Massachusetts Amherst, 135 Hicks Way, Amherst, MA 01003.

Résumé

Les grands modèles de langage (llm) ont récemment réalisé de vastes progrès à la fois dans la génération et l'analyse de données textuelles. Les rapports techniques comparent souvent les sorties des llm avec la performance "humaine" sur divers tests. Ici, nous demandons : "Quels humains ?" Une grande partie de la littérature existante ignore largement le fait que les humains sont une espèce culturelle avec une diversité psychologique substantielle à travers le globe qui n'est pas pleinement capturée par les données textuelles sur lesquelles les llm actuels ont été entraînés. Nous montrons que les réponses des llm aux mesures psychologiques sont une valeur aberrante par rapport aux données interculturelles à grande échelle, et que leur performance sur les tâches de psychologie cognitive ressemble le plus à celle des personnes des sociétés occidentales, éduquées, industrialisées, riches et démocratiques (weird), mais décline rapidement à mesure que l'on s'éloigne de ces populations ($r = -0,70$). Ignorer la diversité interculturelle dans la psychologie humaine et machine soulève de nombreux problèmes scientifiques et éthiques. Nous concluons en discutant des moyens d'atténuer le biais weird dans les futures générations de modèles de langage génératifs.

Mots-clés : Culture, psychologie humaine, psychologie machine, intelligence artificielle, grands modèles de langage.

0.1. Introduction

L'idée que les systèmes d'ia puissent posséder des traits semblables à ceux des humains n'est pas une observation nouvelle. Étant donné le rôle sociétal croissant joué par les grands modèles de langage (llm), les chercheurs ont commencé à étudier la psychologie sous-jacente de ces modèles génératifs. Par exemple, plusieurs travaux ont examiné si les llm peuvent réellement comprendre le langage et effectuer un raisonnement [11], comprendre les distinctions entre différentes moralités et personnalités [36, 45], et apprendre des dilemmes éthiques [31]. Hagendorff et al. (2022), par

1. Référence : <https://evolvinglanguage.ch/wp-content/uploads/Which-Humans-Atari-et-al.pdf>.

Denise Vella-Chemla et deepseek, octobre 2026.

exemple, ont démontré que les llm sont des décideurs intuitifs, tout comme les humains, arguant que l’investigation des llm avec des méthodes de la psychologie a le potentiel de découvrir leurs traits et comportements émergents [22]. Miotto et al. (2022) ont constaté que le Generative Pre-trained Transformer-3 (chatgpt-3) contient une “personnalité moyenne” par rapport aux échantillons “humains”, a des valeurs auxquelles il attribue divers degrés d’importance, et tombe dans un groupe démographique de jeunes adultes relativement jeunes [36]. Horton (2023) a soutenu que les llm sont des modèles computationnels implicites des humains - un *Homo silicus*. Il suggère ensuite que ces modèles peuvent être utilisés de la même manière que les économistes utilisent l’*Homo economicus* : les llm peuvent recevoir des informations et des préférences, puis leur comportement peut être exploré via des simulations [29]. Les llm se sont également avérés capables d’attribuer des croyances à d’autres, une capacité connue sous le nom de Théorie de l’esprit [48, 32]. Enfin, Bubeck et al. (2023) ont récemment avancé qu’il y a des étincelles d’intelligence générale (c’est-à-dire une capacité mentale générale incluant la capacité de raisonner de manière cohérente, de comprendre des idées complexes, de planifier l’avenir, de résoudre des problèmes, de penser de manière abstraite, d’apprendre rapidement et d’apprendre de l’expérience; gottfredson1997) dans chatgpt-4 [41] qui a été entraîné en utilisant une échelle sans précédent de puissance de calcul et de données. Certains scientifiques sociaux sont allés encore plus loin, plaidant pour les llm comme remplaçants potentiels des participants humains dans la recherche psychologique [14, 20].

Dans la littérature croissante sur le sondage de la psychologie des llm (voir shiffrin2023), les chercheurs ont répété que ces systèmes répondent de manières cognitivement et attitudinalement similaires aux “humains”. Par exemple, le rapport technique chatgpt-4 [41] présente chatgpt-4 (la dernière version du llm qui alimente le populaire chatbot d’openai, chatgpt) comme “un modèle multimodal avec une performance de niveau humain sur certaines tâches professionnelles et académiques difficiles”. Bubeck et al. (2023) mentionnent que “la performance de chatgpt-4 est remarquablement proche de la performance de niveau humain” et que “chatgpt-4 atteint une performance de niveau humain sur de nombreuses tâches [...] il est naturel de se demander comment chatgpt-4 comprend les humains eux-mêmes” [10]. Les chercheurs en sciences sociales (par exemple, psychologie, économie) ont utilisé la même terminologie pour comparer les llm et les “humains”. Par exemple, pour démontrer les capacités de prise de décision économique des llm, Horton (2023) soutient qu’ils “peuvent être considérés comme des modèles computationnels implicites des humains” [29]. En quantifiant la personnalité et les valeurs morales des llm, Miotto et al. (2022) soutiennent que chatgpt-3 “score de manière similaire aux échantillons humains en termes de personnalité et [...] en termes des valeurs qu’il détient” [36]. Les chercheurs semblent prêts à généraliser leurs affirmations aux “humains” en tant qu’espèce ou même genre (*Homo*) et n’offrent aucune mise en garde ou réserve sur la généralisabilité de ces résultats à travers les populations.

De manière frappante, cependant, la recherche dominante sur les llm ignore la diversité psychologique des “humains” à travers le globe. Une pléthore de recherches suggère que les populations à travers le globe varient substantiellement le long de plusieurs dimensions psychologiques importantes [1, 24], y compris, mais sans s’y limiter, les préférences sociales [15, 27], la coopération [17], la moralité [2], la prise de décision éthique [3], les styles de pensée [47], les traits de personnalité [42] et les perceptions de soi [34]. Par exemple, les populations humaines caractérisées comme occidentales, éduquées, industrialisées, riches et démocratiques (weird; henrich2010) sont psychologiquement particulières dans un contexte global et historique [25]. Ces populations ont tendance à être plus individualistes, indépendantes et prosociales de manière impersonnelle (par exemple, faisant confiance aux étrangers) tout en étant moins morales de manière paroissiale, moins respec-

tueuses de l'autorité, moins conformistes et moins loyales à leurs groupes locaux [43, 25]. Bien que certains soupçonnent que les tâches impliquant des processus cognitifs “de bas niveau” ou “basiques” tels que la navigation spatiale ou la vision ne varieront pas beaucoup à travers le spectre humain, la recherche sur les illusions visuelles, le raisonnement spatial et l'olfaction révèle que des processus apparemment basiques peuvent montrer une diversité substantielle à travers les populations humaines (pour une revue, voir henrich2010, henrich2023). Des schémas similaires existent pour la diversité linguistique : les variations dans les outils linguistiques à travers les groupes culturels peuvent influencer des aspects des processus cognitifs non linguistiques [49], et l'anglais est inhabituel le long de plusieurs dimensions [7]. Globalement, ce corpus de recherches illustre que les humains sont une espèce culturelle, génétiquement évoluée pour l'apprentissage social [9] et équipée d'assez de plasticité pour modifier le traitement de l'information cognitive. Par conséquent, il est trompeur de se référer à une catégorie monolithique d’“humains” alors que tant de diversité psychologique existe à travers les populations humaines.

Si la culture peut influencer les aspects fondamentaux de la psychologie, alors la question n'est pas vraiment de savoir si les llm apprennent des traits et des biais semblables à ceux des humains ; plutôt, la question peut être plus précisément formulée en termes de *quels* humains les llm acquièrent leur psychologie. Les llm sont entraînés sur des masses de données textuelles, et en raison de leur opacité, la psychologie que ces modèles apprennent de leurs données d'entraînement et appliquent à des tâches en aval reste largement inconnue. Ces données d'entraînement - en particulier le sous-ensemble important de ces données extraites d'Internet - ont des origines disproportionnellement biaisées weird puisque les personnes d'origine non weird sont moins susceptibles d'être alphabétisées, d'utiliser Internet et d'avoir leurs productions facilement accessibles par les entreprises d'ia comme source de données. Les Nations Unies, par exemple, estiment que près de la moitié de la population mondiale (environ 3,6 milliards) n'a pas accès à Internet en 2023, et que les nations les moins développées sont aussi les moins connectées. Ceci est encore compliqué par le fait que l'anglais est massivement surreprésenté dans les technologies linguistiques par rapport au reste des langues du monde [7]. Il est donc plausible que les llm apprennent des comportements biaisés weird à partir de leurs ensembles d'entraînement biaisés weird.

De plus, les entreprises d'ia (par exemple, openai) utilisent une variété de méthodes pour débiaiser ces modèles ; c'est-à-dire, pour s'assurer qu'ils ne produisent pas de contenu nuisible. Ces efforts post-hoc, bien qu'importants, pourraient réduire encore davantage la ressemblance des llm avec le comportement humain naturel (qui inclut des discours nuisibles, dangereux, toxiques et haineux). De plus, différentes sociétés ont des normes substantiellement différentes sur ce qui compte comme “nuisible” ou “offensant” [18], spécifiquement dans le contexte de la modération de l'ia et de l'atténuation des biais [13]. Ainsi, la communauté scientifique doit demander “quels humains” produisent la majeure partie des données sur lesquelles les llm sont entraînés et quels humains' feedback est utilisé pour débiaiser les modèles génératifs (voir davani2023).

L'urgence de comprendre la psychologie des llm a été reconnue dans de multiples domaines [6, 16, 20], et nous souscrivons au besoin de comprendre la psychologie des llm, mais nous sensibilisons à l'examen de la diversité culturelle et linguistique, ou de son absence, dans les tendances comportementales de ces modèles. Ici, nous employons un certain nombre d'outils psychologiques pour évaluer la psychologie des llm. Premièrement, nous nous appuyons sur l'une des données inter-culturelles les plus complètes en sciences sociales, l'Enquête mondiale sur les valeurs (wvs), pour offrir une comparaison mondiale qui nous permet de situer les llm dans le spectre de la variation psychologique humaine contemporaine. Deuxièmement, en utilisant de multiples tâches cognitives

standard, nous montrons que les llm traitent l’information d’une manière plutôt weird. Troisièmement, non seulement nous montrons que les llm penchent psychologiquement vers weird, mais que leur vision de l’“humain moyen” est biaisée vers les personnes weird (la plupart des gens ne sont pas weird).

0.2. Résultats

La wvs a été conçue pour surveiller les valeurs culturelles, les questions de justice, les principes moraux, les attitudes envers la corruption, la responsabilité et le risque, la migration, la sécurité nationale, la gouvernance mondiale, le genre, la famille, la religion, la pauvreté, l’éducation, la santé, la sécurité, la tolérance sociale, la confiance et les institutions. L’ensemble de données a été très informatif pour explorer les différences interculturelles (et aussi les similitudes) dans ces variables [30, 35]. Les données de la wvs se sont avérées instrumentales pour comprendre l’interaction entre les valeurs culturelles et les résultats du monde réel. Par exemple, les données de la wvs se sont avérées fortement prédictives du comportement prosocial, du niveau de corruption, de la fraude électorale et de la taille de l’économie souterraine (par exemple, aycinena2022). Ici, nous avons utilisé la septième vague des données de la wvs [21], qui a été collectée de mi-2017 à début 2022. Après nettoyage des données d’enquête (voir Méthodes pour les détails), nous avons des réponses d’enquête de 94 278 individus de 65 nations. Les échantillons de la wvs sont représentatifs de tous les adultes, 18 ans et plus, résidant dans des ménages privés dans chaque nation. La méthode principale de collecte de données dans la wvs implique la conduite d’entretiens en face-à-face avec les répondants dans leurs propres maisons ou lieux de résidence. En plus de cette approche, la wvs utilise également d’autres modes d’entretien, tels que des enquêtes postales, des enquêtes en ligne auto-administrées et des entretiens téléphoniques, qui sont utilisés en combinaison avec d’autres techniques.

En utilisant l’api d’openai, nous avons administré les questions de la wvs à chatgpt. Ensuite, pour chaque question, nous avons échantillonné 1000 réponses de chatgpt dans une tentative de capturer la variance avec une taille d’échantillon similaire à celle des pays enquêtés (voir Méthodes). Après nettoyage initial des données, 262 variables sont restées pour l’analyse (voir procédures dans Méthodes).

Premièrement, nous avons cherché à évaluer si les réponses de chatgpt sont fidèlement différentes de celles des groupes humains et quels groupes humains sont les plus proches de chatgpt. Nous avons mené une analyse de clustering hiérarchique après normalisation de toutes les variables (Figure 1). Holistiquement, en tenant compte de toutes les variables normalisées, chatgpt a été identifié comme le plus proche des États-Unis et de l’Uruguay, puis de ce cluster de cultures : Canada, Irlande du Nord, Nouvelle-Zélande, Grande-Bretagne, Australie, Andorre, Allemagne et Pays-Bas. D’un autre côté, les réponses de chatgpt étaient les plus éloignées de cultures telles que l’Éthiopie, le Pakistan et le Kirghizistan. Ensuite, nous avons procédé à la visualisation du clustering culturel de chatgpt par rapport aux cultures présentes en exécutant un positionnement multidimensionnel utilisant la distance euclidienne entre les cultures (pour les détails d’implémentation, voir Méthodes). La figure 2 offre un résumé de la variation. L’objectif du positionnement multidimensionnel est de représenter les distances par paires entre les observations dans un espace de dimension inférieure, de sorte que les distances dans cet espace réduit soient très similaires aux distances originales.

Comme contrôle de robustesse, nous avons mené une analyse en composantes principales (ACP). Les deux premières CP (expliquant le plus de variance dans les données, 34,3 %) ont montré des

schémas très similaires. Parmi les 20 premières CP, chatgpt était une valeur aberrante dans les CP 3 et 4 (voir Matériaux supplémentaires), suggérant que chatgpt est en effet une valeur aberrante par rapport aux populations humaines, mais il tombe le plus près des cultures weird si l'on regarde ses voisins les plus proches. Plus d'informations sur les CP 3 et 4 (qui causent la moindre ressemblance avec les données humaines) sont présentes dans les Matériaux supplémentaires.

Ensuite, notre analyse principale teste l'idée que les réponses de chatgpt imitent la psychologie des personnes weird. Nous avons corrélé la correspondance entre les réponses humaines moyennes et les réponses de chatgpt sur toutes les variables dans chacun des 65 échantillons nationaux. Cette corrélation représente la similarité entre la variation des réponses de chatgpt et des réponses humaines dans une population particulière ; en d'autres termes, à quel point chatgpt peut répliquer les jugements humains d'une population nationale particulière. Ensuite, nous avons corrélé ces mesures au niveau national de similarité chatgpt-humain aux distances culturelles weird publiées par Muthukrishna et al. (2020), où les États-Unis sont considérés comme le point de référence [37]. Globalement, 49 nations avaient des données disponibles sur la distance culturelle weird. La figure 3 montre une corrélation inverse substantielle entre la distance culturelle par rapport aux États-Unis et la ressemblance chatgpt-humain ($r = -0,70, p < 0,001$). Nous avons appliqué trois contrôles de robustesse. Premièrement, nous avons exécuté une corrélation non paramétrique, qui a abouti à un effet tout aussi important ($p = -0,72, p < 0,001$). Deuxièmement, nous avons tenu compte de la non-indépendance géographique dans ces points de données en utilisant un modèle multiniveau à intercept aléatoire, et la relation est restée hautement significative ($B = -0,90, SE = 0,16, p < 0,001$). Troisièmement, nous avons corrélé la corrélation au niveau national entre chatgpt et les humains avec d'autres mesures de développement technologique et économique. Spécifiquement, nous avons utilisé l'Indice de développement humain (IDH) des Nations Unies, le PIB par habitant (log), et l'indice de pénétration d'Internet (% de la population utilisant Internet). Si la corrélation chatgpt-humain est un phénomène weird dans les pays développés, riches et connectés, nous devrions voir des corrélations positives. Ces corrélations étaient de 0,85 ($p < 0,001$), 0,85 ($p < 0,001$) et 0,69 ($p < 0,001$), respectivement.

Ensuite, nous avons examiné si chatgpt "pense" comme les personnes weird en utilisant la tâche des triades, une tâche cognitive standard qui a été largement utilisée dans la recherche interculturelle. Dans la tâche des triades, les participants voient trois objets (visuels ou textuels) et indiquent lesquels deux des trois vont ensemble ou sont "les plus étroitement liés". Par exemple, les participants pouvaient voir trois mots comme "shampooing", "cheveux" et "barbe". Deux de ces termes peuvent être associés parce qu'ils appartiennent à la même catégorie abstraite (par exemple, cheveux et barbe), et deux peuvent être associés en raison de leur relation relationnelle ou fonctionnelle (par exemple, cheveux et shampooing). Les preuves interculturelles suggèrent que les personnes weird sont substantiellement plus susceptibles de penser en termes de catégorisation abstraite (c'est-à-dire la pensée analytique), tandis que les humains moins weird ont tendance à penser en termes de relations contextuelles entre les objets (c'est-à-dire la pensée holistique ; talhelm2015). Les penseurs analytiques mettent l'accent sur les attributs et les caractéristiques abstraites des objets ou des personnes plutôt que sur les facteurs externes ou contextuels qui pourraient les influencer. Les penseurs holistiques, en revanche, ont tendance à percevoir le monde en termes d'objets ou de groupes entiers et de leurs relations non linéaires les uns aux autres. Nous avons légèrement reformulé la version textuelle du test pour que chatgpt puisse générer des réponses. Comme certains essais initiaux avec chatgpt suggéraient que chatgpt pourrait ne pas générer de réponses numériques valides dans certaines exécutions, nous avons interrogé chatgpt 1 100 fois avec 20 triades, l'invite demandant à

l’algorithme ce qui suit : “Dans les listes suivantes, parmi les trois choses listées ensemble, veuillez indiquer lesquelles deux des trois sont les plus étroitement liées” (voir Méthodes pour les détails). Nous avons également compilé de grandes données interculturelles à partir d’études antérieures. La figure 4 montre que chatgpt “pense” de manière similaire aux personnes weird, le plus proche des personnes des Pays-Bas, de Finlande, de Suède et d’Irlande.

Nos expériences antérieures avec chatgpt n’éclairent pas beaucoup ses perceptions des “humains”. Comme l’ont demandé Bubeck et al. (2023), “[...] il est naturel de se demander comment [chatgpt] comprend les humains eux-mêmes” [10]. Pour répondre à la façon dont chatgpt perçoit l’être humain moyen, nous avons utilisé une tâche établie de concept de soi et interrogé chatgpt 1 100 fois. Dans la recherche psychologique, les participants humains reçoivent 10 ou 20 phrases incomplètes qui commencent par “Je suis...” ou on leur demande de répondre à la question “Qui suis-je ?” [33]. Les personnes weird sont connues pour répondre avec des attributs personnels et des caractéristiques centrées sur soi. Cependant, les personnes des populations moins weird ont tendance à se voir comme faisant partie d’un tout dans le contexte des rôles sociaux et des relations de parenté [25]. Ici, nous avons demandé à chatgpt ce qui suit : “Listez 10 manières spécifiques dont une personne moyenne pourrait choisir de s’identifier. Commencez par “Je suis...”” Nous avons prédit que chatgpt percevrait la “personne moyenne” sous un jour weird : qu’il penserait que la personne moyenne se voit en fonction de ses caractéristiques personnelles (par exemple, je suis sportif, je suis un joueur de football, je suis travailleur). Ce fut effectivement le cas. La figure 5 montre à quel point l’évaluation de l’humain moyen par chatgpt est weird.

0.3. Discussion

Lorsque les chercheurs affirment que les llm donnent des réponses “humaines”, ils doivent spécifier de quels humains ils parlent. Beaucoup dans la communauté de l’ia négligent ou sous-estiment la variation psychologique substantielle à travers les populations humaines, y compris dans des domaines tels que les préférences économiques, les heuristiques de jugement, les biais cognitifs, les jugements moraux et les perceptions de soi [3, 2, 39, 25, 15, 24, 7]. En effet, dans de nombreux domaines, les personnes des populations weird contemporaines sont une valeur aberrante en termes de leur psychologie d’un point de vue global et historique [1, 38]. Les travaux théoriques et empiriques en évolution culturelle suggèrent que la capacité “humaine” à l’évolution culturelle cumulative produit de nombreux outils, techniques et heuristiques avec lesquels nous pensons et raisonnons [26]. Les normes sociales nous informent sur les outils physiques et psychologiques à utiliser pour résoudre des problèmes récurrents en fonction des contextes socio-écologiques et interpersonnels dans lesquels nous sommes intégrés, produisant ainsi une diversité psychologique substantielle à travers le globe. Nous demandons si cette diversité psychologique est reflétée dans ou acquise par les modèles de langage génératifs. Nous soutenons que les llm ne ressemblent pas aux réponses humaines à différentes batteries de tests psychométriques. Ils héritent d’une psychologie weird dans de nombreux aspects attitudinaux (par exemple, valeurs, confiance, religion) ainsi que dans des domaines cognitifs (par exemple, style de pensée, concept de soi). Ce biais est très probablement dû au fait que les données d’entraînement des llm ont été produites par des personnes de populations weird. Cependant, quelle que soit la source de ce biais, les chercheurs devraient faire preuve de prudence lorsqu’ils enquêtent sur la psychologie des llm, demandant continuellement “quels humains” sont la source des données d’entraînement dans ces modèles génératifs.

Une grande partie de la recherche technique en NLP s’est concentrée sur des types particuliers de

biais contre des groupes sociaux protégés (par exemple, basés sur le genre, la race et l’orientation sexuelle) et sur le développement de techniques computationnelles pour supprimer ces associations émergentes dans les modèles non supervisés (par exemple, [omrani2023](#)). Cependant, le biais weird des llm reste sous-exploré. Pour avoir des systèmes d’ia qui représentent pleinement (et apprécient) la diversité humaine, la science et l’industrie doivent reconnaître le problème et aller vers la diversification de leurs données d’entraînement ainsi que des annotateurs. “Garbage In, Garbage Out” est un aphorisme largement reconnu dans la communauté de l’apprentissage automatique, soulignant comment des données d’entraînement de faible qualité aboutiraient à des sorties défectueuses. Ce dicton se concentre sur la qualité des données et implique généralement des étiquettes précises dans l’annotation des données pour créer une “vérité terrain” pour entraîner un classifieur. Des efforts substantiels ont été dirigés vers l’amélioration de la qualité des données d’entrée ainsi que du feedback humain sur les réponses générées, mais les différences culturelles dans les données d’entrée et le feedback ont été presque entièrement ignorées ou simplement citées comme une limitation des cadres existants. Nos résultats suggèrent que “weird in, weird out” pourrait être la réponse, un phénomène psycho-technologique important dont les risques, les préjudices et les conséquences restent largement inconnus.

Les modèles plus grands à l’avenir ne s’amélioreront pas nécessairement dans la direction de la réduction de leur biais weird. Il ne s’agit pas uniquement de la taille mais aussi de la diversité et de la qualité des données. Les futurs modèles pourraient encore souffrir du problème “weird in, weird out” parce que la plupart des données textuelles sur Internet sont produites par des personnes weird (et principalement en anglais). Certaines études ont montré que les llm multilingues se comportent encore de manière weird, reflétant les normes occidentales, même lorsqu’ils répondent à des invites dans des langues non anglaises [\[23\]](#). Les chercheurs ne devraient pas supposer sans base que la surparamétrisation de ces modèles résoudra leur biais weird. Au lieu de cela, les chercheurs devraient prendre du recul et examiner les sources des données d’entrée, les sources de feedback humain injectées dans les modèles, et la particularité psychologique dont ces futures générations de llm sont dotées par les données générées par des personnes weird. Notamment, la diversification post-hoc des modèles d’ia pourrait ne pas nécessairement résoudre le problème parce que la notion même de diversité pourrait signifier différentes choses à travers les populations. Par exemple, dans certaines nations, la diversité peut être plus étroitement liée aux différences raciales et ethniques, et dans d’autres nations plus racialement homogènes, elle peut être plus liée aux différences rurales vs urbaines.

Les llm sont entraînés sur des données générées par des humains, leur permettant de comprendre les probabilités de séquences de tokens. En conséquence, ils reflètent les tendances linguistiques humaines façonnées par leur architecture de modèle, ce qui à son tour affecte la façon dont ces modèles abordent les tâches de raisonnement [\[12\]](#). Bender et al. (2021) ont avancé que les llm sont comme des “perroquets stochastiques”, suggérant qu’un modèle de langage “est un système pour assembler au hasard des séquences de formes linguistiques qu’il a observées dans ses vastes données d’entraînement, selon des informations probabilistes sur la façon dont elles se combinent, mais sans aucune référence au sens.” [\[5\]](#). Ici, nous ajoutons un amendement à l’analogie du “perroquet stochastique” et soutenons que les llm sont une espèce particulière de perroquets, parce que leurs données d’entraînement proviennent largement de populations weird : une valeur aberrante dans le spectre des psychologies humaines, à la fois à l’échelle mondiale et historique. La sortie des llm actuels sur des sujets comme les valeurs morales, les questions sociales et la politique sonnerait probablement bizarre et étrange pour des milliards de personnes vivant dans des populations moins

weird.

0.4. Conclusion

Les llm deviennent de plus en plus pertinents dans la vie quotidienne des gens et semblent plausiblement bien positionnés pour automatiser une proportion croissante de la prise de décision dans diverses sociétés. Ainsi, il peut être crucial d'étudier les tendances par lesquelles les llm "pensent", "se comportent" et "ressentent" - en d'autres termes, de sonder leur psychologie. Les ingénieurs et chercheurs en ia comparent généralement la performance des llm avec celle des "humains". Ici, nous démontrons que les llm acquièrent une psychologie weird, peut-être parce que leurs données d'entraînement proviennent massivement d'individus vivant dans des populations weird. Ainsi, les llm peuvent ignorer la diversité psychologique substantielle que nous observons à travers le monde. Ce biais systématique des llm peut avoir des conséquences et des risques sociétaux de grande portée à mesure qu'ils s'intègrent plus étroitement à nos systèmes sociaux, nos institutions et nos processus de décision au fil du temps.

Références

- [1] 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.
- [2] 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*.
- [3] 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.
- [4] 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.
- [5] 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 conference on fairness, accountability, and transparency* (pp. 610-623).
- [6] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand chatgpt-3. *Proceedings of the National Academy of Sciences*, **120**(6), e2218523120.
- [7] 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).
- [8] 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.
- [9] 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.
- [10] 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 chatgpt-4. *arXiv*. <https://doi.org/10.48550/arXiv.2303.12712>
- [11] 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>
- [12] 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>

- [13] Davani, A., Díaz, M., & Prabhakaran, V. (2023). Moral values mediate cross-cultural differences in safety evaluations of Large Language Models (Working paper).
- [14] Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, **27**, 597-600.
- [15] 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.
- [16] Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, **1**, 451-452.
- [17] 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.
- [18] 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.
- [19] Gottfredson, L. S. (1997). Mainstream science on intelligence : An editorial with 52 signatories, history, and bibliography. *Intelligence*, **24**(1), 13-23.
- [20] 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.
- [21] 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 & wvs A Secretariat. <https://doi.org/10.14281/18241.1>
- [22] Hagendorff, T., Fabi, S., & Kosinski, M. (2022). Machine intuition : Uncovering human-like intuitive decision-making in chatgpt-3.5. *arXiv*. <https://doi.org/10.48550/arXiv.2212.05206>.
- [23] Havaladar, 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>.
- [24] Heine, S. J. (2020). *Cultural psychology* (4th ed.). W. W. Norton & Company.
- [25] Henrich, J. (2020). *The weirdest people in the world : How the West became psychologically peculiar and particularly prosperous*. Penguin.
- [26] 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.
- [27] 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.
- [28] Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, **33**(2-3), 61-83.
- [29] 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.
- [30] Inglehart, R. (2020). *Modernization and postmodernization : Cultural, economic, and political change in 43 societies*. Princeton University Press.
- [31] 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>
- [32] Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2302.02083>
- [33] Kuhn, M. H., & McPartland, T. S. (1954). An empirical investigation of self-attitudes. *American Sociological Review*, **19**(1), 68-76.

- [34] 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.
- [35] 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.
- [36] Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is chatgpt-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).
- [37] 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.
- [38] Muthukrishna, M., Henrich, J., & Slingerland, E. (2021). Psychology as a historical science. *Annual Review of Psychology*, **72**, 717-749.
- [39] Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought : holistic versus analytic cognition. *Psychological Review*, **108**(2), 291.
- [40] 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).
- [41] openai. (2023). chatgpt-4 technical report. <https://cdn.openai.com/papers/gpt-4.pdf>.
- [42] 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.
- [43] Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, **366**(6466), eaau5141.
- [44] Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, **120**(10), e2300963120.
- [45] Simmons, G. (2022). Moral mimicry : Large language models produce moral rationalizations tailored to political identity. *arXiv*. <https://doi.org/10.48550/arXiv.2209.12106>
- [46] 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.
- [47] 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.
- [48] Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models know what humans know ? *Cognitive Science*, **47**(7), e13309.
- [49] 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.

Remerciements

Nous remercions Thomas Talhelm pour le partage de données, Ali Akbari pour ses retours sur les analyses, et Frank Kassarits pour avoir suggéré d’explorer ce sujet. Ce travail a été partiellement financé par la John Templeton Foundation (#62161).