Artificial social research? AI's capabilities and risks in predicting values and attitudes

by *Caterina Ambrosio, Ciro Clemente De Falco, Domenico Trezza* [*]

This study explores the potential and limitations of generative artificial intelligence, with a particular focus on ChatGPT-4, in reproducing human values and attitudes based on socio-demographic profiles. By comparing real data from the European Social Survey (ESS) with AI-generated data, the study assesses the ability of AI to reflect public opinion trends. While AI-generated responses exhibit a general alignment with survey data, they show limited variability and some inconsistencies in group-level analyses. In conclusion, despite its promising aspects, the findings suggest that AI cannot replace empirical research.

Keywords: artificial intelligence, experiment, predictive capacity, validity and reliability.

Ricerca sociale artificiale? Capacità e rischi dell'IA nella previsione di valori e atteggiamenti

Questo studio esplora le potenzialità e i limiti dell'intelligenza artificiale generativa, con particolare riferimento a ChatGPT-4, nella riproduzione di valori e atteggiamenti umani sulla base di profili socio-demografici. Attraverso il confronto tra dati reali provenienti dall'European Social Survey (ESS) e dati generati da ChatGPT, si valuta la capacità dell'IA di rispecchiare le tendenze dell'opinione pubblica. Sebbene le risposte generate dall'IA mostrino un allineamento generale con i dati dell'indagine, esse presentano una ridotta variabilità e alcune incongruenze nelle analisi di gruppo. In conclusione, nonostante gli aspetti promettenti, i risultati suggeriscono che l'IA non può sostituire la ricerca empirica.

Parole chiave: intelligenza artificiale, esperimento, capacità predittiva, validità e affidabilità.

Introduction

In recent years, generative artificial intelligence (AI-gen) has increasingly become central to technological progress and research. Like all major technological innovations, it raises ethical, social, and cultural questions. The

DOI: 10.5281/zenodo.17297559

[*] Università di Napoli Federico II. caterina.ambrosio@unina.it, ciroclemente.defalco@unina.it, domenico.trezza@unina.it.

ability of these generative systems to understand and replicate the nuances of human behavior has become – among various issues related to the relationship between AI and society – an intensely debated topic in the social sciences, particularly with respect to the integration of social and cultural competencies into computational models (Floridi, Cowls, 2022). On the one hand, AI-gen promises to offer innovative tools for the analysis and prediction of social phenomena; on the other hand, numerous questions emerge regarding the validity of such approaches and the implications of their application (Mokander, Schroeder, 2022).

This reflection could be a part of a broader and likley epistemological transformation within social research, which has recently seen the rise of big data as a predominant object of study. While big data promises – for many, it threatens – the "end of theory" (Amaturo, Aragona, 2019; Anderson, 2008), today AI-gen and advanced language models seem to present similarly unprecedented scenarios for social research: the potential to generate credible simulations of opinions, attitudes, and values without the need to collect real-world data raises a provocative – and likely exaggerated – question: Are we facing the end of empirical research as we know it?

Building on these premises, our study explores AI's ability to predict personal orientations on values and attitudes solely from socio-demographic data. We ask if a language model can adopt a social category and replicate its opinions and values. By comparing AI-generated responses with European Social Survey (ESS) data using identical profiles, we examine the alignment between AI and human insights. Focusing on value elements and gender issues, our preliminary experiment an existing, yet not yet predominant, ability to predict social opinions. The paper is divided into five sections covering previous research, methodology, results, discussion, and future perspectives.

1. Literature review

The predictive capabilities of AI are generating cross-disciplinary interest, engaging in all fields of study, including the humanities and social sciences (Fan *et al.,* 2024). A recent interest in this field is the ability to simulate human responses to sociological stimuli. This area of research encompasses studies in which the generated data are commonly referred to as "synthetic data" or "silicon samples" (Argyle *et al.,* 2023). The goal of such studies is to determine whether responses simulated by AI to questions on attitudes and opinions are comparable to those collected from real participants.

The study by Argyle and colleagues (2023) was among the first to propose the use of LLMs to simulate responses on political opinions. Specifically, GPT-3 was tested for its ability to replicate voting choices for either the Republican or Democratic candidate based on the socio-demographic characteristics of human respondents. Using data from the 2012, 2016, and 2020 editions of ANES, GPT-3 was conditioned on variables such as gender, ethnicity, age, political ideology, political interest, and state of residence. The model was asked to complete prompts like "In 2016, I voted for...". After comparing GPT-3's responses with real data, the authors concluded that the AI could replicate patterns with high accuracy, but limited variability emerged within specific subgroups.

Two years later, building on this study, Bisbee and colleagues (2024) proposed a similar study. In this case, the authors tested ChatGPT 3.5, a more advanced version. Once again, the comparison dataset consisted of real responses collected from the 2016 and 2020 ANES editions. The authors input the descriptions of 7,530 human respondents into the chatbot using an automated program. For each of these profiles, 30 possible responses were requested. The model then simulated various responses for each human profile. After comparing synthetic data with real data, the results demonstrated an apparent similarity to real data; indeed, the mean scores were highly similar. However, what stood out again was the reduced variability

The studies highlight LLMs' potential and limitations in social research. Both find that ChatGPT generates survey-like synthetic data but warn of risks. Low response variability, biases in training data, and prompt sensitivity make LLMs unreliable for replacing traditional surveys.

The present study pursues the same goal as previous works but aims to overcome some of their limitations. First, it uses the latest publicly available version, ChatGPT-4. This version allows the provision of documents in various formats (docx, Excel, etc.). This enables a more quantitative and structured approach. Second, while the analyzed articles relied on a single approach in prompts to obtain responses, this experiment tested different approaches. Finally, unlike previous studies, which primarily focused on simple and specific questions, such as voting preference, our study selected items from attitude scales designed to explore more complex and nuanced dimensions of personality and human opinions.

2. Methodology

2.1. Research design

The aim of this study was to investigate the ability of generative AI - specifically, the GPT chatbot – to reproduce patterns of values and opinions based on specific socio-demographic profiles. To achieve this goal, two real datasets were compared with two "synthetic" datasets. Specifically, through an experiment that compares AI-generated responses with data from the international European Social Survey (ESS) – Round 11, using identical socio-demographic profiles, we aim to assess the degree of alignment between human-generated and AI-generated data.
The version of ChatGPT used for this study was ChatGPT-4, available as a paid service since May 2024. The 2023 ESS dataset for Italy and Great Britain was used for comparison. These countries were selected because ChatGPT, like other LLMs, is mainly trained on English data (Brown *et al.,* 2020), which could influence its outputs. Italy was included due to expertise in its context. The variables considered in this study include socio-demographic variables as well as variables concerning attitudes and opinions for 1,000 Italian individuals and 1,000 British individuals, sampled from a ESS dataset using simple random sampling. For attitudes and opinions, 2 items were selected from the "Human Values Scale" module and 8 items from the "Gender in Contemporary Europe**"** module. The variables included are presented in the table below (Tab.1)

Tab 1. Variables included in the research.

Variable Description	Response Categories / Coding Details
SOCIO-DEMOGRAPHIC	
Gender	1: Male, 2: Female, 9: No answer
Highest level of education[1]	ISCED codes (0: No ISCED completed to 800: Doctoral degree), 5555: Other, 7777: Refusal, 8888: Don't know, 9999: No answer

[1] For the purpose of the analysis, this variable was recoded as follows: 1 = Low level; 2 = Medium level; 3 = High level.

Household's total net income, all sources[2]	1-10: Income deciles, 77: Refusal, 88: Don't know, 99: No answer
HUMAN VALUES SCALE	
Important to be rich, have money and expensive things (renamed *money*)	1: Very much like me, 6: Not like me at all, 66-99: Missing values
Important to have a good time (renamed (*goodtime*)	1: Very much like me, 6: Not like me at all, 66-99: Missing values
GENDER IN CONTEMPORARY EUROPE	
Good or bad for family life if equal number of men and women are in paid work (renamed *family*)	0: Very bad, 6: Very good, 7-9: Missing values (Refusal, Don't know, No answer)
Good or bad for politics if equal number of men and women hold leadership positions (renamed *politics*)	0: Very bad, 6: Very good, 7-9: Missing values (Refusal, Don't know, No answer)
Good or bad for businesses if equal number of men and women are in higher management (renamed *business*)	0: Very bad, 6: Very good, 7-9: Missing values (Refusal, Don't know, No answer)
Good or bad for economy if men and women receive equal pay for the same work (renamed *economy*)	0: Very bad, 6: Very good, 7-9: Missing values (Refusal, Don't know, No answer)
Dividing the number of seats in parliament equally between men and women (renamed *parliament*)	1: Strongly in favour, 5: Strongly against, 7-9: Missing values
Requiring both parents to take equal periods of paid leave (renamed *periods*)	1: Strongly in favour, 5: Strongly against, 7-9: Missing values
Firing employees who make insulting comments directed at women (renamed *insulting*)	1: Strongly in favour, 5: Strongly against, 7-9: Missing values
Fining businesses when men are paid more than women for the same work (renamed *business*)	1: Strongly in favour, 5: Strongly against, 7-9: Missing values

The various prompts and their corresponding outputs were executed in January 2025. For each country, ChatGPT was provided with a matrix containing completed socio-demographic variables and incomplete attitudinal variables. To avoid computational slowdowns, which occurred during

[2] For the purpose of the analysis, this variable was recoded as follows: 1 = Low income; 2 = Medium income; 3 = High income.

preliminary tests, the two datasets were split into 4 datasets of 250 cases each. The ESS codebook was also provided to aid in understanding the variable labels. Once the matrix and the codebook were supplied, ChatGPT was instructed to perform the following tasks for each session:

1. Confirm understanding of the dataset and codebook. Verify the presence of completed variables (socio-economic and demographic characteristics) and incomplete variables (attitudes and opinions).
2. Identify the socio-economic and demographic profile of the subjects. Extract and interpret specific information for each subject based on the complete variables.
3. Impute responses to the missing variables. Generate responses to the questions on attitudes and values for each individual, based on their defined socio-economic and demographic profile. The responses had to be coherent and substantiated, following the descriptions provided in the codebook and avoiding randomness.

Given the demonstrated sensitivity of AI-generated responses to the prompts used (Argyle *et al.,* 2023; Bisbee *et al.,* 2024), three different prompts were tested for each country, each employing a different approach. A new chat session was initiated for each approach. The three approaches were as follows:

- Predictive approach. In this approach, the AI was asked to predict coherent responses to all the questions on attitudes and values based on the subject's socio-demographic profile.
- Interviewer approach. This approach simulated a direct interview with the subjects represented in the dataset. The AI assumed the role of an interviewer and filled in the questionnaire cells for each socio-demographic profile provided in the Excel file.
- Researcher approach. In this approach, the AI was made aware that it was participating in a social research experiment. The aim was to determine whether it could identify with a reference social category and accurately reproduce the typical patterns of values and opinions of that group.

2.2. Analysis Procedure

To compare real responses with those generated by the three AI models (Section 3.1), we first standardized the scores across the three scales to ensure consistency. We then calculated average scores for each item, analyzing variability through standard deviations. This helped assess whether AI could

replicate the diversity of human responses. Differences in average scores were presented using tables and visualizations, highlighting areas where AI responses aligned with real data.
To examine how well the AI models reflected values and attitudes based on social categories (Section 3.2), we grouped education and income into three levels (low, medium, high). The analysis of gender-related items and values followed a structured approach. For the two value-related items, we conducted an ANOVA with post-hoc tests for all three AI models in Italy and the UK, incorporating socio-demographic factors. The ANOVA identified significant relationships, while post-hoc tests clarified their direction and strength.
For gender-related items, we complemented the ANOVA with a multiple regression model. We aggregated responses into composite scores to reflect overall attitudes, creating an additive index as the dependent variable. Socio-demographic factors were converted into dummy variables, using "high" education and income as reference categories, along with "male" for gender. This approach helped isolate the specific influence of each factor on attitudes toward gender issues.

3. Analysis of results

In this paragraph the ability of AI-generated responses to replicate human survey data will be examined. Section 3.1 analyzes score distributions across different items, comparing human responses (ORIG) with three AI simulation approaches (PRED, INTERV, RESEARCH). Section 3.2 extends the analysis by testing whether AI-generated responses reflect socio-demographic patterns observed in the original data. Using ANOVA and regression models, we assess the consistency of AI predictions with human attitudes across variables such as gender, education, and income.

3.1. Analysis of item scores

This section examines how original survey scores compare with those generated by our three AI simulation approaches. The survey items are grouped into three scales: two on gender equality (GENDER_A and GENDER_B, each with four items) and one on personal and social values (VALUE, with two items). Since each scale uses different metrics, we normalized the scores for fair comparison (tab.2).

At first glance, differences emerge. Human responses show a lively spread (standard deviation: 0.256), reflecting diverse opinions. In contrast, AI scores are more uniform, with deviations of 0.079 for PRED, 0.069 for RESEARCH, and just 0.011 for INTERV. This suggests AI struggles to capture the variability of human thought.
A closer look reveals further nuances. For GENDER_A – celebrating equal inclusion in work, politics, and management – survey scores are consistently higher than AI's, suggesting a strong real-world consensus that AI fails to replicate. Conversely, for GENDER_B – corrective measures like quotas and penalties – AI rates them more favorably, possibly overestimating public support. On the VALUE scale, covering personal aspects like enjoyment and wealth, AI aligns more closely with human responses, indicating a better grasp of universal sentiments.

Tab.2 Mean of normalized item scores

codebook	Item	ORIG	PRED	INTERV	RES
0=bad 1=good	Bad or good for family life in [country] if equal numbers of women and men are in paid work	0.808	0.582	0.501	0.642
0=bad 1=good	Bad or good for politics in [country] if equal numbers of women and men are in positions of political leadership	0.805	0.676	0.507	0.551
0=bad 1=good	Bad or good for businesses in [country] if equal numbers of women and men are in higher management positions	0.825	0.582	0.507	0.637
0=bad 1=good	Bad or good for economy in [country] if women and men receive equal pay for doing the same work	0.882	0.507	0.486	0.573
0=strongly in favour 1=strongly against	Dividing the number of seats in parliament equally between women and men	0.353	0.514	0.508	0.457

0=strongly in favour 1=strongly against	Require both parents to take equal periods of paid leave to care for their child	0.325	0.514	0.487	0.493
0=strongly in favour 1=strongly against	Firing employees who make insulting comments directed at women in the work-place	0.278	0.51	0.51	0.454
0=strongly in favour 1=strongly against	Making businesses pay a fine when they pay men more than women for doing the same work	0.24	0.514	0.5	0.51
0=strongly agree 1=strongly disagree	Important to have a good time	0.45	0.411	0.482	0.508
0=strongly agree 1=strongly disagree	Important to be rich, have money and ex-pensive things	0.608	0.411	0.512	0.474
	DEV.STANDARD	0.256	0.07	0.01	0.06

These differences are visualized in Table 3 using grayscale shading – lighter cells indicate lower scores and closer alignment with original data. GENDER_A items show the biggest discrepancies, GENDER_B items fall in the middle, and VALUE items are simulated most accurately. Notably, the statement "Important to have a good time" has minimal differences, ranging only from 0.03 to 0.05.

Among approaches, RESEARCH performs best, with an overall differential of 0.183, suggesting that when AI is explicitly informed about research objectives, it better mirrors human responses. PRED (0.207) and INTERV (0.226) perform slightly worse, likely due to the challenge of simulating complex individual interactions.

Interestingly, AI struggles most with items featuring intricate linguistic constructions, especially those on economic aspects of gender issues. While modern language models excel in processing language, they still miss subtle nuances in topics like pay equity. In summary, RESEARCH best captures gender equality sentiments, while INTERV slightly outperforms personal values. These findings highlight the importance of context and clear instructions in improving AI's ability to predict social opinions with human-like variability.

Tab.3 Absolute differences with original scores

Issue	Item	PRED	INTERV	RES	tot.avg
GENDER_A	Bad or good for family life in [country] if equal numbers of women and men are in paid work	0,226	0,308	0,166	**0,233**
	Bad or good for politics in [country] if equal numbers of women and men are in positions of political leadership	0,129	0,299	0,254	**0,227**
	Bad or good for businesses in [country] if equal numbers of women and men are in higher management positions	0,243	0,318	0,188	**0,250**
	Bad or good for economy in [country] if women and men receive equal pay for doing the same work	0,375	0,396	0,308	**0,360**
GENDER_B	Dividing the number of seats in parliament equally between women and men	0,162	0,155	0,104	**0,140**
	Require both parents to take equal periods of paid leave to care for their child	0,189	0,162	0,168	**0,173**
	Firing employees who make insulting comments directed at women in the workplace	0,233	0,233	0,176	**0,214**
	Making businesses pay a fine when they pay men more than women for doing the same work	0,274	0,260	0,270	**0,268**
VALUE	Important to have a good time	0,039	0,032	0,058	**0,043**
	Important to be rich, have money and expensive things	0,197	0,096	0,134	**0,142**
	tot. Avg.	***0,207***	***0,226***	***0,183***	**0,205**

3.2. Group analysis

In this section, we will test the approaches' ability to reproduce the values and attitudes of social categories chosen for the analysis. The approach taken for this analysis and described in the 2.2 section produced numerous outputs that have been summarized in the tables 4,5,6. Each table shows the results of a single variable. Each table is organised as follows: in the rows are the

gender-related items and values while in the columns are the datasets for the three approaches used and the original db. Each cell represents the result of the ANOVA between a specific item and a given approach on a single socio-demographic variable. Empty cells indicate that the Anova did not reach the level of statistical significance ($p < 0.05$), while cells with significant results directly report post-hoc test results. Finally, Table 7 reports the results of the multiple regressions. The outputs include the classic regression parameters, and only the results for which the t-test was significant were presented.

The first results we are going to comment on concern gender (table 4), where the original data from Italy (ITA) show significant relationships in the expected direction: the female sex has a more favourable attitude toward gender issues. This trend is observed on both scales. Analyses using ANOVA and regression models confirm these relationships. Of the three IA approaches, the "INTERV" approach shows no similarity with the original data. In particular, no significant relationships emerge between sex and scores on individual items, as evidenced by both ANOVA and the regression coefficient. In contrast, the third approach (RES) shows consistent relationships with the original data in three out of four cases on the first scale. This partial consistency is confirmed by the regression coefficient.

Tab 4. Anova on Gender

ITEM	*Original*	*Prev*	*Interv*	*Research*
family	F	F	-	-
politics	F	F (x)	-	F
business	F	F	-	F
economy	F	F	-	F
parliament	M	F (x)	-	-
periods	-	XX F)	-	-
insulting	M	F (x)	-	-
business	M	F (x)	-	-
goodtime	F	F (x)	-	-
money	-	F (x)	-	-

The first AI approach (PREV), on the other hand, is distinguished by complex and not always consistent behaviours. Significant relationships are observed, but with divergent directions between the two scales: on the first scale, the female sex is significantly more attentive to gender issues, while on the second scale, they are less attentive. This results are consistent with

the original data on the first scale but discordance on the second. An important methodological aspect concerns the presence of the "X" in the tables, indicating cases where Anova was not performed due to the homogeneity of within-group scores. Finally, the regression model confirms what has been observed: the regression coefficient is significant on both scales but shows signs consistent with the original data only on the first scale, while on the second scale, it shows opposite directions. What was observed on the sex variable is also repeated, in part, for educational attainment. In Italy, there are significant relationships in the expected direction for most items and for both scale summary indices.

Tab 5. Anova on Education

ITEM	*Original*	*Prev*	*Interv*	*Research*
family	L<M=H	L<H	-	-
politics	L<M=H	L>M=H	-	-
business	L<M<H	L<H	-	-
economy	L<M=H	L>M=H	-	-
parliament	L>M=H	L>M=H	-	-
periods	L>M=H	L>M=H	-	-
insultin	-	L>M=H	-	-
business	L>M>H	L>M=H	-	-
goodtime	L>M=H	L>M=H	-	-
money	L>M=H	L>M=H	-	-

The "INTERV" approach shows no similarity with the original data in Italy as confirmed by the regression coefficients on both scales. The RESEARCH approach IA shows no significant relationships between education and attitudes toward gender issues, except for two items on the first scale, where the direction of the relationship is opposite to that expected. For the PREV approach, in Italy, on the first scale, the regression coefficient is consistent with the original data, but only two out of four items show the expected direction. On the other two items, the relationship exists but is of opposite sign. On the second scale, there are relationships consistent with the original data on three out of four items. However, in the regression, the effect of education is cancelled out by the influence of sex. Finally, regarding income, significant relationships in the expected direction are observed on the first scale.However, the PREV approach shows no significant relationship between income and attitudes toward gender issues.

Tab 6. Anova on income

ITEM	*Original*	*Prev*	*Interv*	*Research*
family	L<H	-	L>M=H	-
politics	L<M=H	-	-	-
business	L<M=H	-	L<M=H	-
economy	L=M<H	-	-	-
parliament	L>M=H	-	-	-
periods	-	-	-	-
insultin	L>M=H	-	-	-
business	L>M=H	-	-	-
goodtime	L>M>H	-	-	-
money	-	-	-	-

The INTERV approach shows a significant relationship only in Italy but with a coefficient of the opposite sign from the original data. Finally, the RES approach shows. The ninth and tenth items show significant weaknesses in the ability of the AI approach to replicate the original relationships between socio-demographic variables and scores correctly. Focusing on significant relationships, we note that the second approach never shows significant relationships, confirming its general ineffectiveness. The RESEARCH approach also fits into this picture of weakness but with an exceptions on income The PREV approach, however, is notable for the consistent presence of significant relationships on gender, education and scores on items.

Tab 7. Regression model

	Value	*Original*	*Prev*	*Interv*	*Ob_dich*
First Scale	*education*	1,27	0,74		
	gender	-0,94	-7,9		-0,78
	Income	0,93	0,19		
	R_square	0,44	0,957		0,012
	R_square_corr	0,39	0,957		0,007
	T Test	Yes	Yes		0,053
Second Scale	*education*	-0,87			
	gender	1,42	-8		
	Income			0,717	
	R_square	0,66	1	0,013	

R_square_corr	0,61	1	0,008
T Test	Yes		Yes

Conclusion

This study compared real survey data with responses generated by an AI model (based on GPT) to assess the ability of AI to reproduce patterns of values and opinions based on socio-demographic variables. The analysis was conducted in two stages: the first examined differences at the individual question level; the second considered socio-demographic factors.

The results at the item level show that there are differences between the actual and simulated data, especially for questions related to gender and values. These discrepancies may have both linguistic (more complex wording) and semantic (some gender issues are still "difficult" for the model) causes. As found in other empirical studies, the AI-generated responses here also show little variability.

However, AI performance improves when clear objectives and context are provided: in the "RESEARCH" approach, which made the study's purpose and setting explicit, results appeared more consistent with actual data than in the "PRED" (simple prediction) or "INTERV" (interviewer role) modes. Despite this, when socio-demographic variables are introduced, the relationships between them and views on gender are often found to be inconsistent with those observed in the actual data. Overall, the "PRED" approach generated the most significant relationships, but in several cases misaligned with the actual data. The lack of variability in the responses sometimes prevented statistical tests (ANOVA) from being performed, suggesting a potential inconsistent use of certain variables by the AI. Where, on the other hand, results are consistent, high R^2 values suggest overfitting or overestimation of ratios.

Considering this evidence, it can be concluded that although AI is able to capture some general trends on social values and demographic profiles, it still cannot reliably reproduce complex nuances and relationships, especially in sensitive contexts such as gender issues. In addition, the design of the interaction-clear instructions and well-defined goals-has a significant impact on the ability of AI to generate consistent responses. Ultimately, while language models can support social research, they cannot yet completely replace empirical data.

This reflection not only concerns the academic community but is part of a broader debate that also involves the world of work, whose experts are

questioning how far artificial intelligence will be able to replace human labor. (Gmyrek, Berg, Bescond, 2023), in a recent paper for the ILO estimated the impact of generative technologies such as GPT-4 on occupations globally. Even outside the strictly scientific realm, the conclusions converge with the findings of this study: despite the significant improvement in the predictive and linguistic capabilities of the latest generation models, the future in which AI will completely replace human labor is still far off. It is true that the technological leap has meant that it is no longer only simple, manual or repetitive jobs that are potentially exposed to automation, but also certain categories of "cognitive work" particularly administrative professions. However, as the ILO authors point out, this exposure is partial: AI can automate some tasks, but it can hardly replace the full range of activities that characterize a profession. The result, rather than a process of total replacement, seems to take the form of a transformation of work.
Future studies may make use of different AI models and longitudinal analyses to assess how evolution and continuous fine-tuning affect this performance.

References

Amaturo E., Aragona B. (2019). Per un'epistemologia del digitale: note sull'uso di big data e computazione nella ricerca sociale. *Quaderni di Sociologia* [Online], 81- LXIII | 2019, online dal 01 juin 2020, consultato il 09 février 2025. http://journals.openedition.org/qds/3508; DOI: https://doi.org/10.4000/qds.3508

Anderson C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7): 16-07.

Argyle L.P., Busby E.C., Fulda N., Gubler J.R., Rytting C., Wingate D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3): 337-351.

Azzadina I., Huda A.N., Sianipar C.P.M. (2012). Understanding relationship between personality types, Marketing-mix factors, and purchasing decisions. *Procedia-Social and Behavioral Sciences*, *65*: 352-357.

Bisbee J., Clinton J.D., Dorff C., Kenkel B., Larson J.M. (2024). Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, *32*(4): 401-416.

Bodroža B., Dinić B.M., Bojić L. (2024). Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, *11*(10), 240180.

Brown T., Mann B. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*: 1877-1901.

Dillion D., Tandon N., Gu Y., Gray K. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences*, *27*(7): 597-600.

Fan L., Li L., Ma Z., Lee S., Yu H., Hemphill L. (2024). A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, *15*(5): 1-25.

Gmyrek P., Berg J., Bescond D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality. *ILO working paper*, *96*.

Mökander J., Schroeder R. (2022). AI and social theory. AI & SOCIETY, 37(4): 1337-1351.

Patrinos G.P., Sarhangi N., Sarrami B., Khodayari N., Larijani B., Hasanzad M. (2023). Using ChatGPT to predict the future of personalized medicine. *The Pharmacogenomics Journal*, *23*(6): 178-184.

Rafikova A., Voronin A. (2024). AI as a Research Proxy: Navigating the New Frontier of Social Science Inquiry through Language Models.

Vidali M. (2024). Intelligenza Artificiale in Medicina: implicazioni e applicazioni, sfide e opportunità. *biochimica clinica*, *48*(2): 129.