Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

# ASSESSING OCCUPATIONS THROUGH ARTIFICIAL INTELLIGENCE: A COMPARISON OF HUMANS AND GPT-4

Paweł Gmyrek
International Labour Organization

Christoph Lutz
BI Norwegian Business School

Gemma Newlands
University of Oxford

## Introduction and Literature Review

The occupational landscape in many countries has changed profoundly. New occupations have emerged around technologies such as artificial intelligence (AI), big data and social media (Kane, 2017; Newlands & Lutz, 2024b). Such technological shifts have also had a tangible impact on job quality (Makridis & Han, 2021). Job quality is multifaceted (Adamson & Roper, 2019), encompassing not only 'hard' aspects such as working conditions and pay but also 'soft' criteria such as individuals' subjective interpretations of their work. Occupations carry societal values, with society forming subjective views about their prestige and standing (Lambert & Bihagen, 2014). Occupational evaluations are influenced by media, societal narratives, and assumptions about an occupation's characteristics, rewards, and exclusivity (Mejia et al., 2021).

Investigations of societal occupational evaluation predominantly involve surveys with human participants, especially from Treiman's (1977) Standard International Occupational Prestige Scale (SIOPS; see for example Lersch et al., 2020; Pitt & Zhu, 2019). The literature on occupational prestige typically encompasses measurements related to prestige, social standing, social status, and social value. The intersection between prestige and social value is particularly pertinent, given the growing body of research on the perceived (low) social value of occupations, notably in relation to Graeber's (2018) discourse on 'Bullshit Jobs.'

Suggested Citation (APA): Gmyrek, P., Lutz, C., & Newlands, G. (2024, October). *Assessing Occupations Through Artificial Intelligence: A Comparison of Humans and GPT-4*. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

Collectively, existing studies reveal a certain consensus how different occupations are evaluated, even though such evaluations remain dynamic and subject to variation (Avent-Holt et al., 2020; Valentino, 2020; Valentino, 2021). We follow this idea by differentiating two distinct axes of occupational evaluation: occupational prestige (OP) and occupational social value (OSV). The former captures whether an occupation is seen as prestigious (i.e., whether it has high esteem), whereas the latter includes perceptions of social utility (i.e., whether an occupation is seen as useful).

Our study links such evaluations of occupations with the rapid advancement of AI. There is an ongoing debate about AI's role in complementing or replacing human labour (Frey & Osborne 2017; Gmyrek et al., 2023). The use of algorithms in workforce management, often termed 'algorithmic management' (Jarrahi et al., 2021), exemplifies this transformation, with diverging assessments of the very concept (Berg et al., 2018; Jarrahi et al., 2022; Shestakofsky, 2017). Particularly, Large Language Models (LLMs), such as GPT-4 and Llama, represent a pivotal advancement, due to their capacity of content analysis and creation across unstructured textual, visual, and audio-visual media.

Emerging literature shows that these new LLMs not only process data, but also form discernible opinions on societal aspects. Since they are trained on historical data, LLMs reflect and perpetuate societal biases (Saetra, 2023). We contribute to the rapidly expanding body of literature examining LLMs and their social perceptions (Argyle et al., 2023a, 2023b), basing our discussion on the concept of a 'technological construction of society' (Authors, 2024). This idea responds to STS perspectives by Klein and Kleinman (2002) and Pinch and Bijker (1984) regarding the social construction of technology. Technological construction of society suggests inverted roles. It refers to the notion that technology, especially digital technology, exerts substantial influence on the configuration and perception of various aspects of modern society. Viewing technology as a potential shaper of human perceptions and communication raises important questions about the rules for its integration into different areas of human activity such as the world of work. To respond to this challenge, we provide an in-depth analysis of the OP and OSV perceptions of GPT-4, contrasting the LLM's scores with those of human respondents.

## Methods, Results and Discussion

We rely on a list of 576 occupation titles, aligned with the International Standard Classification of Occupations (ISCO-08; ILO, 2008), and four non-occupational roles (Unemployed, Retiree, Student, Homemaker). We matched every occupation title to one ISCO-08 unit group. The full occupation list also includes a core list, with exactly one occupation title for each of the 130 ISCO-08 minor groups, selected to be highly known and institutionalised (e.g., Taxi Driver, Butcher, Journalist, Dentist).

To measure OP and OSV comprehensively, we developed a scalable approach, where occupational titles are scored on a 0-100 scale with a slider (Newlands & Lutz, 2024a, 2024b). We did not explain the concepts of OP and OSV, in order not to prime the

respondents. For the recruitment of survey participants, we relied on Prolific. We collected OP and OSV assessments in March 2022, using Prolific's representative sample option for the UK, where the platform selects the respondents across age, sex, and ethnicity to mirror the population distribution. The study reward was £2.50, with a median response time of 19 minutes (SD = 15 minutes). OP and OSV were assessed in separate surveys as we did not want the same respondents to score occupations on these two dimensions concurrently to avoid priming effects and to maintain statistical independence. Consequently, the surveys were launched sequentially with screening out for previous participation in any data collection. Our final sample size is 2429 respondents (1219 for OP and 1210 for OSV). 48.7% of the respondents identify as male, 50.6% as female, and the remaining 0.7% as non-binary. The average age is 44 years (SD = 16 years). Ethnicity and education match the UK Office of National Statistics distribution.

GPT scores were generated using a Python script that accesses GPT-4 through the OpenAI library. The script is organised as a loop of sequential API calls that process each of the 580 occupations individually, with an exponential backoff and retry option set to handle any API response errors. We request a written justification of each score in the first round of predictions (see Appendix).

We proceeded in four steps for the empirical analyses.

1) We compared GPT-4 raw scores with survey scores. Despite similarities, a simple visual breakdown across the main demographic characteristics confirms that GPT-4 overestimates scores across the board, when compared to human scores (Figure 1 for OSV, same for OP).

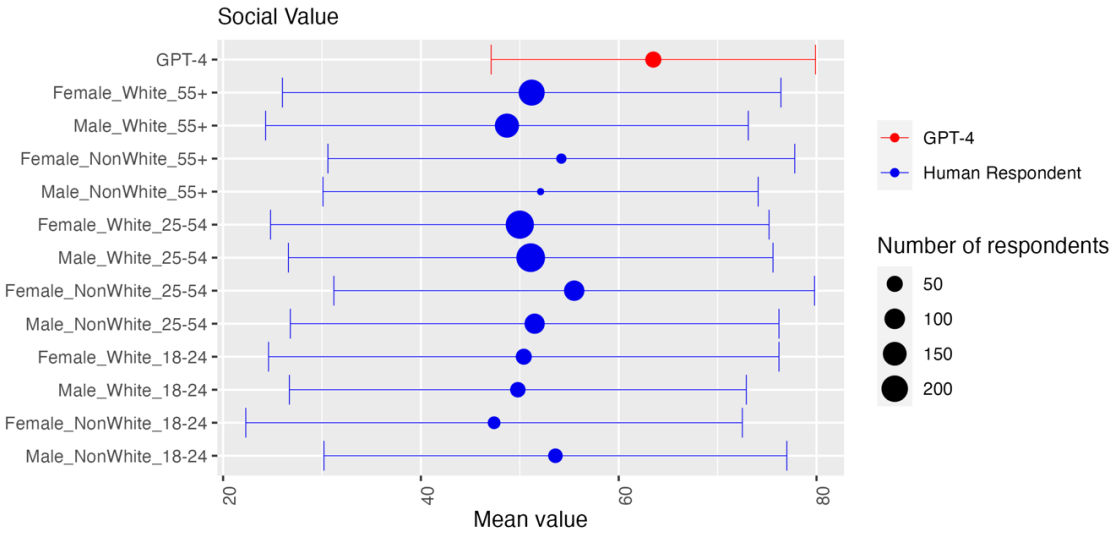


*Figure 1: GPT-4 OSV scores vs. respondent scores*

2) The scores were z-standardized and GPT-4 and human scores were subsequently correlated. This showed high consistency between humans and GPT-4 (correlations mostly between 0.7 and 0.9).

3) We focused on outlier occupations, where the absolute difference in z-scores between humans and GPT-4 exceeds 0.5 (see Figure 2 for OSV, similar patterns for OP).

4) We analysed which demographic groups are best/worst captured by GPT-4 scores by splitting the human survey across age (18-24/25-54/55+), gender (male/female) and ethnic majority/minority status (White/Non-White). GPT-4 is most aligned with white men and women above 25 years and least aligned with non-white men of all age groups, and non-white women 55+.

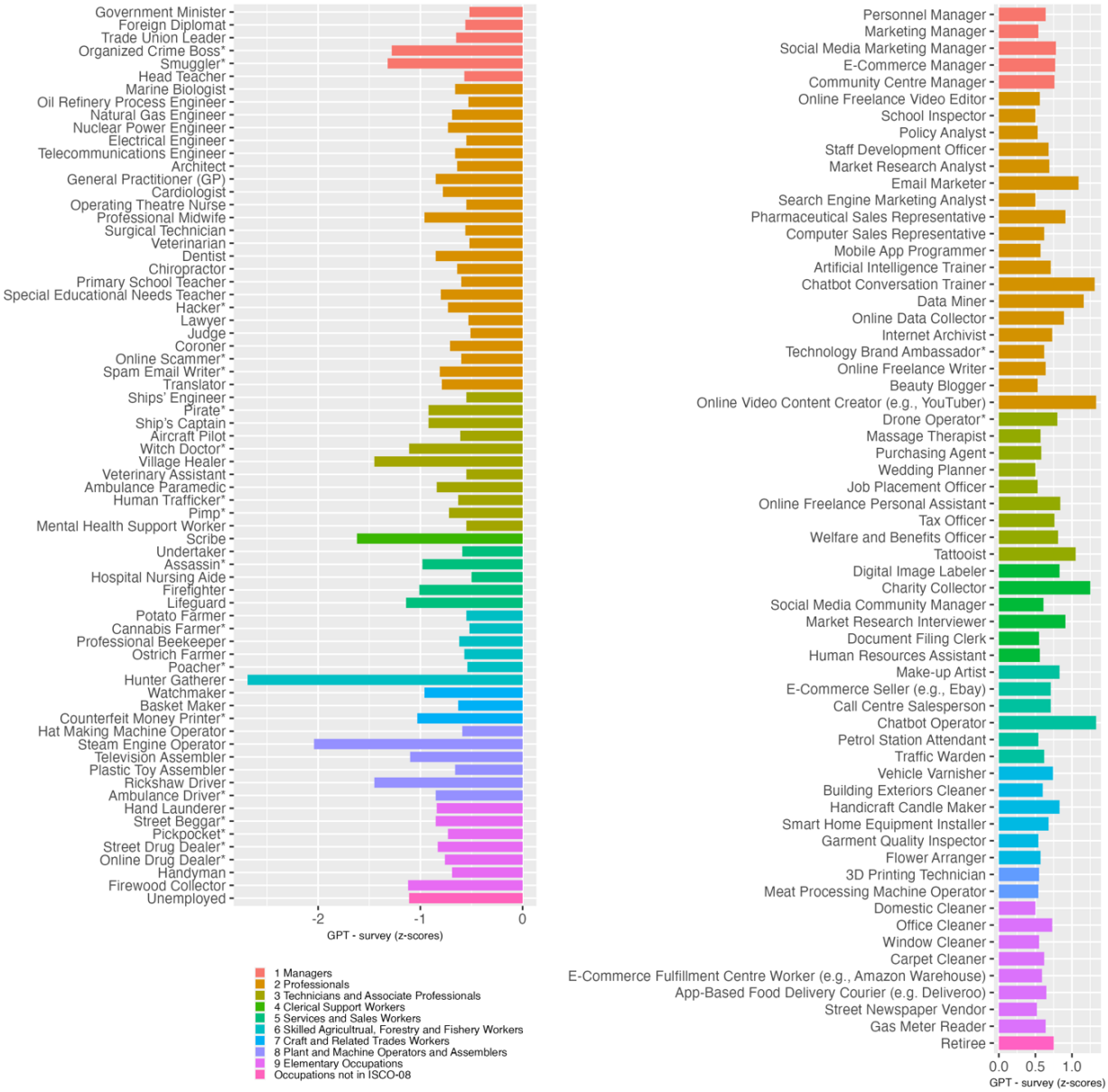


*Figure 2: Occupations with largest OSV difference between GPT-4 and humans*

Overall, our research shows that GPT-4 accurately reflects the OP and OSV perceptions of UK residents. However, the model overestimates OP and OSV scores compared to human respondents, an issue that can be corrected through data transformations but remains a limitation. It also struggles with capturing nuanced opinions for specific occupations, particularly those related to the digital economy and the shadow sector, aligning with research on LLM biases (Cheng et al., 2023). GPT-4's evaluations poorly represent minority groups. These disparities may be influenced by historical biases and systemic inequalities where LLMs primarily reflect WEIRD perspectives (Atari et al., 2023). While GPT-4 is a useful tool for probing social perceptions, it cannot fully replace nuanced human surveys.

A more detailed description of the methods, results and implications is available in a peer-reviewed article that has in the meantime been published (Gmyrek et al., 2024).

**References**


Adamson, M., & Roper, I. (2019). 'Good' jobs and 'bad' jobs: Contemplating job quality in different contexts. *Work, Employment and Society, 33*(4), 551-559.

Argyle, L. P., Busby, E. C., Fulda, N., et al. (2023a). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337-351.

Argyle, L. P., Busby, E., Gubler, J., et al. (2023b). AI chat assistants can improve conversations about divisive topics. *arXiv Preprint*, 2302.07268. https://arxiv.org/abs/2302.07268

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which humans? *PsyArXiv Preprints*. https://osf.io/preprints/psyarxiv/5b26t

Avent-Holt, D., Henriksen, L. F., Hägglund, A. E., Jung, J., Kodama, N., Melzer, S. M., ... & Tomaskovic-Devey, D. (2020). Occupations, workplaces or jobs? An exploration of stratification contexts using administrative data. *Research in Social Stratification and Mobility, 70*, 100456.

Berg, J., Furrer, M., Harmon, E., Rani, U., & Silberman, M. S. (2018). Digital labour platforms and the future of work. *International Labour Organization* (ILO), Geneva. https://wtf.tw/text/digital_labour_platforms_and_the_future_of_work.pdf

Cheng, M., Durum's, E., & Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv Preprint*, 2305.18189. https://arxiv.org/abs/2305.18189

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? T*echnological Forecasting and Social Change, 114*, 254-280.

Gmyrek, P., Berg, J., & Bescond, D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality. *International Labour Organization* (ILO), Geneva. https://www.ilo.org/global/publications/working-papers/WCMS_890761/lang--en/index.htm

Gmyrek, P., Lutz, C., & Newlands, G. (2024). A technological construction of society: Comparing GPT-4 and human respondents for occupational evaluation in the UK. *British Journal of Industrial Relations*, 1-29. https://doi.org/10.1111/bjir.12840

Graeber, D. (2018). *Bullshit jobs: A theory*. Penguin Books.

ILO (2008). International Standard Classification of Occupations: Structure, group definitions and correspondence tables (ISCO-08). *International Labour Organization* (ILO), Geneva. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_172572.pdf

Jarrahi, M. H., Lutz, C., & Newlands, G. (2022). Artificial intelligence, human intelligence and hybrid intelligence based on mutual augmentation. *Big Data & Society, 9*(2), 1-6.

Jarrahi, M. H., Newlands, G., Lee, M. K., Wolf, C. T., Kinder, E., & Sutherland, W. (2021). Algorithmic management in a work context. *Big Data & Society, 8*(2), 1-14.

Kane, G. C. (2017). Digital maturity, not digital transformation. *MIT Sloan Management Review, 1*(1), 1-15.

Klein, H. K., & Kleinman, D. L. (2002). The social construction of technology. *Science, Technology, & Human Values, 27*(1), 28-52.

Lambert, P. S., & Bihagen, E. (2014). Using occupation-based social classifications. *Work, Employment and Society, 28*(3), 481-494.

Lersch, P. M., Schulz, W., & Leckie, G. (2020). The variability of occupational attainment. *American Sociological Review, 85*(6), 1084-1116.

Makridis, C. A., & Han, J. H. (2021). Future of work and employee empowerment and satisfaction. *Technological Forecasting and Social Change, 173*, 121162.

Mejia, C., Pittman, R., Beltramo, J. M., et al. (2021). Stigma & dirty work: In-group and out-group perceptions of essential service workers during COVID-19. *International Journal of Hospitality Management, 93*, 102772.

Newlands, G., & Lutz, C. (2024a). Occupational prestige and occupational social value in the United Kingdom: New indices for the modern British economy. *Research in Social Stratification and Mobility*, *91*, 100935.

Newlands, G., & Lutz, C. (2024b). Mapping the prestige and social value of occupations in the digital economy. *Journal of Business Research, 180*, 114716.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts. *Social Studies of Science, 14*(3), 399-441.

Pitt, R. N., & Zhu, L. (2019). The relationship between college major prestige/status and post-baccalaureate outcomes. *Sociological Perspectives, 62*(3), 325-345.

Sætra, H. S. (2023). Generative AI: Here to stay, but for good? *Technology in Society, 75*, 102372.

Shestakofsky, B. (2017). Working algorithms: Software automation and the future of work. *Work and Occupations, 44*(4), 376-423.

Treiman, D. J. (1977). *Occupational prestige in comparative perspective*. Academic Press.

Valentino, L. (2020). The segregation premium: How gender shapes the symbolic valuation process of occupational prestige judgments. *Social Forces, 99*(1), 31-58.

Valentino, L. (2021). The heterarchy of occupational status. *Sociological Forum, 36*, 1395-1418.

## Appendix

The GPT-4 prompt used was as follows:

```
{"role": "system", "content": "You are a respondent to a survey in the UK."
"Make your response representative of an UK average resident, taking into
account the social and demographic diversity of the UK." },
{"role": "user", "content": "Look at this occupation title: " +
str(Occupation_Title) +
"Rate the prestige of this occupation on a scale of 0 (the lowest level of
prestige) to 100 (the highest level of prestige), with intervals of 1. "
"Make your response representative of an UK average resident, taking into
account the social and demographic diversity of the UK. "
             "In the first line of response provide only your numerical score,
             no other text."
             "In the following lines provide a justification of your scoring
             decision. "
             "Do not give any score ranges just one score."}
```

Once all occupations have been processed, the procedure repeats 49 times, resulting in a total of 50 estimation cycles. Since we do not generate justifications for these remaining 49 cycles, the “user” part of the script is amended:

```
{"role": "user", "content": […] "In the first line of response provide only
your numerical score,
no other text. Do not give any score ranges just one score."}
```