Synthetic Participants: What Are They for in Business and How Can We Make Them Better?

PAOLA SCHIETEKAT, MARÍA LÓPEZ[1], AND JUAN DE RUS

Neovantas

DARIO KRPAN

London School of Economics and Political Science

Click or scan QR code
to watch video summary

Synthetic participants, commonly defined as artificially created profiles that simulate human responses, behaviours and decisions, are increasingly used as a complementary tool to traditional research methods. Despite their growing adoption, however, their capabilities, limitations and appropriate applications remain poorly systematised. This article develops a structured framework for understanding and deploying synthetic participants in business contexts. First, it introduces a functional classification based on primary objectives: declarative (intent-based), behavioural (action-based) and conversational (multi-agent) models. Second, it reviews key studies within each category, identifying common performance patterns and practical applications. Third, it proposes a differentiated methodological approach grounded in a quantitative-first perspective, integrating behavioural economics through structured libraries and tailoring models to specific use cases. Using a controlled experimental replication, the paper demonstrates that synthetic participants can approximate human data under structured conditions.

Introduction

The rapid proliferation of synthetic participants suggests a potential shift in how market research is conducted, offering clear advantages in scalability, speed and the ability to conduct research in contexts that would otherwise be inaccessible. Yet, despite their promise, the current landscape presents important challenges.

First, in the business context, we have heard of many approaches that rely heavily on the generation of synthetic participants from qualitative inputs, such as high-level buyer persona descriptions derived from small-scale interviews and grounded in descriptive, self-reported data. Furthermore, outputs are sometimes evaluated through subjective expert judgement rather than standardised quantitative metrics. For instance, Schramowski et al. (2022) assess GPT-3-based respondents by primarily using qualitative ratings of plausibility. This matters because decisions made based on plausible-looking but mis-calibrated outputs carry substantial business risk. Second, there is limited integration of psychological and behavioural economics constructs, which might be a significant limitation, as AI systems struggle to reproduce cognitive biases, heuristics and context-dependent preferences without explicit

1 Corresponding author: mlopez@neovantas.com

structuring. Aher et al. (2023) show that LLMs only exhibit human-like loss aversion and fairness when these mechanisms are embedded in the task design as behavioural dictionaries or knowledge bases.

Third, many synthetic panel solutions are built as generalised tools that favour ease of use over rigorous, bespoke, contextualised approaches. This makes their outputs difficult to distinguish from those obtained through ad hoc prompt-based generation with general purpose LLMs.

The significant and widely anticipated advantages, combined with these challenges and the limited number of publicly available business case studies outside academia, contribute to a mix of enthusiasm and a degree of scepticism among organisations considering their adoption.

What Types of Synthetic Participants Exist, and What Are Their Potential Business Applications?

This article aims to provide a simple classification of the types of synthetic participants that can be developed, along with the business areas in which they can be applied (see Table 1).

This classification may evolve as the field develops and new use cases emerge. For now, though, we believe

Table 1: Types of Synthetic Participants: Objectives and Business Applications

Type	Objective	Main Business Applications	Advantages	Challenges
Declarative (Intent-Based) Models	To simulate responses to structured or semi-structured survey instruments. Currently account for most published applications.	Market research, Customer Experience	Scalable survey prototyping, access to hard-to-reach populations, counterfactual scenario testing, hybrid approaches improve depth and efficiency.	Inflated effect sizes, compressed response distributions, high sensitivity to prompt phrasing, inherent intention-behaviour gap.
Behavioural (Action-Based) Models	To replicate specific reactions or behaviours in response to a stimulus, rather than stated intentions.	UX research, Pricing, Marketing	Can match or exceed human specialist performance on usability tasks, no fatigue, detailed interaction logs at scale, persona-aligned journey simulation.	Compressed variability, fails to replicate cognitive inertia or bounded attention, explicit behavioural scaffolding required for realistic behaviour.
Conversational (Multi-Agent) Models	To simulate conversations between two or more agents using a multi-agent architecture guided by scripts or structured instructions.	Sales, Customer Service	Enables testing of sales scripts and communication strategies at scale, allows study of opinion dynamics and information diffusion.	Risk of anthropomorphism, WEIRD-centric cultural bias, value lock-in, plausible outputs may mask mechanistically different cognition.

it offers a practical framework that helps researchers and practitioners grasp the diversity and distinct purposes of synthetic participants.

What Does the Literature Tell Us About the Performance of Synthetic Participants?

Declarative Synthetic Participants

Declarative or intent-based, synthetic participants are designed to provide survey-like responses. They can generate large volumes of responses quickly and support rapid prototyping of questionnaires and experimental designs before human deployment (Aher et al., 2023; Arora et al., 2025; Doudkin et al., 2025; Park et al., 2023; MRS, 2024). Beyond efficiency, these models help explore counterfactual scenarios and hard-to-reach populations (Aher et al., 2023; Lin & Dia, 2025).

A study by Arora et al. (2025), in which a human-LLM hybrid approach ultimately led to efficiency and effectiveness gains in quantitative and qualitative evaluations of a food company, demonstrates clear business applications. Results indicated that qualitative LLM-generated responses were superior in depth and insightfulness.

However, research consistently demonstrates that declarative models inflate effect sizes, with Fisher z values and standardised effects often two to three times larger than those observed in human samples, and produce narrower, smoother response distributions, underrepresenting population heterogeneity (Dillion et al., 2023; Doudkin et al., 2025; Maier et al., 2025; Stamatogiannakis et al., 2025; Ustiyanovych & Krpan, 2025). They are also highly sensitive to prompt phrasing, with small changes significantly altering reported attitudes or preferences (Gao et al., 2024; Tjuatja et al., 2024).

Additionally, although highly useful for accelerating market research, both private companies and public institutions should remain aware that any declarative research method, whether based on real humans or synthetic participants, will exhibit a gap relative to actual behaviour, as shown in Sheeran's meta-analysis, where intentions explain only around 28% of the variance in real-world behaviour (Sheeran, 2002). This is primarily because these methods rely on self-reports of people's views, attitudes and intentions (e.g., "I want to save money every month"), which can diverge from actual behaviour due to contextual or emotional factors, habits or shifts in priorities at the moment of decision-making (e.g., failing to save due to unexpected expenses).

"Any declarative research method, whether based on real humans or synthetic participants, will exhibit a gap relative to actual behaviour."

Behavioural Synthetic Participants

Behavioural – or action-based – synthetic participants focus on what agents do rather than what they say, simulating concrete interactions with interfaces, products or decision environments.

In UX research, they can execute long interaction sequences without fatigue, identify usability issues and provide detailed feedback at scale (Lu et al., 2025; Zhong et al., 2024). For example, Zhong et al. (2024) report that GPT-4, used as an expert evaluator of interface screenshots, detects 73–77% of usability problems versus 57–63% for human specialists, demonstrating that behavioural agents can match or exceed human performance in some metrics.

Behavioural synthetics have evolved beyond UX, though. A study by Mansour et al. (2025), for instance, consisted of mining historical e-commerce data to build Persona Aligned Agentic Retail Shoppers (PAARSs), used to simulate personalised journeys from product discovery all the way to purchase.

Nevertheless, limitations remain. Behavioural synthetics often confirm compressed variability, inflated treatment effects and stereotyped patterns, failing to capture real-world noise, cognitive inertia or bounded attention (Doudkin et al., 2025; Stamatogiannakis et al., 2025; Ustiyanovych & Krpan, 2025). Embedding psychological and behavioural economics constructs, such as heuristics, loss aversion, social norms or context-dependent preferences, is challenging, requiring explicit scaffolding to approximate realistic behaviour (Aher et al., 2023; Maier et al., 2025). Technical safeguards, such as semantic similarity ratings, retrieval-augmented generation and iterative calibration against human benchmarks, can help mitigate bias and overconfident inferences (Cui et al., 2024; Maier et al., 2025; Suh et al., 2025).

Conversational Synthetic Participants

Conversational models simulate dialogues between agents. They are particularly useful for testing sales scripts, customer service interactions and communication strategies (Lu et al., 2025), but they also enable the study of opinion dynamics and information diffusion, as shown in Park et al. (2023) and Qu and Wang (2024).

However, these models raise important epistemological concerns, as they can produce highly plausible conversations while relying on fundamentally different mechanisms from human cognition (Shanahan et al., 2023). This creates a risk of anthropomorphism, where outputs are misinterpreted as genuine human-like understanding (Mitchell & Krakauer, 2023; Vallor, 2024).

Additionally, conversational models are particularly vulnerable to cultural bias and value lock-in. They often reproduce WEIRD-centric perspectives, leading to systematic misrepresentation of non-Western or underrepresented groups (Agnew et al., 2024; Amini, 2025; Santurkar et al., 2023).

Our Approach to Synthetic Participants

Our Vision

When our team at Neovantas began exploring synthetic participants, we quickly became convinced that integrating them as a tool into companies' research and experimentation methodologies could provide a clear, short-term competitive advantage.

As we started working on real projects, it soon became apparent that a 'one-size-fits-all' approach would likely be suboptimal. A declarative synthetic participant model, designed to answer surveys, is fundamentally different from a behavioural model, designed to simulate concrete actions in terms of configuration, prompting, input and output. Furthermore, even within these categories, the business objective, whether UX, pricing or debt collection, requires entirely different approaches. In practice, we are developing specific, specialised synthetic participant models, each with its own logic, tailored to each use case according to the business need. Moreover, we apply a quantitative approach, both in preparing the model's input and in validating the results.

In addition, we hypothesised that enriching synthetic participants' input with a context drawn from behavioural economics and psychology, through structured knowledge libraries, would help synthetic participants' responses and behaviours more closely resemble human reality. This would allow for better output calibration and enrich the reasoning behind the participants' answers or decisions.

"We are developing specific, specialised synthetic participant models, each with its own logic, tailored to each use case according to the business need."

Putting It to the Test

To test this hypothesis, we conducted the same exercise with declarative synthetic participants, comparing results with and without this enriched input from the libraries. The original exercise we selected was a randomised controlled trial conducted in 2017 with 199 Spanish working adults aged 35 to 65, who were surveyed about their willingness to save (WTS) for retirement and their impulsiveness. The study aimed to evaluate whether exposing a treatment group to a pension calculator would increase their WTS by making the benefits of saving more tangible. The intervention led to a statistically significant WTS increase of 25% ($p = 0.044$).

Our goal was to see whether exposing 199 synthetic participants, matched to the original population in sociodemographic characteristics such as age and gender, and financial attributes such as income and financial literacy, to the same pension calculator stimulus would produce similar results, with WTS as the outcome. We also aimed to assess whether enriching the input with structured knowledge libraries on the most relevant behavioural economics and psychology principles for saving decisions would bring the outcomes even closer to the original results.

To achieve this, we created two libraries:

Behavioural library. This dictionary described each bias, its mechanism and manifestation in the context of WTS, its interaction across sociodemographic variables and under a saving saliency treatment. Interaction rules guided the direction and intensity of activation. Each bias was drawn from research on behavioural economics and financial behaviour.

Impulsiveness library. Impulsiveness scores were generated for each synthetic participant using the

Barratt Impulsiveness Scale, BIS-11, a 30-item instrument. Responses were sampled from the empirical distributions observed in the original dataset, and total scores were computed accordingly.

In this case, both libraries were created as JSON files and retrieved by the model through prompting.

The 'basic' solution, made up of synthetic participants generated without a behavioural knowledge base, produced the weakest distributional fit. KS statistics[2] in this condition ranged from 0.82 to 0.83, and real distributions diverged substantially across nearly the entire outcome range. This pattern is consistent with the literature, in that 'one-size-fits-all' generative models tend to produce WTS estimates that are both more uniform and more concentrated around the centre of the scale than the real data, thus underrepresenting the natural spread of the human distribution.

> "Introducing a behavioural library produced a meaningful improvement representing a reduction in distributional divergence of approximately 15% relative to the baseline."

Introducing a behavioural library produced a meaningful improvement. The GPT-5.4-mini model achieved a KS statistic of 0.70 in the library-only

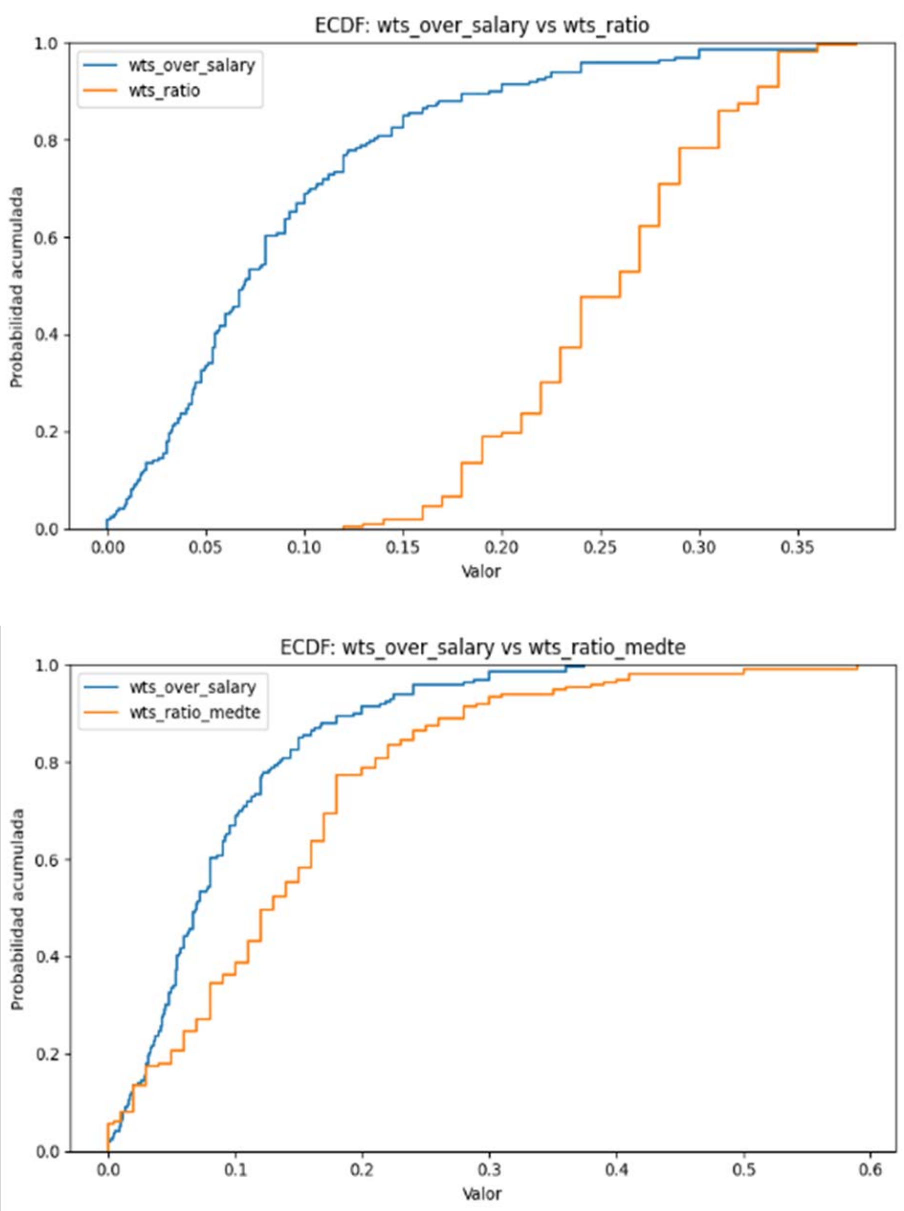

Figure 1: Comparison between synthetic models generated without dictionaries (top) and with behavioural and impulsiveness dictionaries (bottom).

2 The Kolmogorov-Smirnov (KS) statistic is a measure of the biggest difference between two datasets.

condition, representing a reduction in distributional divergence of approximately 15% relative to the baseline (even though it still indicates meaningful residual divergence). Including the Impulsiveness library yielded a KS statistic of 0.54 with the standard prompt configuration, reflecting a substantive gain (see Figure 1).

Despite these findings, a systematic tendency to overestimate WTS persisted, particularly due to occasional high-value outliers that distorted the upper tail of the distribution. This suggests an inherent fault within generative models, i.e., they are unable to naturally reproduce the inertia that characterises real savings behaviour.

As synthetic users are still in an embryonic stage, exercises grounded in known human distributions are essential to assess fidelity. For transparency, the original experimental results were not used as model inputs or prompts, thereby avoiding potential bias. Instead, the behavioural libraries were constructed from a comprehensive review of the academic literature, modelling the effects of stimuli such as the calculator on WTS.

> “The behavioural libraries were constructed from a comprehensive review of the academic literature, modelling the effects of stimuli such as the calculator on WTS.”

While this approach is innovative, it requires further validation. We are therefore pursuing two next steps: replicating the study with a new sample and synthetic population to test consistency, and using known results to develop guidelines for building reliable synthetic participants when human reference data is unavailable.

Conclusion

These results support a differentiated and modular approach to building synthetic participants by separating the sociodemographic buildup (who the participant is) from the behavioural library (what cognitive principles are likely active, and how strongly). Many solutions offer quick and adaptable synthetic models that include sociodemographic data but overlook behavioural profiling and contextualisation. The limitations of this approach have been widely discussed in the literature, even before the development of AI models; demographic profiling and segmentation without a behavioural layer often paints an incomplete picture of the subjects of study, especially when gauging a behavioural outcome.

Synthetic participants are best suited to contexts where outputs will be used to inform rather than replace human judgement. They are most defensible as an augmentation tool useful for hypothesis generation, rapid prototyping and scenario testing, and they are least suited to situations involving high emotional complexity, culturally specific behaviour or underrepresented populations. Embedding behavioural economics constructs is feasible where the relevant decision environment is sufficiently understood, whilst biases can be operationalised as structured knowledge libraries that modulate how synthetic participants respond across profiles and conditions – as our experiment demonstrates. The same logic extends to adaptation over time: as economic conditions, social norms and digital behaviours evolve, static libraries will drift from reality. Periodic recalibration against human reference data is therefore not an optional refinement but a structural requirement.

Synthetic participants are increasingly useful tools for specific, well-defined tasks when their construction is grounded in empirical data, explicit psychological theory and transparent calibration procedures. The framework proposed in this article is offered as a starting point for that kind of responsible deployment: one that treats the limitations of synthetic participants as design parameters to be managed rather than as obstacles to be minimised.

> “Embedding behavioural economics constructs is feasible where the relevant decision environment is sufficiently understood, whilst biases can be operationalised as structured knowledge libraries that modulate how synthetic participants respond.”

Three practical implications stand out for practitioners. First, quantitative calibration against human benchmark distributions should be treated as an essential check, as distributional fit metrics make deviations visible before they influence decisions. Second, anchoring synthetic outputs to empirically grounded knowledge bases, through structured libraries or retrieval-augmented approaches, can reduce the idealisation bias that unconstrained generative models exhibit. Third, the question of what synthetic participants are actually good for should remain at the centre of deployment decisions. Used within aforementioned boundaries, they are a powerful addition to the behavioural researcher's toolkit; used beyond them, they carry risks that are well-documented and avoidable.

THE AUTHORS

María López is a behavioural economics manager and Head of Innovation at Neovantas. She holds a University Expert in Behavioral Economics from Behavior & Law and UDIMA, Spain. At Neovantas, María leads international projects focused on creating specialised synthetic participant models enriched with behavioural input. She also works on designing new products that integrate AI with behavioural science insights.

Paola Schietekat is a senior behavioural economics consultant at Neovantas. She holds a master's degree in Public Policy from Oxford University, where she specialised in Behavioural Science and Economic Development. At Neovantas, she works on integrating behavioural science insights into AI solutions.

Juan de Rus is the managing partner at Neovantas. He holds an MSc in Behavioural Science with Distinction from the London School of Economics & Political Science (LSE), an MBA from Universidad Carlos III de Madrid and a Master of Arts with Distinction from the University of Dundee (UK). He also serves as an adjunct professor of Behavioural Science and Marketing at IE Business School, Universidad Carlos III and the University of Barcelona.

Dario Krpan is Associate Professor of Behavioural Science at the London School of Economics and Political Science. His research examines the transformative power of behavioural science in the age of AI and new technology. In this context, he explores the intersection of behavioural science and AI, with a particular emphasis on synthetic participants as tools for simulating human behaviour and decision-making. Dario is also a senior advisor to Neovantas, where he contributes to shaping AI and data-driven strategies aimed at transforming organisations through the use of behavioural data.

REFERENCES

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), 286.* https://doi.org/10.1145/3613904.3642703

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning, 202*, 337-371. https://proceedings.mlr.press/v202/aher23a.html

Amini, A. (2025). Survey transfer learning: Recycling data with silicon responses. *arXiv*. https://arxiv.org/abs/2501.06577

Arora, N., Chakraborty, I., & Nishimura, Y. (2025). AI–human hybrids for marketing research: Leveraging large language models as collaborators. *Journal of Marketing, 89*(2), 43-70. https://doi.org/10.1177/00222429241276529

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), 337-355. https://doi.org/10.1017/pan.2023.2

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. https://doi.org/10.1017/pan.2024.5

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Cui, Z., Li, N., & Zhou, H. (2024). Can large language models replace human subjects? A large-scale replication of scenario-based experiments. *arXiv*. https://arxiv.org/abs/2409.00128

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences, 27*(7), 597-600. https://doi.org/10.1016/j.tics.2023.04.008

Doudkin, A., Pataranutaporn, P., & Maes, P. (2025a). AI persuading AI vs. AI persuading humans: Differential effectiveness in promoting pro-environmental behavior. *arXiv*. https://arxiv.org/abs/2503.02067

Doudkin, A., Pataranutaporn, P., & Maes, P. (2025b). From synthetic to human: The gap between AI-predicted and actual behavior change after chatbot persuasion. *Proceedings of the ACM Conference on Conversational User Interfaces* (CUI '25), 71. https://doi.org/10.1145/3719160.3736610

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2024). Take caution in using LLMs as human surrogates: Scylla ex machina. *arXiv*. https://arxiv.org/abs/2410.19599

González-Bustamante, B., Verelst, N., & Cisternas, C. (2025). Emulating public opinion: Synthetic survey responses for the Chilean case. *arXiv*. https://arxiv.org/abs/2509.09871

Haider, S. A., Borna, S., Trabilsy, M., Bagaria, S., & Forte, A. J. (2025). Synthetic patient–physician conversations simulated by large language models. *Sensors, 25*(14), 4305. https://doi.org/10.3390/s25144305

Lin, Z. (2025). Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science, 8*(3), 1-19. https://doi.org/10.1177/25152459251357566

Lin, Z., & Dai, Y. (2025). Fostering epistemic insights into AI ethics through a constructionist pedagogy: An interdisciplinary approach to AI literacy. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*. 29171-29177. https://doi.org/10.1609/aaai.v39i28.35190

Lu, Y., Yao, B., Gu, H., Huang, J., Wang, J., Li, Y., Gesi, J., He, Q., Li, T. J.-J., & Wang, D. (2025). UXAgent: An LLM agent-based usability testing framework. *arXiv*. https://arxiv.org/abs/2502.12561

Maier, B. F., Aslak, U., Fletcher, K., Rismal, N., Dow, R., Pappas, K., Fiaschi, L., Luhmann, C. C., & Wiecki, T. V. (2025). LLMs reproduce human purchase intent via semantic similarity elicitation. *arXiv*. https://arxiv.org/abs/2510.08338

Mansour, S., Perelli, L., Mainetti, L., Davidson, G., & D'Amato, S. (2025). PAARS: Persona-aligned agentic retail shoppers. *arXiv*. https://arxiv.org/abs/2503.24228

Market Research Society Delphi Group. (2024). *Using synthetic participants for market research: Part two of the BEST framework for generative AI.* MRS Evidence Matters. https://www.mrs.org.uk/pdf/MRS_Delphi_synthetic.pdf

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

O'Grady, C. (2024). AI-generated "participants" can lead social science experiments astray. *Science, 383*(6674), 830-832. https://www.science.org/doi/10.1126/science.aec9020

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv*. https://arxiv.org/abs/2304.03442

Qu, Y., & Wang, J. (2024). Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications, 11*, 1095. https://doi.org/10.1057/s41599-024-03609-x

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review, 46*(1), 192-210. https://doi.org/10.5465/amr.2018.0072

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proceedings of the 40th International Conference on Machine Learning, 202*, 29971-30004. https://proceedings.mlr.press/v202/santurkar23a.html

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*(7987), 493-498. https://doi.org/10.1038/s41586-023-06647-8

Sheeran, P. (2002). Intention-behavior relations: A conceptual and empirical review. *European Review of Social Psychology, 12*(1), 1-36. https://doi.org/10.1080/14792772143000003

Shrestha, P., Krpan, D., Koaik, F., Schnider, R., Sayess, D., & Binbaz, M. S. (2025). Beyond WEIRD: Can synthetic participants substitute for humans in global policy research? *Behavioral Science & Policy, 10*(2), 26-45.

Stamatogiannakis, A., Ghodsinia, A., Etminanrad, S., Gonçalves, D., & Santos, D. (2025). How human is the machine? Evidence from large-scale LLM conversations. *arXiv*. https://arxiv.org/abs/2510.07321

Suh, J., Jahanparast, E., Moon, S., Kang, M., & Chang, S. (2025). Fine-tuning language models on survey data for public opinion prediction. *arXiv*. https://arxiv.org/abs/2502.16761

Tjuatja, L., Chen, V., Wu, T., Talwalkar, A., & Neubig, G. (2024). Do LLMs exhibit human-like response biases? *Transactions of the Association for Computational Linguistics, 12*, 1011-1026. https://doi.org/10.1162/tacl_a_00685

Ustiyanovych, J., & Krpan, D. (2025). The potential of synthetic twin agents for personalized behavioural interventions at scale. https://doi.org/10.31234/osf.io/vs2mk_v5

Wang, Z., Lu, Y., Li, W., Sun, B., Bart, Y., Lyu, W., Gesi, J., Wang, T., Huang, J., Su, Y., Ehsan, U., Alikhani, M., Li, T. J.-J., Chilton, L., & Wang, D. (2025). See, think, act: Simulating online shopper behavior with multimodal agents. *arXiv*. https://arxiv.org/abs/2501.06577

Zhong, R., McDonald, D. W., & Hsieh, G. (2024). Synthetic heuristic evaluation: A comparison between AI- and human-powered usability evaluation. *arXiv*. https://doi.org/10.48550/arXiv.2507.02306