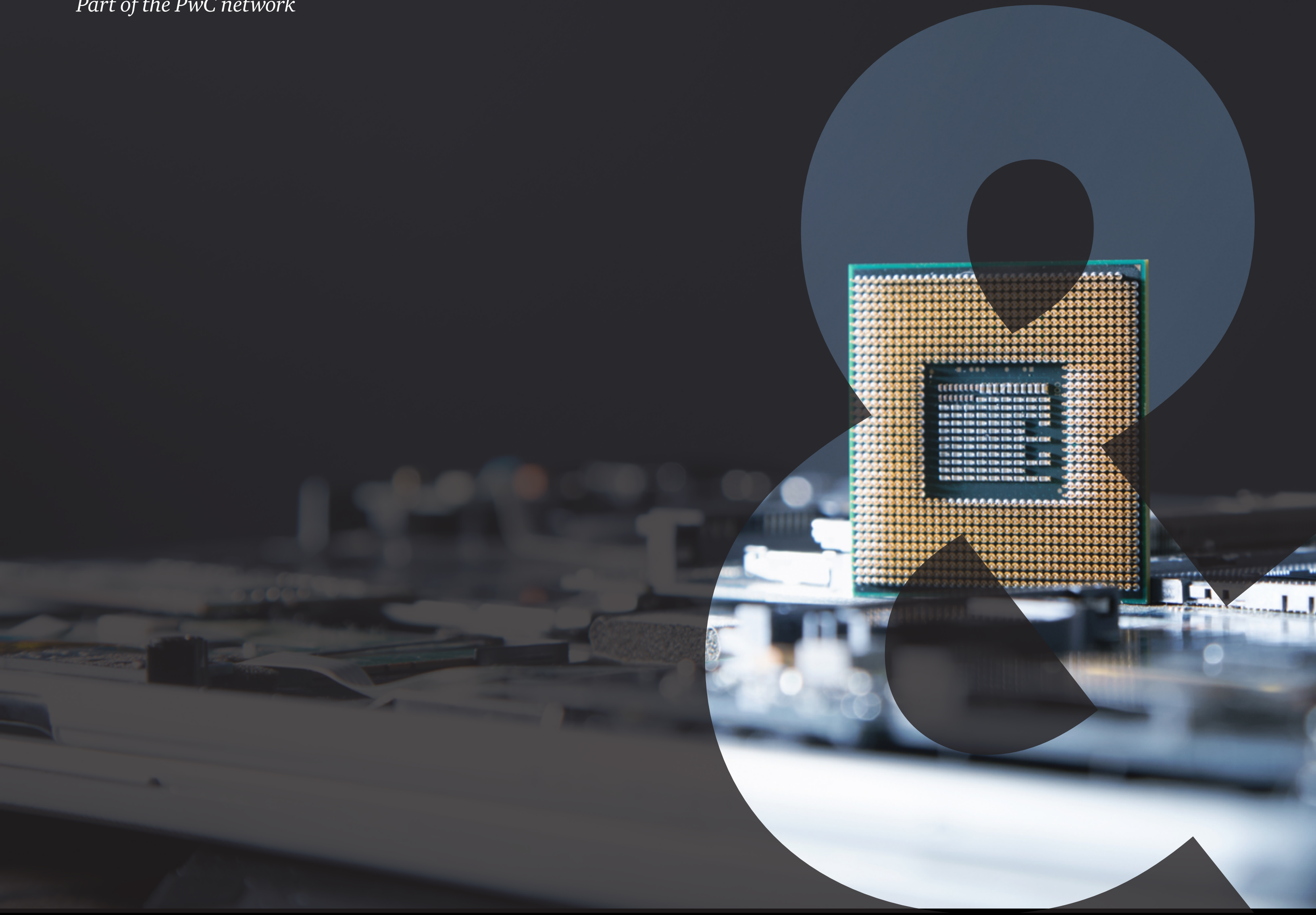
strategy&
Part of the PwC network



IDEATION CENTER INSIGHT

Generative AI to augment behavioral research—evidence from the Middle East

Contacts

Dubai

Dima Sayess
Partner, director of the
Ideation Center
+971-4-436-3000
dima.sayess
@strategyand.pwc.com

Dr. Fatima Koaik
Behavioral economics director
+971-4-436-3000
fatima.koaik
@strategyand.pwc.com

ABOUT THE AUTHORS

Dima Sayess is a partner with Strategy& Middle East, part of the PwC network, and the director of the Ideation Center, the leading think tank for Strategy& in the Middle East. She has more than 20 years of experience in public-sector consulting in the region, including socioeconomic development, quality of life, government of the future, and innovation in policymaking. She was formerly an advisor to the minister of finance and education in Lebanon, and the Executive Council of Dubai.

Dr. Fatima Koaik is the behavioral economics director at the Ideation Center. She works at the intersection of applied research and consulting projects in the region. She completed her Ph.D. in the Psychology and Behavioral Science Department at London School of Economics and Political Science (LSE), has led the establishment of several behavioral units, and has conducted more than 50 experiments. Her previous experience includes working as a behavioral scientist at the World Bank's Mind, Behavior and Development Unit (eMBeD) and at the United Nations Development Programme (UNDP).

Pujen Shrestha is a behavioral and data scientist at the Ideation Center, where he focuses on applying behavioral science to address policy challenges in the Middle East. His work includes designing and implementing behavioral experiments related to entrepreneurship, labor participation, and consumer behavior. Previously, he was a behavioral science researcher at the Behavioral Insights Team in the United Kingdom.

Dr. Robin Schnider is a senior behavioral economist at the Ideation Center. As part of the behavioral economics team, he applies behavioral insights in order to conduct experiments and evidence-based research advising clients and policymakers on strategic challenges. He previously worked as a postdoctoral researcher at the University of Zurich, where he earned his Ph.D. in management and economics, applying methods of experimental economics to various topics such as corporate social responsibility, legitimacy, and pro-social behavior.

Raneem Alturki is a behavioral scientist at the Ideation Center, where she applies behavioral insights to various policy and financial challenges. She holds a master's degree in behavioral finance from the University of Reading. Previously, she was part of the Behavioral Science Unit at the Center of Government in Riyadh, Saudi Arabia, where she conducted interdisciplinary research in public policy, financial well-being, and environmental studies.

May Saad Bin Baz has over 18 years of experience across public and private entities. She works at the intersection of behavioral science and economics, spanning research and consultancy, with a focus on behavioral change, decision-making, and improving well-being for individuals, organizations, and government entities. With an MBA from Al Faisal University and an M.Sc. in behavioral science from the LSE, she has expertise in translating evidence-based insights into actionable strategies for sustainable outcomes and impact.

Dr. Dario Krpan is an assistant professor of behavioral science in the Department of Psychological and Behavioral Science at the LSE and is affiliated with the LSE's Data Science Institute. He completed his academic training at the University of Cambridge, earning both an M.Phil. and a Ph.D. in psychology. His primary research focus is on transformative behavioral change—how individuals can adapt their lifestyles to address some of the world's most pressing challenges, including climate change and the rise of AI.

This report is an adapted version of a scientific paper first published in *Behavioral Science & Policy* (SAGE Publishing), titled "Beyond WEIRD: Can Synthetic Survey Participants Substitute for Humans in Global Policy Research?"

The Ideation Center

The Ideation Center is the leading think tank for Strategy& Middle East, part of the PwC network. We aim to promote sustainable growth in the region by helping leaders across sectors translate socioeconomic trends into actions and better business decisions. Combining innovative research, analysis, and dialogue with hands-on expertise from the professional community in the private and public sectors, the Ideation Center delivers impactful ideas through our publications, website, and forums. The end result is one that inspires, enriches, and rewards. The Ideation Center upholds Strategy&'s mission to develop practical strategies and turn ideas into action. At the Ideation Center, we enjoy the full support of all practices in the Middle East. Together we bring unsurpassed commitment to the goal of advancing the interests of the Middle East region. Find out more by visiting www.ideationcenter.com.

EXECUTIVE SUMMARY

Behavioral researchers have begun to explore whether large language models (LLMs) such as Open AI's GPT (which stands for *generative pre-trained transformer*) can be used to create "synthetic" research participants—artificial agents that can respond to surveys in a manner similar to that of humans. Studies have found that such synthetic participants can indeed mimic human decisions and respond much like their human counterparts, even replicating previous research findings. This raises the question: Could artificial intelligence (AI) models replace humans in testing behavioral policy interventions?

To date, research has focused primarily on Western countries, with limited participation from the Middle East and North Africa (MENA) region. To study the accuracy of synthetic participants across contexts, we examined the similarity between human and synthetic participants from samples in three countries—Saudi Arabia, the United Arab Emirates (UAE), and the U.S.—in three policy domains: sustainability, financial literacy, and female labor force participation. Across these domains, we assessed attitudes about policies and measured the impact of several interventions on self-reported behaviors from both human and synthetic participants.

In summary, we found the synthetic participants created by GPT produced responses similar to those of their human counterparts across the three policy domains we assessed. However, the effects of the behavioral interventions we tested varied between human and synthetic participants. We also observed two primary differences in Saudi Arabian and UAE responses compared with those from the United States. First, the correlations were stronger for U.S. participants—when human responses in the U.S. increased or decreased, synthetic responses reflected them more closely. Second, for the U.S., GPT exhibited higher levels of positive bias (overestimating human participants' support for various policy proposals), and for Saudi Arabia and the UAE exhibited higher levels of negative bias (underestimating participants' support). This report highlights the main policy implications of these findings and makes practical recommendations for researchers.

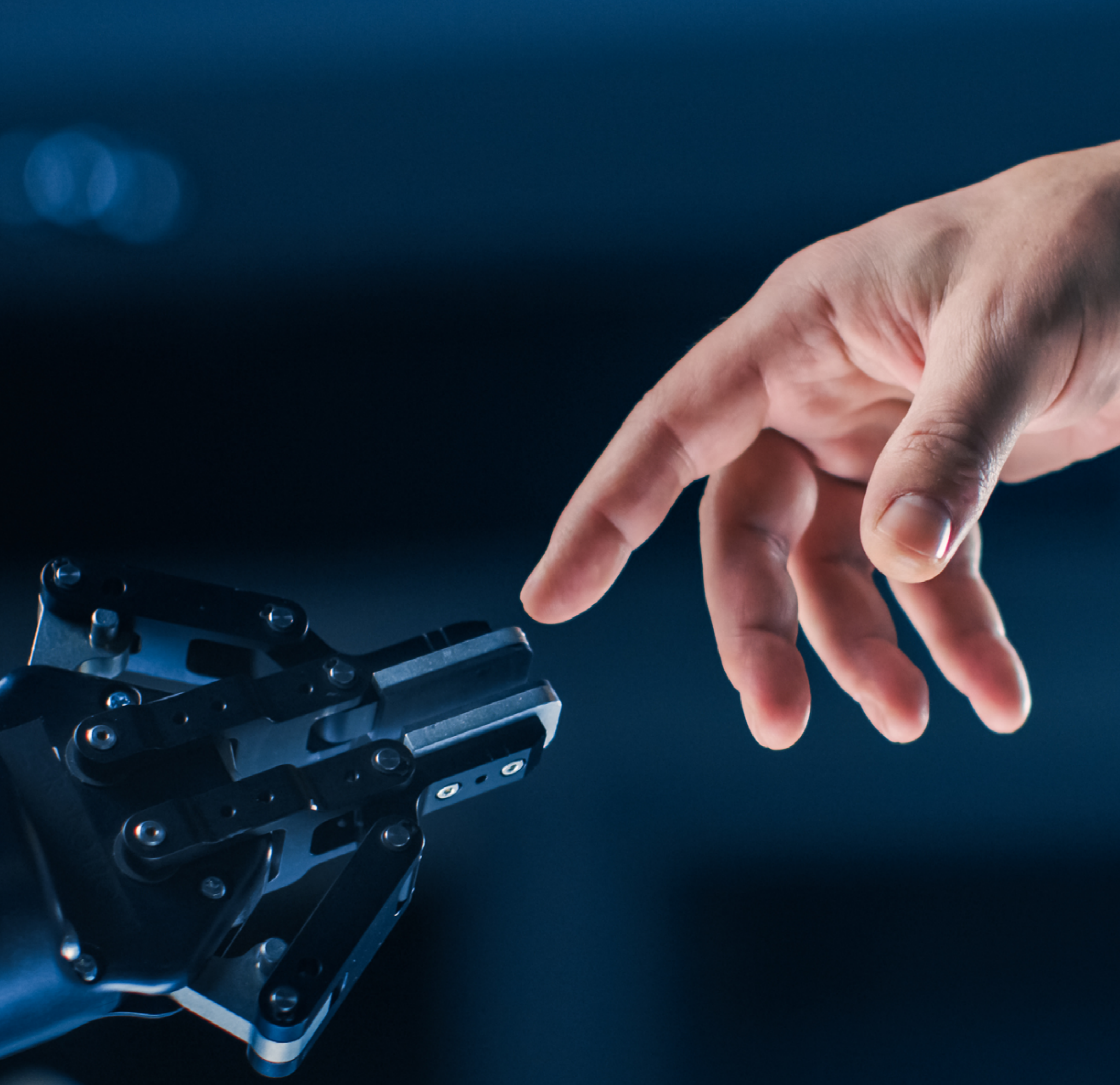
CHALLENGE

ChatGPT's launch on November 30, 2022,[1] caused a huge spike in interest among stakeholders around the world about how artificial intelligence (AI) could reduce the burden of labor-intensive tasks on the workforce. It has been estimated that AI's impact on the global economy will reach US$15.7 trillion by 2030.[2] In the Gulf Cooperation Council (GCC) countries,[3] Saudi Arabia announced a plan to create a $40 billion fund dedicated to AI investments.[4]

Within behavioral science, researchers are exploring whether large language models (LLMs)[5] can mimic humans. Significant discussion has taken place about whether synthetic participants—artificial agents that can respond to surveys much as humans do—could replace humans in domains where assessing public opinion is crucial.[6,7]

Although the application of AI to behavioral science could be transformational, questions remain. First, research has primarily examined whether synthetic participants can replicate previous experimental results and exhibit traits and values akin to those of human participants.[8,9] Thus far, the use of synthetic participants to generate new policy insights has been overlooked. Second, research has focused on non-MENA populations, with limited data and insights relevant to countries such as Saudi Arabia and the UAE.[10,11] In the MENA region, recruiting human participants for policy research is often challenging due to underrepresentation of diverse local demographics on popular recruitment platforms and higher costs associated with specialized recruitment agencies. Therefore, addressing the accuracy of LLMs such as OpenAI's GPT-4 in MENA contexts is critical; synthetic participants could provide invaluable insights into local policy issues and help bridge the gap in research output.

This report examines the use of synthetic participants in relevant regional policy challenges. It highlights opportunities and challenges with AI's use in behavioral science and provides evidence-based guidance for public policy in the region.



> It has been estimated that AI’s impact on the global economy will reach US$15.7 trillion by 2030. In the Gulf Cooperation Council (GCC) countries, Saudi Arabia announced a plan to create a $40 billion fund dedicated to AI investments.

CONTEXT

Synthetic participants are modeled after humans with specific characteristics.[12,13] For example, a human participant could be a 30-year old female from Saudi Arabia with a master's degree who is employed, married, and extroverted. The corresponding synthetic participant would be created by instructing GPT to answer survey questions from the perspective of a person with these characteristics.[14,15] Therefore, synthetic participants in essence are programmed for GPT or other LLMs to mimic the responses of humans with different profiles.

Research on synthetic participants has typically investigated whether they exhibit psychological processes similar to those of human participants and can replicate previous research findings. For example, synthetic participants have been able to demonstrate moral judgments that mirror those of humans[16] and even display similar "big five" personality traits—openness (curiosity and creativity), conscientiousness (organization and reliability), extroversion (sociability and assertiveness), agreeableness (cooperativeness and empathy), and neuroticism (emotional instability).[17]

Previous research has often overlooked attitudes about public policy, including people's views on societal challenges and the actions that could be taken to address them. One study found the opinions of 60 U.S. demographic groups on topics as diverse as abortion and automation were misaligned with the opinions of their synthetic counterparts,[18] whereas another study revealed alignment of voting intentions and political views.[19] An additional oversight is the lack of research on non-Western respondents; the few studies conducted found weaker correlations between synthetic and human participants from these regions compared with the United States.[20] This pattern was evident in the World Values Survey (WVS), which measures values and beliefs about topics such as gender equality and attitudes toward work.[21]

Overall, although previous research suggests synthetic participants may resemble humans and offer policy insights, gaining a further understanding of generative AI's potential is critical to exploring outcomes related to policies and extending research to populations in the MENA region.

> “Significant discussion has taken place about whether synthetic participants—artificial agents that can respond to surveys much as humans do—could replace humans in domains where assessing public opinion is crucial.

Design

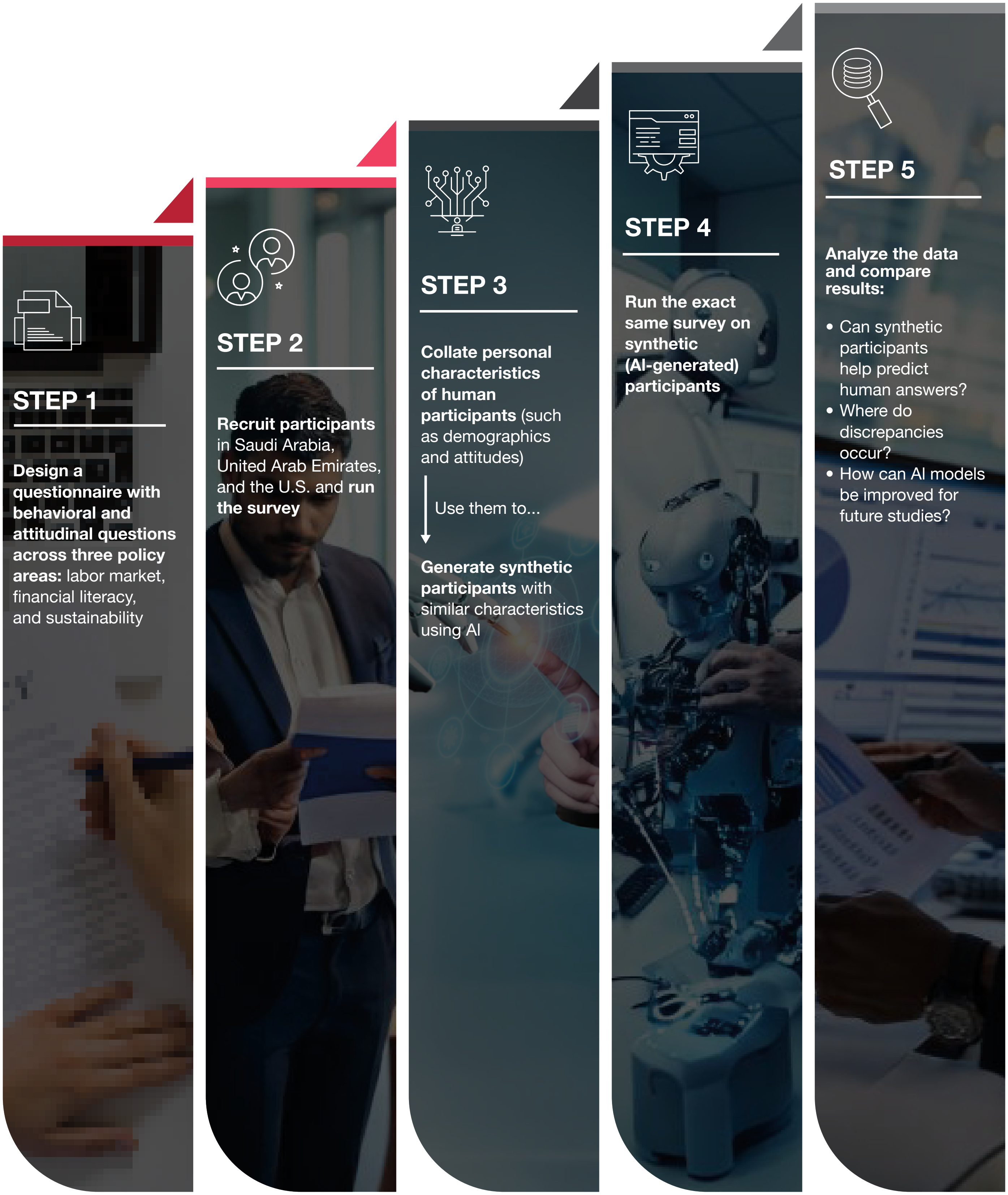
We created synthetic agents of human participants from Saudi Arabia, the UAE, and the U.S. using a variety of demographic and psychological traits (such as age, gender, nationality, employment status, and educational attainment). We exposed both human and synthetic participants to questions related to three policy domains: sustainability, financial literacy, and female labor force participation. For each domain, participants received one of two types of questions. “Behavioral” questions focused on hypothetical scenarios and asked how respondents or the described characters would act (such as offsetting carbon emissions through donation, saving versus investing versus spending money, and returning to work after having children). Meanwhile, “attitudinal” questions measured views on various corresponding issues (for example, actions designed to protect the environment, plan for one’s financial future, and promote gender equality in the workplace).

This design (*see Exhibit 1*) allowed us to explore the following questions:

- Are synthetic participants **able to predict human responses to attitudinal questions,** and does AI’s ability to replicate responses differ by country?
- Are synthetic participants **able to predict human reactions to interventions,** and does AI’s ability to predict the reactions differ by country?

EXHIBIT 1

Study design for using AI in behavioral research



STEP 1

Design a questionnaire with behavioral and attitudinal questions across three policy areas: labor market, financial literacy, and sustainability

STEP 2

Recruit participants in Saudi Arabia, United Arab Emirates, and the U.S. and **run the survey**

STEP 3

Collate personal characteristics of human participants (such as demographics and attitudes)

Use them to...

Generate synthetic participants with similar characteristics using AI

STEP 4

Run the exact same survey on synthetic (AI-generated) participants

STEP 5

Analyze the data and compare results:

- Can synthetic participants help predict human answers?
- Where do discrepancies occur?
- How can AI models be improved for future studies?

Results

The results are divided into four components: correlation, precision, bias, and the findings of the behavioral experiments.

1. **Correlation**

   Correlation refers to the degree to which the responses from human and synthetic participants moved in the same direction. For this research, we computed aggregate correlations, which represent the similarity between average human and synthetic responses across all 43 variables we assessed. A strong correlation (for example, $r \geq .50$)[22] would indicate that when human responses increased or decreased, synthetic responses did so as well. Our results found that the correlations between the human and synthetic responses were indeed strong for all three samples (Saudi Arabia, the UAE, and the U.S.). On average, human and synthetic responses strongly covaried, which means they increased or decreased in similar ways. Nevertheless, the correlations for the U.S. sample were consistently highest (r = .86), followed by those of the UAE (r = .75) and those of Saudi Arabia (r = .65).

2. **Precision**

   Precision refers to how closely the average responses of the synthetic participants matched those of the human participants. Responses of both types of participants broadly moved in the same direction, but they did not match with a high degree of accuracy, indicating a medium level of precision. For example, if synthetic participants had positive attitudes regarding questions on a selected topic, human participants also generally exhibited positive attitudes. Therefore, GPT can guess the direction of human responses, but its precision could be improved in estimating the exact mean values of human responses.

3. **Bias**

   The bias result captures the degree to which GPT tended to overestimate (positive bias) or underestimate (negative bias) the support of human participants for various policy proposals. GPT was more likely to exhibit positive bias and less likely to indicate negative bias for the U.S. sample compared with the Saudi Arabia and UAE samples. This pattern was particularly evident for sustainable attitudes and behaviors, followed by financial literacy. For example, in the U.S. sample, synthetic participants were generally more supportive of sustainability-related issues, such as being more willing to pay higher prices for goods and services in order to protect the environment, compared with their human counterparts. This trend was reversed in the Saudi and UAE samples, where human participants showed greater support than their synthetic counterparts.

4. **Findings of the behavioral experiments**

   These findings refer to how reliably GPT could predict the impact of our interventions on the self-reported behaviors we examined. The results showed the interventions generally yielded similar effect sizes (small, medium, or large) for human and synthetic participants, which indicates GPT was able to estimate the strength of intervention effects. It could not, however, always predict whether an intervention would have a statistically significant influence on self-reported behavior. Although GPT was able to estimate how large the effects of the interventions would be with reasonable accuracy, it could not consistently identify which interventions might significantly affect the behaviors.

POLICY RECOMMENDATIONS

Our analysis identified four core recommendations for policymakers and practitioners in the MENA region.

1. **Use GPT in preliminary testing of views of policies and piloting interventions in Saudi Arabia and the UAE.**

   Human and synthetic participants were reasonably well aligned, as their responses strongly covaried and generally moved in the same direction. In addition, the effects of our experimental interventions did not differ significantly between the two types of participants. Therefore, in the initial stage of policy development and testing, GPT synthetic participants can serve as a good approximation of human participants

2. **Use human participants in more advanced stages of policy development and testing.**

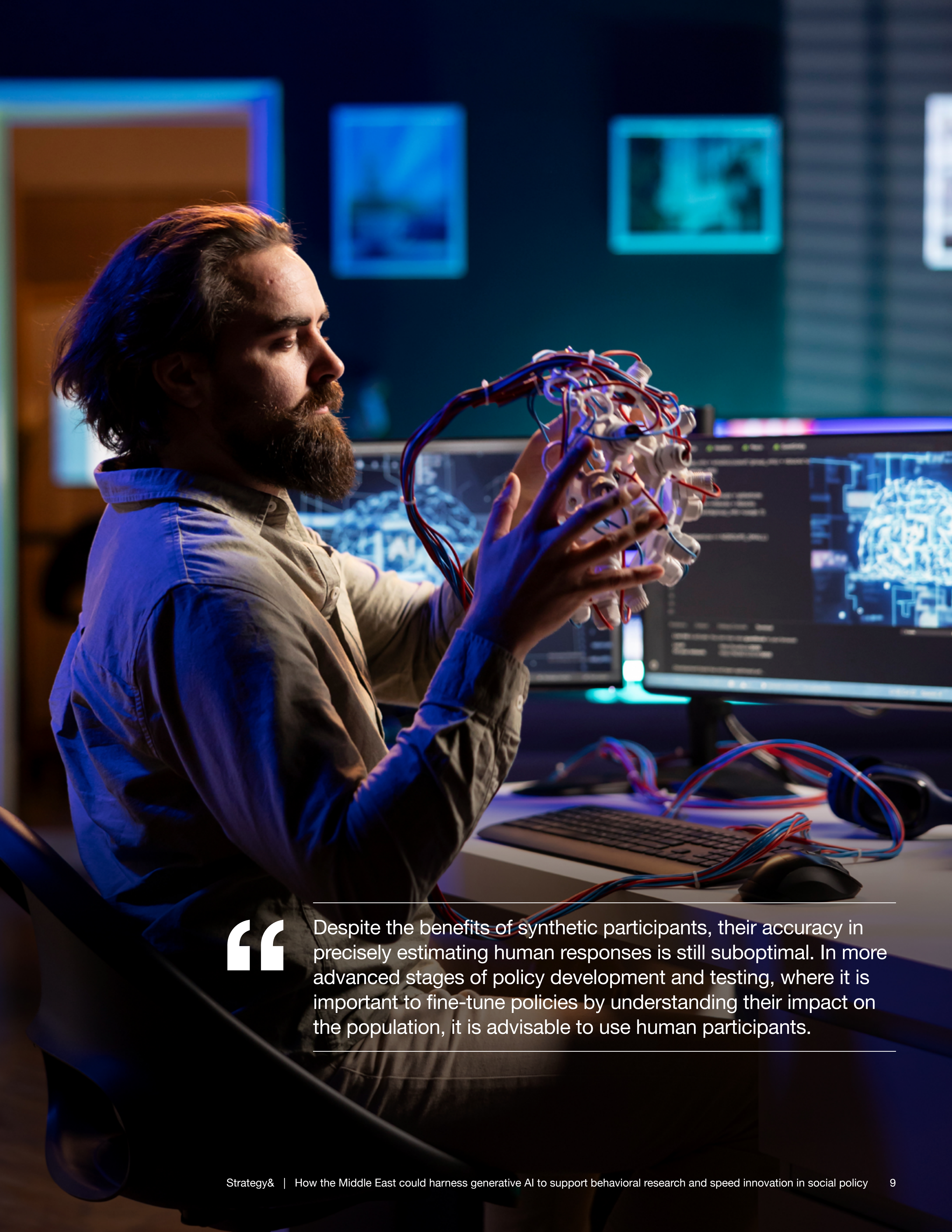
   Despite the benefits of synthetic participants, their accuracy in precisely estimating human responses is still suboptimal. In more advanced stages of policy development and testing, where it is important to fine-tune policies by understanding their impact on the population, it is advisable to use human participants.

3. **When using synthetic participants in policy research, be mindful of potential biases.**

   Biases are a concern with synthetic participants. In sustainability, for example, GPT's synthetic participants are likely to produce more progressive responses compared with those of humans for the U.S. sample, whereas this pattern is the opposite for the Saudi Arabia sample.

4. **When creating synthetic participants, use the simplest prompting strategy for optimal results.**

   A straightforward approach to creating synthetic participants (for example, providing GPT-4 with basic demographic traits such as age, gender, and employment status and instructing it to generate responses based on those traits) will produce findings that are either comparable to, or sometimes more accurate than, the ones from more advanced prompting (such as going beyond basic demographics to provide GPT with additional traits). Overall, our research shows synthetic participants can be used in policy research even when only basic demographic data is available.



> Despite the benefits of synthetic participants, their accuracy in precisely estimating human responses is still suboptimal. In more advanced stages of policy development and testing, where it is important to fine-tune policies by understanding their impact on the population, it is advisable to use human participants.

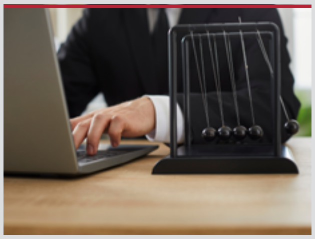
ADVANCING BEHAVIORAL RESEARCH IN THE GCC REGION USING GENERATIVE AI

Although we are excited about the potential of generative AI in behavioral research and policy development, policymakers and practitioners need to be aware of and address biases in GPT, particularly its negative bias in Saudi Arabia and the UAE (in particular, for sustainability) as well as its consistent positive bias in the U.S. across all policy domains we tested.

Our results highlight the importance of recognizing that tools like OpenAI's GPT-4 application programming interface (API) may offer default, or out-of-the-box, solutions that might not be suitable for all contexts. The development of tools that rely on pretrained models may lead to unintended consequences for end-users, particularly those residing in certain regions. For example, a chatbot designed to help users engage in sustainable practices may recommend less environmentally responsible behavior if the model believes it is interacting with someone living in Saudi Arabia or the UAE. Researchers and practitioners should therefore explore ways to recalibrate LLMs to account for these biases and consider withholding non-Western contexts (such as the country, or region, of a generated synthetic participant) in prompts to achieve potentially more universal results.

Two opportunities emerge for researchers and practitioners in the region:

- **Exploring GPT's biases in additional areas and contexts:** On the basis of our findings, it seems likely that GPT responses could demonstrate bias across other areas and contexts. Additional policy-relevant areas and contexts should be explored beyond the three topics covered by our study, and the origin of these biases should be examined further. GPT's biases might reveal stereotypes within existing training data that could still be prevalent among the creators or consumers of this data (including Western populations). If so, GPT's cultural bias could be used as a tool to detect common stereotypical beliefs in Western contexts toward non-Western countries.



- **Training models based on data from the MENA region:** GPT's negative bias displayed for Saudi Arabia and the UAE, as well as its limited awareness with respect to the behaviors and policy issues prevalent in non-Western populations (for example, financial literacy challenges and low savings rates in GCC countries) highlight the need for training LLMs on regional data. Thus, committing funding and conducting research into the development of MENA-centric LLMs seem necessary if behavioral scientists are to consider applying GPT-like models to the region's specific policy challenges.

CONCLUSION

Generative AI has sparked the imagination of policymakers and the research community for its potential to dramatically accelerate research into public policy. Our research found GPT could be useful in gauging the public's reaction to prospective policies, but it is still premature to consider using it in more advanced stages of policy development or the testing of behavioral interventions. Further advances are needed to precisely estimate human responses and remove biases against GCC populations. Understanding generative AI's promise and limitations will be crucial to unlocking its full power in the future.

ACKNOWLEDGMENTS

If you would like to learn more about the study this report is based on, please visit **https://journals.sagepub.com/doi/10.1177/23794607241311793** to find the peer-reviewed paper published in the journal *Behavioral Science & Policy*. The full paper was developed by the Strategy& Ideation Center's Behavioral Science Unit in partnership with experts at the London School of Economics and Political Science and behavioral science practitioners from Saudi Arabia. This study was fully funded by Strategy&'s Ideation Center. To learn more about the efforts of the Ideation Center in promoting evidence-based public policy in the MENA region, please visit the website at **ideationcenter.com.**

ENDNOTES

1. OpenAI, “Introducing ChatGPT,” November 30, 2022 (https://openai.com/blog/chatgpt).

2. PwC, “Sizing the Prize: What’s the Real Value of AI for Your Business and How Can You Capitalise?” 2017 (https://www.pwc.com/gx/en/issues/artificial-intelligence/publications/artificial-intelligence-study.html).

3. The GCC countries are Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates.

4. Maureen Farrell and Rob Copeland, “Saudi Arabia Plans $40 Billion Push into Artificial Intelligence,” *New York Times*, March 19, 2024 (https://www.nytimes.com/2024/03/19/business/saudi-arabia-investment-artificial-intelligence.html).

5. LLMs are advanced AI systems trained on vast amounts of text data to understand and generate human-like language. Examples include OpenAI’s GPT, Google’s Bard, and Meta’s LLaMA.

6. D. Dillion, N. Tandon, Y. Gu, and K. Gray, “Can AI Language Models Replace Human Participants?” *Trends in Cognitive Sciences*, Vol. 27, No. 7, 2023 (https://doi.org/10.1016/j.tics.2023.04.008).

7. M. Hutson, “Guinea Pigbots: Doing Research with Human Subjects Is Costly and Cumbersome. Can AI Chatbots Replace Them?” *Science*, Vol. 381, No. 6654, 2023 (https://doi.org/doi: 10.1126/science.adj7014).

8. D. Dillion, N. Tandon, Y. Gu, and K. Gray, “Can AI Language Models Replace Human Participants?” *Trends in Cognitive Sciences*, Vol. 27, No. 7, 2023 (https://doi.org/10.1016/j.tics.2023.04.008).

9. P.S. Park, P. Schoenegger, and C. Zhu, “Diminished Diversity-of-Thought in a Standard Large Language Model,” *Behavior Research Methods*, 2024 (https://doi.org/10.3758/s13428-023-02307-x).

10. M. Atari, M.J. Xue, P.S. Park, D. Blasi, and J. Henrich, “Which Humans?” *PsyArXiv*, 2023 (https://doi.org/10.31234/osf.io/5b26t).

11. L.P. Argyle, E.C. Busby, N. Fulda, J.R. Gubler, C. Rytting, and D. Wingate, “Out of One, Many: Using Language Models to Simulate Human Samples,” *Political Analysis*, Vol. 31, No. 3, 2023 (https://doi.org/10.1017/pan.2023.2).

12. D. Dillion, N. Tandon, Y. Gu, and K. Gray, “Can AI Language Models Replace Human Participants?” *Trends in Cognitive Sciences*, Vol. 27, No. 7, 2023 (https://doi.org/10.1016/j.tics.2023.04.008).

13. P.S. Park, P. Schoenegger, and C. Zhu, “Diminished Diversity-of-Thought in a Standard Large Language Model,” *Behavior Research Methods*, 2024 (https://doi.org/10.3758/s13428-023-02307-x).

14. L.P. Argyle, E.C. Busby, N. Fulda, J.R. Gubler, C. Rytting, and D. Wingate, “Out of One, Many: Using Language Models to Simulate Human Samples,” *Political Analysis*, Vol. 31, No. 3, 2023 (https://doi.org/10.1017/pan.2023.2).

15. J. de Winter, T. Driessen, and D. Dodou,“The Use of ChatGPT for Personality Research: Administering Questionnaires Using Generated Personas,” *Personality and Individual Differences*, 2024 (https://www.researchgate.net/publication/374415968_The_use_of_ChatGPT_for_personality_research_Administering_questionnaires_using_generated_personas).

16. D. Dillion, N. Tandon, Y. Gu, and K. Gray, “Can AI Language Models Replace Human Participants?” *Trends in Cognitive Sciences*, Vol. 27, No. 7, 2023 (https://doi.org/10.1016/j.tics.2023.04.008).

17. J. de Winter, T. Driessen, and D. Dodou, “The Use of ChatGPT for Personality Research: Administering Questionnaires Using Generated Personas,” *Personality and Individual Differences*, 2024 (https://www.researchgate.net/publication/374415968_The_use_of_ChatGPT_for_personality_research_Administering_questionnaires_using_generated_personas).

18. S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, “Whose Opinions Do Language Models Reflect? Proceedings of the 40th International Conference on Machine Learning,” 2023 (https://proceedings.mlr.press/v202/santurkar23a.html).

19. L.P. Argyle, E.C. Busby, N. Fulda, J.R. Gubler, C. Rytting, and D. Wingate, “Out of One, Many: Using Language Models to Simulate Human Samples,” *Political Analysis*, Vol. 31, No. 3, 2023 (https://doi.org/10.1017/pan.2023.2).

20. M. Atari, M.J. Xue, P.S. Park, D. Blasi, and J. Henrich, “Which Humans?” *PsyArXiv*, 2023 (https://doi.org/10.31234/osf.io/5b26t).

21. C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen, “World Values Survey: Round Seven—Country-Pooled Datafile,” 2020, Madrid, Spain & Vienna, Austria: JD Systems Institute and WVSA Secretariat, 7, 2021 (https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp).

22. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*, 1988. Lawrence Earlbaum Associates.

Strategy&

Strategy& is a global strategy consulting business uniquely positioned to help deliver your best future: one that is built on differentiation from the inside out and tailored exactly to you. As part of PwC, every day we're building the winning systems that are at the heart of growth. We combine our powerful foresight with this tangible know-how, technology, and scale to help you create a better, more transformative strategy from day one.

As the only at-scale strategy business that's part of a global professional services network, we embed our strategy capabilities with frontline teams across PwC to show you where you need to go, the choices you'll need to make to get there, and how to get it right.

The result is an authentic strategy process powerful enough to capture possibility, while pragmatic enough to ensure effective delivery. It's the strategy that gets an organization through the changes of today and drives results that redefine tomorrow. It's the strategy that turns vision into reality. It's strategy, made real.

Read the latest Ideation Center insights

ideationcenter.com

Connect with Strategy& Middle East

twitter.com/strategyandme

linkedin.com/company/strategyandme

strategyand.pwc.com/me

Connect with Strategy&

twitter.com/strategyand

linkedin.com/company/strategyand

youtube.com/user/strategyand

strategy&

Part of the PwC network

www.strategyand.pwc.com/me

Disclaimer: This content is for general purposes only, and should not be used as a substitute for consultation with professional advisors.