A MEGA-STUDY OF DIGITAL TWINS REVEALS STRENGTHS, WEAKNESSES AND OPPORTUNITIES FOR FURTHER IMPROVEMENT

Tiany Peng, Columbia Business School, tianyi.peng@columbia.edu

George Gui, Columbia Business School, zg2467@gsb.columbia.edu

Daniel J. Merlau, Columbia Business School, djm2199@columbia.edu

Grace Jiarui Fan, Columbia Business School, jfan29@gsb.columbia.edu

Malek Ben Sliman, Columbia Business School, mab2343@mail.gsb.columbia.edu

Melanie Brucks, Columbia Business School, mb4598@gsb.columbia.edu

Eric J. Johnson, Columbia Business School, ejj3@gsb.columbia.edu

Vicki Morwitz, Columbia Business School, vgm2113@gsb.columbia.edu

Abdullah Althenayyan, Columbia Business School, AAlthenayyan27@gsb.columbia.edu

Silvia Bellezza, Columbia Business School, sbellezza@gsb.columbia.edu

Dante Donati, Columbia Business School, dd3137@gsb.columbia.edu

Hortense Fong, Columbia Business School, hf2462@gsb.columbia.edu

Elizabeth Friedman, Columbia Business School, emf2198@gsb.columbia.edu

Ariana Guevara, Barnard College, arianaregguevara@gmail.com

Mohamed Hussein, Columbia Business School, mh4547@columbia.edu

Kinshuk Jerath, Columbia Business School, kj2323@gsb.columbia.edu

Bruce Kogut, Columbia Business School, bk2263@gsb.columbia.edu

Kristen Lane, Columbia Business School, kel2167@gsb.columbia.edu

Hannah Li, Columbia Business School, hannah.li@columbia.edu

Patryk Perkowski, Yeshiva University, patryk.perkowski@yu.edu

Oded Netzer, Columbia Business School, onetzer@gsb.columbia.edu

Olivier Toubia, Columbia Business School, ot2107@gsb.columbia.edu

ABSTRACT

Do ‘digital twins’ capture individual responses in surveys and experiments? We run 19 pre-registered studies on a national U.S. panel and their digital twins (constructed based on previously-collected extensive individual-level data) and compare twin and human answers across 164 outcomes. The correlation between twin and human answers is modest (approximately 0.2 on average) and twin responses are less variable than human responses. While constructing digital twins based on rich individual-level data improves our ability to capture heterogeneity across participants and predict relative differences between them, it does not substantially improve our ability to predict the exact answers given by specific participants or enhance predictions of population means. Twin performance varies by domain and is higher among more educated, higher-income, and ideologically moderate participants. These results suggest current digital twins can capture some degree of relative differences but are unreliable for individual-level predictions and sample mean and variance estimation, underscoring the need for careful validation before use. Our data and code are publicly available for researchers and practitioners interested in optimizing digital twin pipelines.

INTRODUCTION

Accurately measuring and predicting people's opinions, preferences and reactions to interventions is critical to public policy, development in the social sciences, and encouraging the successful adoption of innovations. Surveys, field experiments and new product testing requires days if not months, and per-respondent costs are increasing as response rates decline.

As a result, there is considerable interest among academics and practitioners in leveraging large language models (LLMs) to simulate human responses (Argyle et al., 2023; Dillion et al, 2023). One particularly promising approach relies on creating silicon samples of digital twins of humans based on extensive individual-level data. If each twin is able to simulate the behavior of its human counterpart, digital twins would offer "faster" (e.g., hours rather than weeks), "cheaper" (e.g., a few dollars rather than hundreds or thousands of dollars) and even potentially "better" (e.g., no respondent fatigue, within-subject experiments in which the response to each stimulus is not influenced by previous stimuli) data for researchers and managers, enabling more rapid and robust development of knowledge, products, and content, among others.

Other research has tested more generally the validity of leveraging LLMs to simulate social science surveys and experiments, without specifically creating digital twins based on real individual-level data. When it comes to opinion surveys, several studies have shown that LLMs often struggle to express opinions that are consistent with those of human populations (e.g., Santurkar et al., 2023; Motoki et al., 2024). When it comes to replicating experimental results, Ashokkumar et al. (2024) attempted to replicate treatment effects found in published and unpublished studies. While they found a high (r=0.85) correlation between predicted and simulated treatment effects, they also found that raw predictions derived from GPT-4 systematically overestimated actual effects, leading to a large root mean squared error. Cui, Li and Zhou (2025) found that LLMs were able to replicate about three quarters of the main effects they tested (replication was lower on interaction effects); however, LLMs tended to overestimate effect sizes. Other research has developed LLMs dedicated to predicting and capturing human cognition. For example, Binz et al. (2025) and Kolluri et al. (2025) developed LLMs tailored for human behavior predictions, by fine-tuning open-source base LLMs with extensive datasets covering a wide range of experiments. Such effort is complementary to the creation of digital twins. Models like Centaur, Socrates-Llama or Socrates-Qwen trained on large, cross-sectional datasets (i.e., participants are not identified across experiments), may serve as base models for the creation of digital twins. That is, they can be combined with additional panel data where participants are tracked across multiple experiments or studies, allowing the creation of twins associated with specific individuals.[1]

The creation of a panel of digital twins of real individuals offers several potential benefits over silicon samples based on hypothetical personas. First, to the extent that digital twins are able to accurately mimic the behavior of their human counterparts, one may expect digital twins to reduce or eliminate systematic biases found when simulating

[1] Our efforts to use these models as a base model for digital twins have not been successful so far.

human responses using more generic, hypothetical personas. Second, creating a panel that maps directly onto a human panel should, in principle, help capture heterogeneity in the target population. Third, creating AI agents designed to simulate the behavior of specific individuals unlocks multiple use cases, from agentic AI (e.g., having a digital twin negotiate, work, or date on your behalf) to more novel applications such as allowing individuals to understand themselves better by interacting with digital copies of themselves.

Despite this potential, little is known about the validity of digital twins. Park et al. (2024) created digital twins of over 1,000 individuals based on proprietary interview transcripts and reported encouraging results (relative accuracy of 85% on the General Social Survey, based on the ratio of digital twin accuracy to test-retest accuracy). Toubia et al. (2025) created a publicly available dataset, Twin-2K-500, which permits the creation of digital twins of over 2,000 individuals based on their answers to over 500 questions. While they reported relatively high accuracy of digital twins on holdout data (average accuracy of 72%, relative accuracy of 88% based on the ratio of digital twin accuracy to test-retest accuracy), they found that digital twins replicated only about half of the between- and within-subject effects they tested.

In addition to ambiguity on the overall performance of digital twins, there is a lack of clarity on *how* digital twins should be built (e.g., the input that should be used and the specific ways in which twins should be programmed) and on how the accuracy of digital twins varies across behavioral domains and types of human participants. Answering these questions in a transparent and replicable manner calls for the creation of large, publicly available datasets that cover a wide range of domains.

We fill these gaps by performing a mega-study consisting of 19 pre-registered (https://researchbox.org/4145) sub-studies run on the Twin-2K-500 human sample and their digital twins. Each digital twin is created based on answers provided by its human counterpart to over 500 questions (approximately 128K characters) covering demographics, personality traits, cognitive abilities, economic preferences, as well as responses to heuristics and biases experiments and a pricing study (see Figure 1 for an overview and SI for detail).

Our 19 sub-studies were developed by a large group of co-authors with various backgrounds and interests. They cover a wide range of domains and topics including creativity, politics, privacy preferences, storytelling, fairness perceptions, interactions with technology platforms, luxury consumption, news consumption, and labor market preferences, among others. See SI for details. Previous research typically tested digital twins by replicating well-known, published studies (Park et al., 2024, Toubia et al., 2025). While some of our sub-studies also replicated well-known studies, many were either based on unpublished research or they were developed specifically for this project. Some sub-studies explored various behavioral economics or psychological processes. Other developed personality scales. Some of our sub-studies compared treatment effects between twins and humans in various domains, while others employed within-subject designs or mixed-designs, and others were correlational studies. In other words, as a set, our sub-studies reflect how a diverse group of scholars would leverage digital twins as they are available today to conduct their research; hence, our mega-study provides an ecologically valid test of digital twins.

Each sub-study was run on a sub-sample of the original Twin-2K-500 participants (recruited from Prolific) and their digital twins, i.e., we match the answer of each digital twin to each question with the actual answer given by their human counterpart to the same question. This enables us to test, across a wide range of domains, the accuracy of digital twins at the *individual level* as well as the correlation between responses from the digital twins and their human counterparts. Our digital twins are based on a representative sample of the U.S. population. Hence, we can explore not only how performance varies across types of domains and questions, but also which groups are well represented by digital twins and which are not. By making our data (https://huggingface.co/datasets/LLM-Digital-Twin/Twin-2K-500-Mega-Study) and code (https://github.com/TianyiPeng/Twin-2K-500-Mega-Study) publicly available, we encourage follow-up work that will improve the performance of digital twins. Note that for privacy reasons we are unable to share the

Prolific IDs of our participants; therefore, other researchers are not able to run new studies on the original Twin-2K-500 participants as we are.

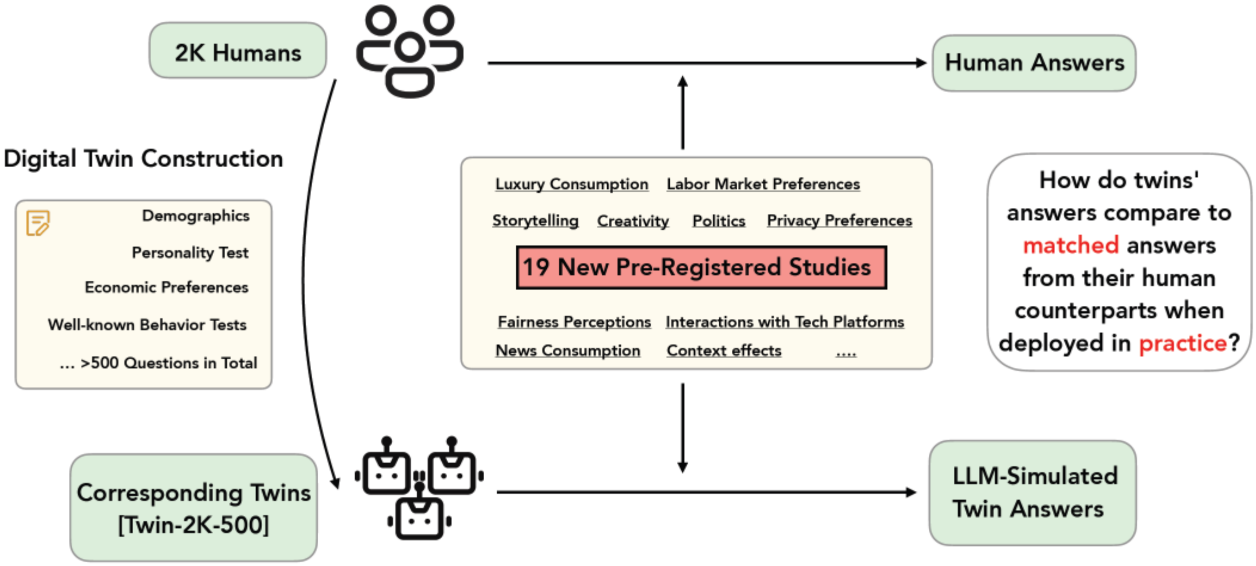

2K Humans
Human Answers
Digital Twin Construction
Demographics
Personality Test
Economic Preferences
Well-known Behavior Tests
... >500 Questions in Total
Luxury Consumption
Labor Market Preferences
Storytelling
Creativity
Politics
Privacy Preferences
19 New Pre-Registered Studies
Fairness Perceptions
Interactions with Tech Platforms
News Consumption
Context effects
....
How do twins' answers compare to matched answers from their human counterparts when deployed in practice?
Corresponding Twins [Twin-2K-500]
LLM-Simulated Twin Answers

Figure 1: Mega-Study Overview. We run 19 pre-registered studies on digital twins from the Twin-2K-500 dataset and their human counterparts. The studies were proposed by a diverse group of scholars and cover a wide range of behaviors and domains. As a set, they represent how digital twins may be leveraged today by social scientists. We match the answer of each digital twin to each question with the answer from their human counterpart, allowing us to explore the performance of digital twins both at the individual and population levels.

RESULTS

Pre-Registered Comparisons

Across our 19 sub-studies, we pre-registered 164 outcomes, for example binary choice on specific questions, and 31 comparisons between digital twins and humans based on these outcomes (1-4 comparisons per sub-study). These comparisons evaluated whether twins responded similarly to humans on average to a particular question or set of questions, whether the variance in their responses was similar, whether they replicated known treatment effects, among others. A complete list of these comparisons and their results are available in the Supplementary Information (SI). Overall, across the 31 comparisons, we found that twins aligned with humans in six cases, did not align in 20, and showed mixed results in five.

Some patterns emerge based on these 31 comparisons. Twins seemed relatively well-aligned with humans in social domains, such as assessing the fairness of various practices. On comparisons related to preferences and attitudes, twins struggled to capture human preferences, in particular in the political domain, where they expressed views that were often more “pro-humans” than those expressed by humans. For example, compared to their human counterparts, twins were more likely to believe that people are fair and can be trusted, that people should take care of themselves, and they were relatively favorable to people who donate to both political parties. They were also more favorable to government regulations of fees and surcharges in the marketplace, and more willing to pay taxes to improve healthcare for all people. In the cognitive domains, twins tended to display perfect knowledge on questions with objectively-correct answers, and they tended to behave more rationally than their human counterparts. When evaluating content, twins were not significantly correlated with humans when rating the creativity of ideas,[2] but they provided more reliable input when evaluating and predicting the content of book chapters. Finally, several of our comparisons related to human-technology interactions. There, twins tended to provide answers that were consistent with a view of technology as a safe tool under the control of humans. For example, twins were relatively more accepting of algorithmic hiring compared to their human counterparts, they perceived online targeting as less intrusive on their privacy, and they under-reported usage of platforms like Netflix and TikTok.

Digital Twin Performance

In order to analyze our data more systematically and with better statistical power, we switch our unit of analysis from the 31 comparisons to the 164 outcomes on which the comparisons were based.

[2] Interestingly, when told whether the ideas came from humans or from their twins, humans showed AI aversion but twins showed human aversion

A key advantage of our study is the ability to directly match the response of each digital twin to that of their human counterpart. We focus on performance measures that leverage this feature of the data. Specifically, for each outcome, we match each twin's response to its corresponding human's response and compute two metrics:

-Accuracy: Following previous research (Kolluri et al., 2025, Toubia et al., 2025), we measure accuracy as 1-*MAD*/*range*, where *MAD* is the mean absolute deviation between a human's and their twin's response, and *range* is the natural range of the outcome.[3] We compute the average **individual-level accuracy** across participants, where a higher value indicates greater accuracy.

-Correlation: We calculate the **correlation** between human and twin responses across participants for each outcome. This reflects the degree to which twins capture individual-level heterogeneity in human responses. To compute an overall average correlation across outcomes, we first apply a Fisher z-transformation to individual correlations, average them, and then convert back using an inverse transformation (Park et al., 2024).

In addition to these metrics based on matched human-twin data, we compute two distribution-level performance metrics that compare the aggregate *distributions* of responses from twins and humans for each outcome:

-Comparison of averages: We compare the **average response** from humans to that of their twins using the absolute value of Glass's delta.[4] Higher values indicate larger differences between group means (expressed in standard deviations).

-Comparison of standard deviations: We compute the **ratio of standard deviations** between humans and twins for each outcome.

[3] We exclude outcomes without a defined range, such as open-ended price estimates, resulting in 161 observations for this metric.

[4] Glass's delta is an effect size measure calculated as: $\Delta = (\bar{X}_{twin} - \bar{X}_{human}) / s_{human}$, where $\bar{X}_{twin}$ and $\bar{X}_{human}$ are the sample means of the twin and human groups respectively, and s_{human} is the standard deviation of the human group. It is similar to Cohen's d, but it relies on the standard deviation from the human sample only, as the assumption of equal variances between the two samples is not valid in our case.

See SI for histograms of all four performance metrics across outcomes. Across all outcomes, the average accuracy is 0.748. This compares favorably to the average accuracy of 0.717 reported by Toubia et al. (2025) using the same sample of human respondents. This is also in line with the test-retest accuracy of 0.817 found in that study, based on multiple questions being repeated. However, we note that a random benchmark achieves an accuracy level of 0.629, suggesting that the absolute levels of accuracy should be interpreted with caution. As an illustration, if the true responses were uniformly distributed on [0,1], always predicting 0 would yield an accuracy of 0.5. This does not imply that variations in accuracy are not meaningful, but this does stress the importance of comparing accuracy across relevant benchmarks, which we do next. The correlation between twin and human responses is positive in 157 of the 164 outcomes (95.7%), and it is significantly positive with $p<0.05$ in 128 of these cases. The overall average correlation is 0.197.[5] Average twin responses differ from human responses by 0.352 standard deviations on average;[6] a paired t-test finds a statistically significant difference ($p<0.05$) between humans and digital twins in 125 out of 164 outcomes (76.2%). Finally, the standard deviation of the twin responses is lower than that of human responses in 154 of 164 cases (93.9%), indicating under-dispersion in twin responses. This difference is statistically significant ($p<0.05$) in 146 of those 154 cases.

The performance of digital twins is influenced both by the base LLM (GPT4.1 in this case) as well as the individual-level data used to create each twin. To assess the gains from leveraging individual-level data to create digital twins, we test how different levels of personal detail influence performance, holding the base LLM constant. The five benchmarks are (see SI for more detail):

1. *Full Persona*: detailed description from the Twin-2K-500 dataset (approximately 128K characters). This is our default approach for creating digital twins.

[5] Note that Park et al. (2024) report a higher correlation in their digital twin study. However, they compute the correlation across questions for each participant, while we compute the correlation across participants for each outcome.

[6] This means the MSE of the digital twin's mean estimate is about, $(0.352\ \sigma)^2 \approx 0.12\sigma^2$, which approximately equal to the MSE of the sample mean from 10 humans, $0.1\sigma^2$.

2. *Persona summary:* a concise, statement-based summary by simplifying the questions and summarizing the responses with distributional information (approximately 13K characters).
3. *Demographics only*: includes the 14 demographic variables (region, sex, age, education, race, citizenship, marital status, religion, religious attendance, political party, household income, political ideology, household size, employment status) which are part of the Twin-2K-500 dataset (and therefore also part of the full persona and the persona summary).
4. *Empty Persona*: uses an identical system prompt for all twins (no individual information).
5. *Random Responses*: responses to each outcome are drawn from a uniform random distribution with a support equal to the outcome's range.[7]

Comparing the full persona version to the other versions shows that adding more detailed personal information improves correlations but has only a modest effect on accuracy (see Figure 2).[8] For example, twins built from full personas achieve only slightly higher accuracy than those based only on demographic information (p=0.37). While richer information increases the variance in twin responses, it does not increase the accuracy of their population mean estimate. We also see that the performance of twins is not statistically significantly different when the information is provided in raw vs. summary form.

In other words, enriching digital twins with detailed individual-level information improves our ability to capture heterogeneity across participants and predict relative differences between them, that is, it enhances our ability to distinguish one participant from another. However, it does not substantially improve our ability to reproduce the exact responses given by specific participants, nor does it enhance predictions of population-level averages.

[7] For the three outcomes that do not have a natural range, we use the empirical range in humans' answers as support for the random distribution.

[8] We drop one outcome in these comparisons, the creativity rating of ideas generated by humans and their twins in one of the sub-studies, as it was provided by humans and was only available for ideas from twins based on the full persona approach.

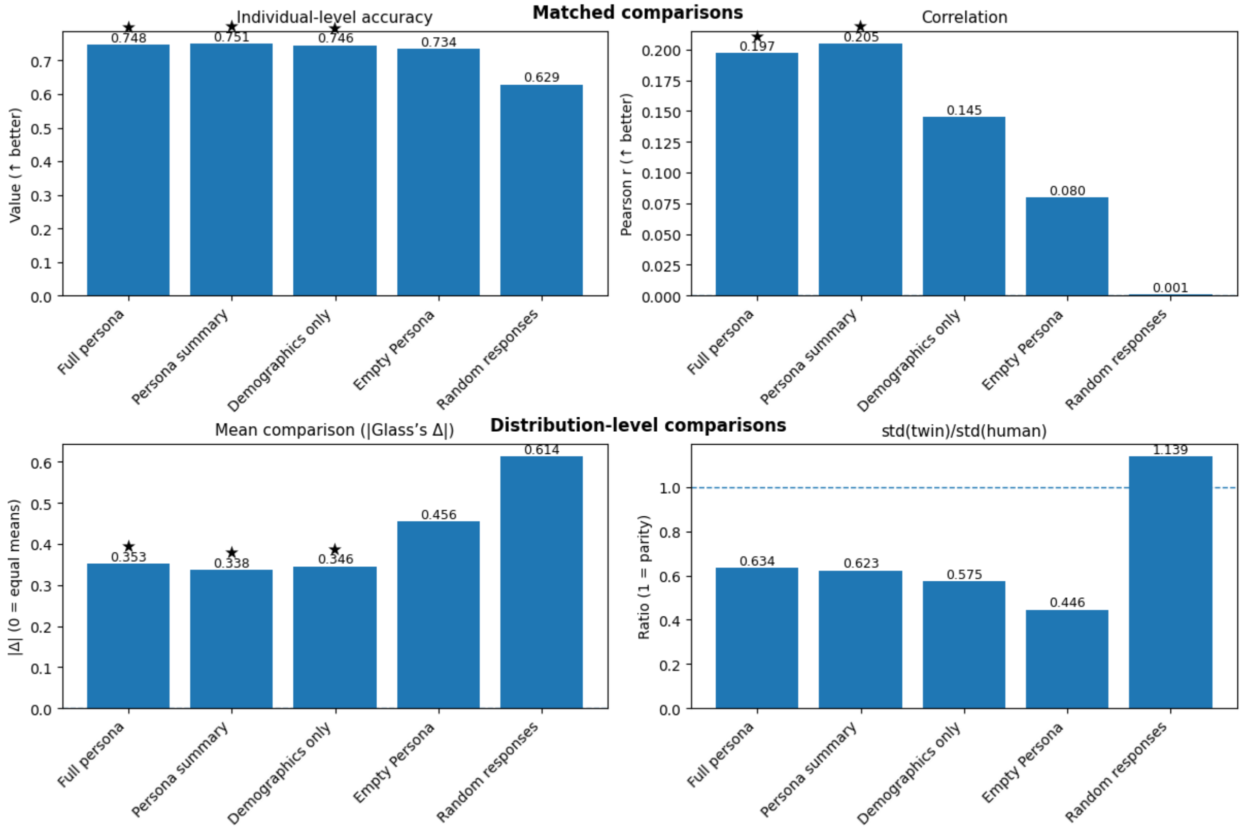


Figure 2: Gains from Leveraging Individual-Level Data. *: best performing benchmark, or not significantly different from best at $p<0.05$ (not applicable to ratio of standard deviations).

This finding aligns with Santurkar et al. (2023)'s argument that LLMs are *steerable* through customized prompting. In our case, adding digital twin information steers responses in ways consistent with observed variation across human participants (as observed by the improved correlation). However, as Santurkar et al. (2023) emphasize, steerability does not guarantee alignment with human responses. In our results, this manifests as modest or nonexistent gains in accuracy. To illustrate the distinction between improving correlation and improving accuracy, we select one outcome (Lack of Control) from one sub-study (Affective Primes), for which the distinction between accuracy and correlation is particularly sharp. Figure 3 reports scatter plots of the predictions from digital twins created using demographics only vs. human responses (left panel), and predictions from digital twins created using full personas vs. human responses (right panel). These plots clearly illustrate how the correlation is much improved when full personas are used

(r_{demo}=0.105 vs. r_{full}=.555). Yet, accuracy is virtually unchanged (accuracy_{demo}=0.892 vs. accuracy_{full}=0.907).

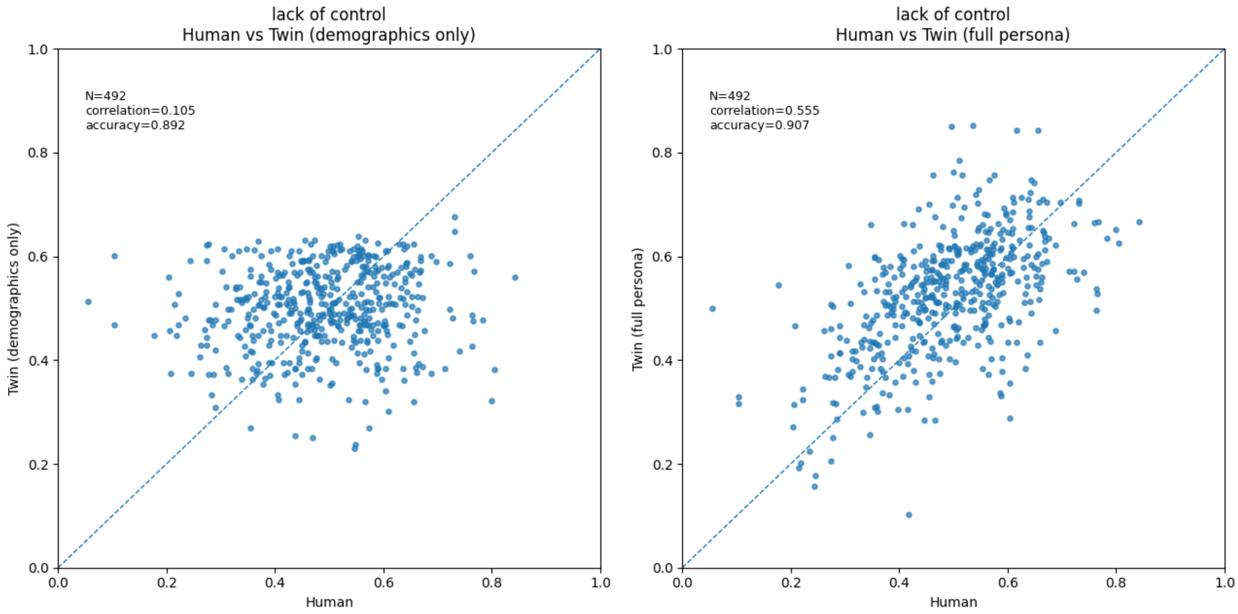


Figure 3: Human Responses vs. Digital Twin Responses based on Demographics Only (Left) and Human Responses vs. Digital Twin Responses based on Full Persona (Right), for One Particular Outcome.

Given these results, we focus on correlation as our key performance metric in the remainder of this section.

Meta-Analysis

We next explore patterns in performance across outcomes using a meta-analytic approach. We conduct a regression with the (z-transformed) correlation for each outcome as the dependent variable (see SI for meta-analysis of the other performance metrics).[9] Our independent variables include a set of categorical labels that characterize each outcome, related to the following domains: social, preferences/attitudes, cognitive skills/rationality, content evaluation, human-tech interactions. We also include mechanical features that might influence performance (e.g., sample size).

[9] We z-transform the correlation due to better statistical properties, e.g., being approximately normally distributed.

Labels are not mutually exclusive; an outcome may be tagged with multiple labels. All labels are binary, except sample size. The labels are as follows:

Social:

-Social domain (104 outcomes): involves social topics

-Conflict-related (15 outcomes): involves intra- or inter-personal conflict

-Pro-social (36 outcomes): involves pro-social topics such as altruism or cooperation

-Social cognition (15 outcomes): involves social cognition, i.e., asking the participant to judge the behavior of others

-Personality measure (6 outcomes): assesses personality traits

-Social desirability (55 outcomes): has a socially desirable answer

Preferences/attitudes:

-Attitude (33 outcomes): captures general attitudes toward a topic, issue, or person

-Political domain (12 outcomes): focuses on political content

-Preference measure (123 outcomes): measures the participant's preferences

-Behavioral intention (44 outcomes): captures future behavioral intentions

Cognitive skills/rationality:

-Cognitive domain (79 outcomes): relates to a cognitive domain

-Test of rationality (10 outcomes): assesses rational decision-making

Content evaluation:

-Evaluating content (62 outcomes): requires evaluation of stimuli

-Valenced (40 outcomes): involves positive vs. negative judgments

Human-tech interactions:

-Human-tech interactions (45 outcomes): concerns human interaction with technology

Mechanical:

-Replicates known human bias (5 outcomes): whether the outcome replicates a well-known human bias

-Different versions of question (82 outcomes): question wording varies across participants

-Scale question (142 outcomes): uses Likert or similar rating scales-Sample size (continuous): the number of participants on which the outcome is measured

We estimate a mixed linear model, including random intercepts for each sub-study. The meta-analysis reveals the following (see Figure 4):

- Digital twins performed relatively better in the social domain. The correlation was higher when outcomes related to conflict, pro-social issues, social cognition, or personality. However, the correlation was significantly *lower* for outcomes where social desirability was salient, suggesting twins are less capable of mimicking human responses in socially-sensitive contexts.
- The labels related to preferences and attitudes were not significant, except for outcomes related to political preferences: correlations were particularly low for outcomes in the political domain. This is consistent with previous findings such as Santurkar et al. (2023) or Motoki et al. (2024).
- Correlations were on average higher in the cognitive domain.
- Twins did not correlate with humans better or worse on average when evaluating content (i.e., the coefficient associated with content evaluation was not significant in the meta-analysis regressions). However, their responses were less correlated with humans when the evaluation was valenced (positive vs. negative).
- Twins generally showed higher correlation with humans on outcomes related to human-technology interactions.
- When questions varied across participants, twin-human correlation dropped, consistent with prior findings that LLMs struggle to treat prompt-level variation as exogenous (Gui and Toubia, 2025).
- Twins showed higher correlation with humans on outcomes that use response scales.

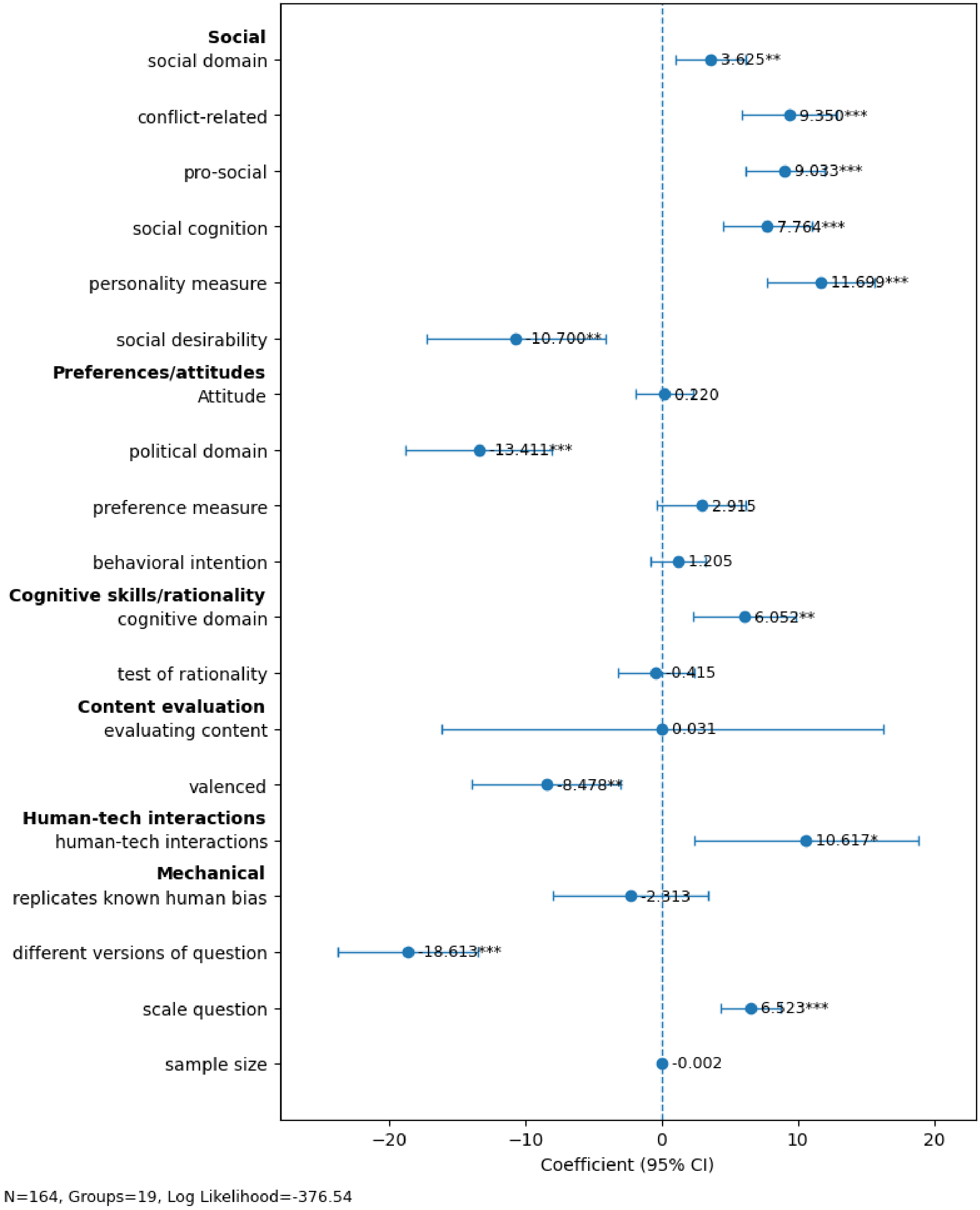


Figure 4: Results from Meta-Analysis (Mixed Linear Model with z-transformed Correlation as Dependent Variable). *, **, ***: p < 0.05, p < 0.01, p < 0.001.

Benchmarking

All results reported so far use GPT4.1 (dated 2025-04-14) with a default temperature of 0.7. We compare models with different temperatures and architectures (e.g., GPT-5, Deepseek, Gemini) and find the best results overall using GPT4.1 at temperature = 0 (e.g., highest correlation, tied for best on individual-level accuracy). Full persona prompts were used in all conditions. See SI for details.

We also compare the performance of digital twins to that of a traditional machine learning approach that would predict each outcome by collecting it from a subset of the population and training a model (e.g., XGBoost) linking personal information to that outcome in order to predict it for the rest of the population. We first consider a case in which the XGBoost model is trained on the full persona information (i.e., same information as digital twins). We find that in order to match the predictive correlation achieved by digital twins using such a traditional machine learning approach, one would need to collect each outcome from approximately 24% of the sample. We also consider an XGBoost model trained on demographic variables only. Such a model never reaches the performance of digital twins, even when using up to 30% of the sample as training sample. This suggests that although the correlation achieved by digital twins is modest in absolute terms, it is higher than what could be obtained by more costly, traditional machine learning approaches. See SI for details.

Participants Characteristics Associated with Digital Twin Performance

We next examine whether demographic characteristics of participants predict the accuracy of their digital twins. For each participant, we compute accuracy (across all outcomes) between the participant's normalized responses and those of their twin.[10] We then analyze how this accuracy relates to participants' demographic characteristics. All demographic variables are dummy-coded, yielding 61 demographic dummy variables

[10] The argument against measuring individual-level accuracy across participants for each outcome does not apply to measuring accuracy across outcomes for each participant. Hence, we switch to accuracy for this analysis as a more natural way to compare the answers between a particular human and their twin across outcomes.

coming from the 14 demographic questions. To identify which participant characteristics are most predictive of digital twin performance, we train an XGBoost (Extreme Gradient Boosting) model to predict twin accuracy using these features (Chen and Guestrin 2012).

To interpret the model's findings, we generate grouped partial dependence plots (PDPs) for each categorical variable. PDPs visualize the average predicted accuracy for each level of a feature while holding all other variables constant. For example, the PDP for political ideology shows predicted accuracy for participants identifying as "Very Conservative," "Liberal," "Moderate," “Liberal,” and “Very Liberal” on a single chart enabling direct comparison across different demographic groups. Plots for education, income, religious attendance, and political views are presented in Figure 5, while the others are in the SI.

Our findings suggest systematic differences in twin accuracy across demographic groups. Twins tend to be more accurate for participants with higher education levels and higher income, indicating a potential bias in favor of more socioeconomically-advantaged individuals. Accuracy also tends to be higher for participants with *moderate* political views compared to those with extreme ideological views. For example, predicted accuracy is highest among moderates and lowest among both very liberal and very conservative participants. We also find that accuracy peaks for participants who attend religious services once or twice a month but is lowest for those who never attend or attend weekly.

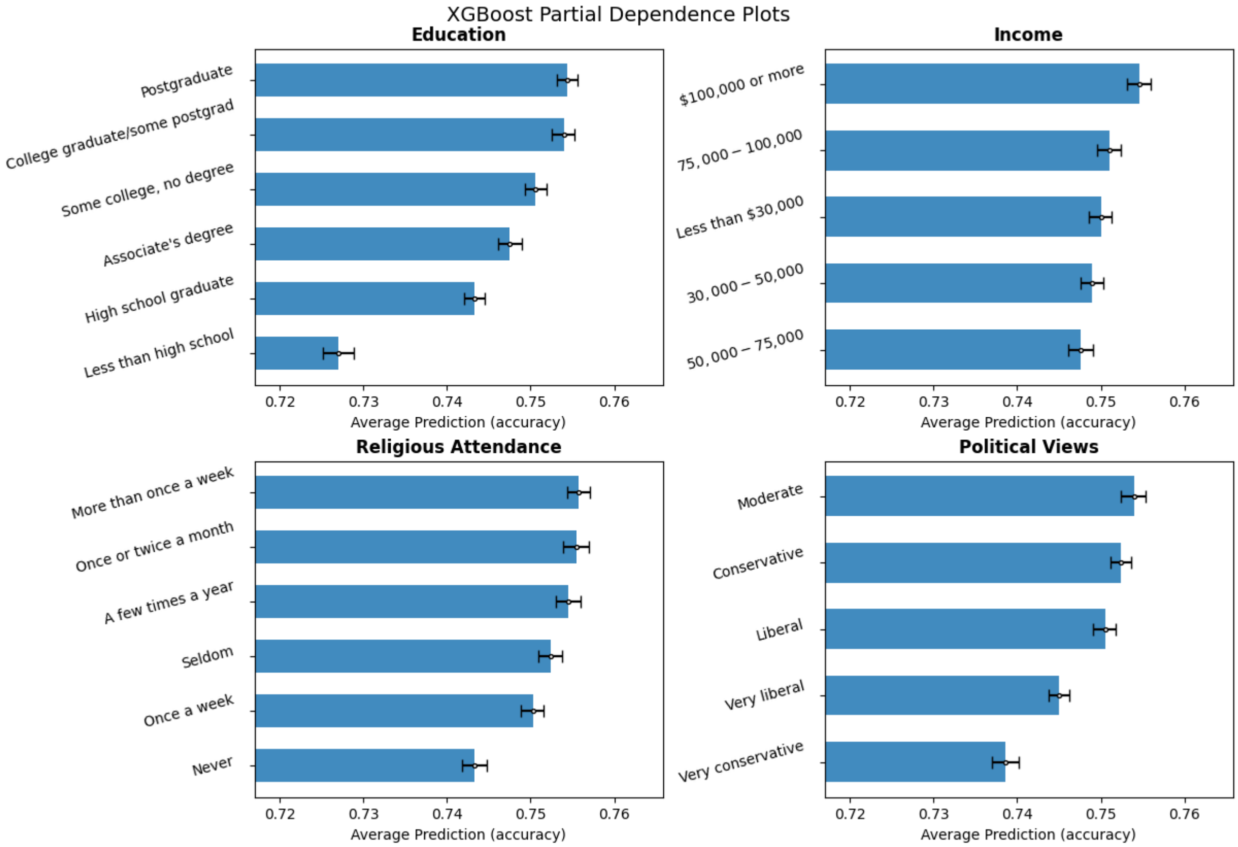


Figure 5: Partial Dependence Plots for Understanding Heterogeneity in Digital Twin Performance

DISCUSSION

Despite the significant interest in leveraging digital twins of humans to simulate surveys and experiments, little is known about how responses from digital twins compare to those of their human counterparts. Further, little is known about the best way to construct digital twins, on the domains of applications in which digital twins are likely to be accurate, or on the profiles of humans whose digital twins are likely to be more accurate. In order to help advance knowledge on digital twins, we contribute a publicly available dataset that: (i) allows matching the answers of digital twins to those of their human counterparts, (ii) covers a wide range of domains of applications, and (iii) involves a relatively large, representative sample of participants.

In a nutshell, we find that while digital twins show promise, they are not fully “ready for prime time” yet, i.e., significant improvements are needed until twins are ready for systematic deployment. We find that the rich individual-level information provided as input to digital twins improves our ability to distinguish between participants and capture heterogeneity, compared to simpler personas based on demographic information. The correlation between the responses from twins and their human counterparts is currently modest (approximately 0.2 on average), but it is nonetheless higher than what could be obtained by more costly, traditional machine learning approaches. In our data and with our current pipelines, we also find that the rich individual-level information contained in digital twins does not improve the accuracy of answers at the individual-level, nor does it improve the accuracy of estimates of population-level averages. Finally, we find that while the individual-level information contained in digital twins helps increase the variance of their answers, they are still under-dispersed compared to human answers.

Our meta-analysis further reveals domains in which digital twin responses are currently more strongly correlated with human responses. We find that digital twins are currently better equipped to approximate responses in social domains, except in cases where social desirability is a potential concern. Correlations in the cognitive domain are also higher, but digital twins are more likely to provide objectively-correct answers on questions that require knowledge or cognitive skills. More research is needed to understand the reliability of twins for content evaluation; our results suggest that twins struggle particularly with valenced evaluations. In general, our comparisons reveal some systematic biases, some of which are likely to come from the base LLM powering the digital twins. Indeed, a digital twin may be viewed as a combination of a base model with individual-level additional data, i.e., the answers are “shrunk” towards a base model, which may show systematic biases. For example, while we find relatively higher correlations on questions related to human-technology interactions, we also find that digital twins are more likely to behave as if they had a pro-technology agenda. And we find digital twins to be further disconnected from their human counterparts in the domain of politics.

Finally, our analysis suggests that the accuracy of digital twins is uneven across demographic groups, with better alignment for participants who are more educated, higher income, and ideologically moderate.

We hope that by making our data and code publicly available, we will help the field develop a data-driven, impartial understanding of the current opportunities and challenges presented by digital twins. We further hope to encourage the development of improved pipelines for digital twins, that will address some of the limitations uncovered in our analysis.

REFERENCES

1. Argyle, Lisa P., et al. "Out of one, many: Using language models to simulate human samples." Political Analysis 31.3 (2023): 337-351.
2. Binz, M. *et al.* A foundation model to predict and capture human cognition. *Nature*, 1–8 (2025).
1. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).
2. Cui, Z., Li, N. & Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 1–8 (2025).
3. Dillion, Danica, et al. "Can AI language models replace human participants?." *Trends in Cognitive Sciences* 27.7 (2023): 597-600.
4. Gui, G. & Toubia, O. The challenge of using LLMs to simulate human behavior: A causal inference perspective. *arXiv* **arXiv:2312.15524** (2023).
5. Hewitt, L., Ashokkumar, A., Ghezae, I. & Willer, R. Predicting results of social science experiments using large language models. Preprint (2024).
6. Kolluri, Akaash, et al. "Finetuning LLMs for Human Behavior Prediction in Social Science Experiments." arXiv preprint arXiv:2509.05830 (2025).
7. Motoki, F., Pinho Neto, V. & Rodrigues, V. More human than human: Measuring ChatGPT political bias. Public Ch*oice* **198**, 3–23 (2024).
8. Park, J. S. *et al.* Generative agent simulations of 1,000 people. *arXiv* **arXiv:2411.10109** (2024).
9. Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).
10. Santurkar, S. *et al.* Whose opinions do language models reflect? In *Proceedings of the International Conference on Machine Learning* (PMLR, 2023).

11. Toubia, O. *et al*. Twin-2K-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science* (2025).

METHODS

Each sub-study was first run on Prolific according to its pre-registered plan (https://researchbox.org/4145). All sub-studies were run from the same Prolific account to ensure consistency. Invitations were sent to Prolific participants from the Twin-2K-500 panel (the Prolific IDs of our participants are not publicly available due to privacy concerns).

Next, each sub-study was conducted on the digital twins of its participants. To set up the simulation, we extracted twin information from the Twin-2K-500 dataset and provided it as prompt input. For each human participant, the completed survey was transformed into text, stripped of its answers, and incorporated into the prompt. Because survey questions varied across participants due to randomization, we preserved these differences when generating inputs for their twins. The LLM was then tasked with producing responses to these surveys. A sample prompt template is provided below. All code is released as open source.

```
You are an AI assistant. Your task is to answer the 'New Survey Question'
as if you are the person described in the 'Persona Profile' (which
consists of their past survey responses).

Remain consistent with the persona's previous answers and stated traits.
Simulate their responses to new questions while accounting for human
cognitive limitations, uncertainty, and biases.

Follow all instructions provided for the new question carefully regarding
the format of your answer.

## Persona Profile (This individual's past survey responses):
{Persona Profile}

## New Survey Question & Instructions (Please respond as the persona
described above):
{Survey Questions}

### Format Instructions:
In order to facilitate the postprocessing, you should generate a JSON
object by filling in Masks with the appropriate values in the following
template for {N} questions, where each question Q1, Q2, ... corresponds to
the question Q1, Q2, ... above.
{Format Templates}
```

For each persona in each sub-study, we construct a prompt using the template above and query the LLM via its API (OpenAI API for OpenAI models, OpenRouter API for others). The JSON outputs are then post-processed to ensure all questions are answered in the required format (e.g., no missing values or out-of-range responses). If validation fails, the API call is retried.

Example of Survey Questions:

```
Q1:
Introduction: You are being invited to participate in a research
study… By selecting the ‘Agree’ option below I am agreeing to
participate, I have not given up any of the legal rights that I would
have if I were not a participant in the study.
Question Type: Single Choice
Options:
  1 - Agree
  2 - Disagree
Answer: [Masked]

Q2:
How important is it to you that your employer actively invests in
environmental sustainability (e.g., reducing carbon emissions,
minimizing waste)?
Question Type: Single Choice
Options:
  1 - Not at all important
  2 - Slightly important
  3 - Moderately important
  4 - Very important
  5 - Extremely important
Answer: [Masked]

Q3:
Would you prefer to work at a company that is outspoken about social
and political issues, or one that remains neutral?
Question Type: Single Choice
Options:
  1 - Strongly prefer outspoken
  2 - Somewhat prefer outspoken
  3 - No preference
  4 - Somewhat prefer neutral
  5 - Strongly prefer neutral
Answer: [Masked]
……
```

Example of Format Templates:

```
{
    "Q1": {
        "Question Type": "Single Choice",
        "Answers": {
            "SelectedByPosition": Masked, // a number from 1 to 2,
corresponding to the option position
        "SelectedText": "Masked" // a string, corresponding to the
option text
        }
    },
    "Q2": {
        "Question Type": "Single Choice",
        "Answers": {
            "SelectedByPosition": Masked, // a number from 1 to 5,
corresponding to the option position
        "SelectedText": "Masked" // a string, corresponding to the
option text
        }
    },
….
```

SUPPLEMENTAL INFORMATION FOR “A MEGA-STUDY OF DIGITAL TWINS REVEALS STRENGTHS, WEAKNESSES AND OPPORTUNITIES FOR FURTHER IMPROVEMENT”

CONTENTS

***Summary of Pre-Registered Comparisons Between Twins and Humans*......................4**

***Distribution of Each Performance Metric Across Outcomes* 12**

***Meta-analysis of Other Performance Metrics*.. 14**

***Details for Persona Construction* 17**

***Benchmarking*... 22**

***Comparison with Traditional Machine Learning Approach*..23**

***Additional Partial Dependence Plots from XGBoost* 25**

***Detailed Information on Each Sub-Study*...26**

Context Effects...26

Main Questions/Hypotheses 26
Methods 26
Results - Pre-registered Analyses 26
Results - Additional Analyses (Non-Preregistered).. 28
Discussion.. 28
References 29

Default Effects 30

Main Questions/Hypotheses 30
Methods 30
Results - pre-registered analyses 30
Results - Additional Analyses (Non-Preregistered).. 32
Discussion.. 32
References 32

Fees Accuracy..33

Main Questions/Hypotheses 33
Methods 33
Results - Pre-registered Analyses 35
Discussion.. 37
References 38

Measures of Creativity 39

Main Questions/Hypotheses 39
Methods 39
Results - Pre-registered Analyses 40
Results - Additional Analyses (Non-Preregistered).. 41
Discussion.. 41
References.. 41

Idea Evaluation..43

Main Questions/Hypotheses 43
Methods 43

Results - Pre-registered Analyses 44
Discussion 46
References 46

Quantitative Intuition 47
Main Questions/Hypotheses 47
Methods 47
Results - Pre-registered Analyses 48
Results - Additional Analyses (Non-Preregistered) 50
Discussion 52
References 52

Promiscuous Donors 53
Main Questions/Hypotheses 53
Methods 53
Results - Pre-registered Analyses 55
Discussion 56
References 57
Appendix - Regression Tables 57

Privacy Preferences 61
Main Questions/Hypotheses 61
Methods 61
Results - Pre-registered Analyses 62
Discussion 63

Heterogenous Story Beliefs 64
Main Questions/Hypotheses 64
Methods 64
Results - Pre-registered Analyses 64
Discussion 65

Targeting Fairness 66
Main Questions/Hypotheses 66
Methods 66
Results - Pre-registered Analyses 66
Results - Additional Analyses (Non-Preregistered) 66
Discussion 67
References 67

User Behavior with Recommendation Systems 68
Main Questions/Hypotheses 68
Methods 68
Results - Pre-registered Analyses 69
Results - Additional Analyses (Non-Preregistered) 70
Discussion 70
References 70

Preferences for Redistribution 71
Main Questions/Hypotheses 71
Results - Pre-registered Analyses 73
Results - Additional Analyses (Non-Preregistered) 74
Discussion 75
References 76

Consumer Minimalism 77

Main Questions/Hypotheses 77
Methods 77
Results - Pre-registered Analyses 79
Results - Additional Analyses (Non-Preregistered) 81
Discussion 81
References 81

Infotainment News Sharing 82
Main Questions/Hypotheses 82
Methods 82
Results - Pre-registered Analyses 84
Results - Additional Analyses (Non-Preregistered) 88
Discussion 89
References 90

Accuracy Nudges for Misinformation 92
Main Questions/Hypotheses 92
Methods 92
Results - Pre-registered Analyses 94
Results - Additional Analyses (Non-Preregistered) 98
Discussion 105
References 106

Digital Certificates for Luxury Consumption 108
Main Questions/Hypotheses 108
Methods 108
Results - Pre-registered Analyses 110
Results - Additional Analyses (Non-Preregistered) 112
Discussion 114
References 114

Affective Primes 116
Main Questions/Hypotheses 116
Methods 116
Results - Pre-registered Analyses 117
Results - Additional Analyses (Non-Preregistered) 120
Discussion 121
References 122

Obedient Twins 124
Main Questions/Hypotheses 124
Methods 124
Results - Pre-registered Analyses 126
Results - Additional Analyses (Non-Preregistered) 127
Discussion 129
References 130

Hiring Algorithms 131
Main Questions/Hypotheses 131
Methods 131
Results - Pre-registered Analyses 134

SUMMARY OF PRE-REGISTERED COMPARISONS BETWEEN TWINS AND HUMANS

(see SI section specific to each sub-study for details)

Sub-study	Pre-registered comparison(s)	Result(s)
Context Effects	Can digital twins demonstrate the attraction effect (Huber et al., 1982) and the compromise effect (Simonson, 1989)?	Humans replicated the attraction effect, but twins did not. Neither humans nor digital twins exhibited a compromise effect.
Defaults Effects	Can digital twins demonstrate default effects (Johnson & Goldstein, 2003; Pichert & Katsikopoulos, 2008)?	Human participants showed a default effect in a green energy adoption paradigm but not in an organ donation paradigm. Twins predicted a strong default effect in the organ donation paradigm but no effect in the green energy paradigm.
Fees Accuracy	(i) How does the knowledge of digital twins regarding fees and surcharges in the marketplace compare to their human counterparts? (ii) How do perceptions of fairness regarding fees and surcharges in the marketplace compare between digital twins and their human counterparts? (iii) How does the variation in perceptions of fairness regarding fees and surcharges in the marketplace compare between digital twins and their human counterparts? (iv) How do beliefs about government regulation regarding fees and surcharges in the marketplace compare between digital twins and their human counterparts? (v) How does the variation in beliefs about government regulation regarding fees and	(i) Twins were much more accurate than humans on knowledge questions. (ii) Twins closely mirrored the average human response on fairness questions. (iii) Twins exhibited narrower variance than humans on fairness perceptions. (iv) Twins were on average more favorable to government regulation compared to humans. (v) For both groups, conservatism was correlated with less support for pricing regulation. However,

	surcharges in the marketplace compare between digital twins and their human counterparts?	twins stereotyped participants based on their political ideology, with exaggerated support for regulation for liberal participants.
Measures of Creativity	(i) How do digital twins' creative abilities compare to those of their human counterparts? (ii) Can digital twins replicate the correlations between creativity tests and idea generation performance observed in humans?	(i) The ideas generated by twins were judged as more creative (and with less dispersion in creativity ratings) than the ideas generated by their human counterparts. Digital twins also performed better (and with less dispersion) at the DAT, a creativity task. (ii) The correlation previously found between DAT and creativity was replicated, both among humans and twins.
Idea Evaluation	(i) How do digital twins' ratings of idea creativity compare to those of their human counterparts? (ii) Do digital twins show human favoritism, AI favoritism, or neither when evaluating ideas?	(i) Creativity ratings coming from twins were not significantly correlated with the ratings coming from their human counterparts. (ii) When no information was provided about the source of the ideas, humans rated twin ideas higher (more creative) than human ideas, and twins rated human ideas higher than twin ideas. When information was provided about the source of ideas, humans displayed AI aversion, and twins displayed human aversion.
Quantitative Intuition Survey	What is the relationship between Quantitative Intuition (QI) measures and demographics, personality traits, and individual measures like overconfidence and numeracy? Can digital twins capture Quantitative Intuition (QI) skills?	We found a moderate correlation between human responses to the individual QI scale and those of their twins, and a weaker relationship in assessments of the QI score of the organization they work(ed) for. Twins failed to mimic responses to two QI

		behavior questions. Twins tended to select the normative, high-QI behavior much more frequently than humans did.
Promiscuous Donors	How do humans and twins differ in their reaction to (i) others who do not donate to either party, (ii) others who donate to both parties, (iii) justification from those who donate to both parties based on pragmatic vs. value-based reasons?	Human respondents penalized promiscuous donors, especially when the behavior appeared motivated by self-interest (e.g., cultivating access or recruiting talent). In contrast, twins were less punitive (sometimes even favorable) toward the very same targets.
Privacy Preferences of Digital Twins	How do humans and their digital twins vary in their perceptions of privacy violations for different data processing and targeting practices in online advertising?	Twins perceived their privacy to be violated less than their human counterparts. We found almost no correlation in preferences for individual-level privacy violation scores for twins compared to humans. However, in aggregate, humans and their twins ranked privacy violations similarly across advertising practices that rely differently on tracking and targeting (high rank correlation).
Heterogeneous Story Beliefs	When reading a book and predicting the emotion (valence and arousal) of the next chapter based on the previous chapter, what is the correlation between the expected valence and expected arousal from humans vs. their digital twins?	We found significant positive correlations for both valence and arousal.
Targeting Fairness Perceptions	How do humans and twins perceive the fairness of demographic-based versus broad targeting?	At the aggregate level, the twins closely matched human judgments of the perceived fairness of demographic targeting, replicating the finding

		that it is viewed as less fair than broad targeting and producing similar average ratings across conditions.
User Behavior with Recommendation Algorithms	(i) How do self-reported average time spent on the platform differ across humans and their digital twin counterparts? (ii) How does knowledge that recommendation algorithms depend on the user's past behavior compare between humans and their digital twin counterparts? (iii) How does the prevalence of strategization (modifying your behavior to influence future recommendations) compare between humans and their digital twin counterparts? (iv) How does the preference for user controls over recommended content compare between humans and their digital twin counterparts?	(i) Twins under-reported platform usage. (ii) Twins were more aware of how recommendation algorithms work. (iii) Twins were more likely to strategize. (iv) Twins preferred more control over algorithms.
Preferences for Redistribution	Can digital twins predict preferences for redistribution and other outcomes like trust, socio-economic background, and beliefs on fairness and social mobility?	Twins showed more extreme redistribution preferences. On average, they were more likely to think "people should take care of themselves" but more willing to pay taxes to improve healthcare for all people. They were more likely to think people are fair and that people can be trusted. They were also more likely to think that hard work and luck are equally important for success. Finally, they showed very little distribution in their fathers' educational attainment, with almost 100% of them reporting their father only completed high school.

Consumer Minimalism	How does the predictive validity of the Consumer Minimalism Scale vary in a sample of human vs. their digital twins? The prediction is that respondents scoring high on the minimalism scale will prefer minimalist home environments more.	The predictive validity of the Consumer Minimalism Scale documented in human samples replicated robustly in the twin sample. The scale predicted consumers' preferences for minimalist versus non-minimalist apartment interiors.
Infotainment News Sharing	When choosing an article to share on social media, will digital twins prioritize the following attributes similarly to their human counterpart: 1. Headline (more vs. less entertaining) 2. Source (more vs. less trustworthy) 3. Content type (entertaining vs. informative) 4. Political lean (conservative vs. liberal) 5. Number of likes (20 vs. 200 vs. 2,000)	While humans and twins both prioritized headline, this consideration mattered much less to twins, who equally relied on number of likes and put far more weight on political lean than their human counterparts. Twins were also sensitive to source trustworthiness, while humans did not differentiate between more or less trustworthy sources.
Accuracy Nudges for Misinformation	Is the effect of prompting people to think about accuracy on sharing of untrustworthy (vs. trustworthy) news similar for twins compared to their human counterparts?	When primed to consider accuracy or reminded of its importance, humans showed no statistically reliable reduction in their willingness to share false versus true entertainment headlines. In contrast, twins responded as the prior work would predict, demonstrating robust improvements in truth discernment.
Digital Certifications for Luxury Consumption	How does the propensity to hold higher status perceptions for a luxury product associated with, vs. not associated with, a digital certification, differ between twins and their human counterparts?	Digital certification via a digital passport increased the value perception of a diffused luxury product for humans. Digital twins' value perceptions, however, were not impacted by the inclusion of a digital passport.

Affective Primes	Research in the social sciences often manipulates emotional states using affective priming, where participants are asked to reflect on a recent event. The broad question here is whether affective priming works with digital twins. Specifically:	
	(i) Do affective priming manipulations induce “states” in digital twins (i.e., does writing about gratitude or lack of control momentarily influence digital twins’ responses?)	(i) Yes. Writing about feeling grateful or out of control significantly increased corresponding feelings of gratitude and lack of control in digital twins, with these induced “affective states” subsequently influencing downstream outcomes.
	(ii) Is the influence of affective priming similar for digital twins and their human counterparts?	(ii) No. Affective priming had stronger effects on digital twins' affective responses compared to humans for both gratitude and lack of control manipulations.
	(iii) Is the influence of affective priming on digital twins dependent on the valence of affective prime (i.e., between positive primes like gratitude or negative primes like lack of control) or the proximal nature of the measures (i.e., does it “spill over” to other, related dimensions or only influence proximal dimensions)?	(iii) Yes. Valence affected twins’ responses, but in a similar way to humans: The gratitude prime had a stronger effect on the manipulation check than the lack of control prime for both humans and their twins.
	(iv) Is there a relationship between what digital twins and their human counterparts say in their response to the manipulation?	(iv) Yes. The semantic content written by each human was significantly more similar to their twin’s content than a randomly selected twin’s content.

Obedient Twins	LLMs are trained to be obedient and deferential. Does this tendency make digital twins more sensitive to survey instructions than their human counterparts? We tested this in three tasks: (i) Self-persuasion (Briñol, McCaslin, and Petty 2012). When prompted to consider the other side, do digital twins abandon their attitudes more readily than their human counterparts? (ii) Scenarios. Do digital twins more “obediently” follow the instructions to imagine themselves in different scenarios, leading to more sensitivity to scenario manipulations? (iii) Absurd Scenarios. Do digital twins “earnestly” respond to instructions that would be non-sensical to their human counterparts?	(i) No. Twins were less sensitive to considering a different point of view compared to humans. (ii) No. Twins were less sensitive overall to these scenario manipulations compared to humans. (iii) No. Humans were more sensitive to the instructions to forecast feelings in absurd scenarios.
Hiring algorithms	Can digital twins predict job candidate preferences for workplace hiring policies? Specifically: (i) Can digital twins predict workplace preferences? (ii) Can experimenting on a population of digital twins recover the true average treatment effects as experimenting on a human population?	(i) Twins tended to express views that were overall more pro-social (sustainability, ESG, work-life balance, transparent leadership, culture and values), and less "career-oriented" (firm prestige, career development) than those expressed by humans. (ii) Twins were overall less averse to hiring algorithms compared to humans.

Table SI.1. Pre-Registered Comparisons and Results.

REFERENCES

Briñol, Pablo, Michael J. McCaslin, and Richard E. Petty. "Self-generated persuasion: effects of the target and direction of arguments." *Journal of Personality and Social Psychology* 102.5 (2012): 925.

Huber, Joel, John W. Payne, and Christopher Puto. "Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis." *Journal of consumer research* 9.1 (1982): 90-98.

Johnson, Eric J., and Daniel Goldstein. "Do defaults save lives?." *Science* 302.5649 (2003): 1338-1339.

Pichert, Daniel, and Konstantinos V. Katsikopoulos. "Green defaults: Information presentation and pro-environmental behaviour." *Journal of environmental psychology* 28.1 (2008): 63-73.

Simonson, Itamar. "Choice based on reasons: The case of attraction and compromise effects." *Journal of consumer research* 16.2 (1989): 158-174.

DISTRIBUTION OF EACH PERFORMANCE METRIC ACROSS OUTCOMES

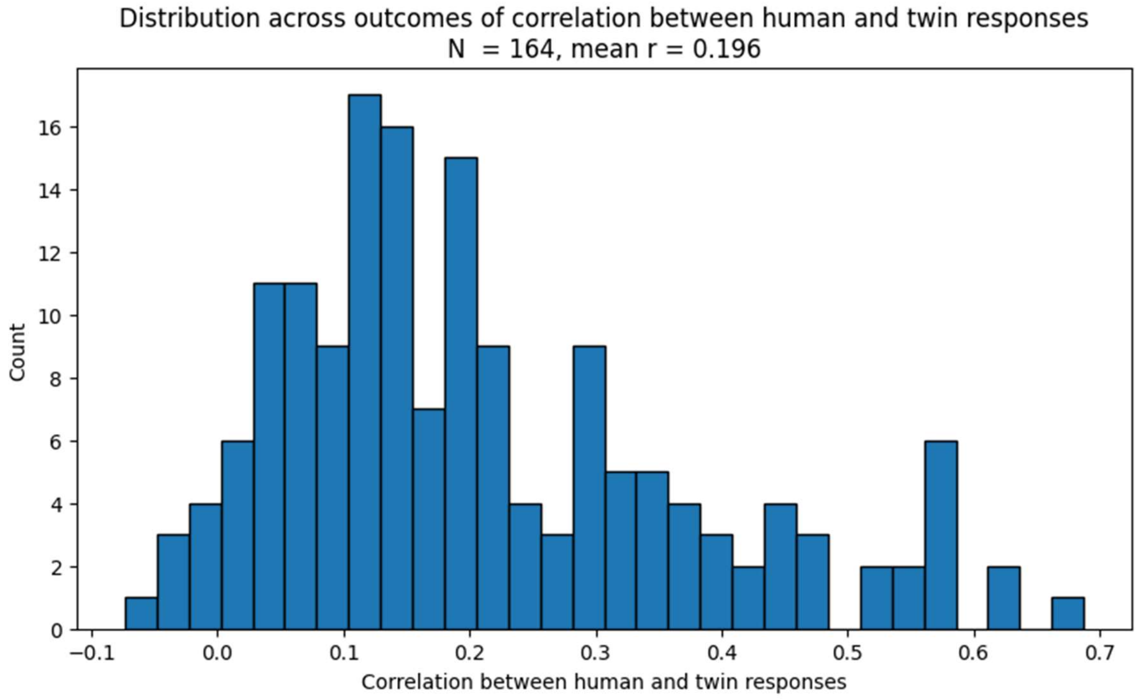


Figure SI.1. Distribution Across Outcomes of Correlation Between Human and Twin Responses.

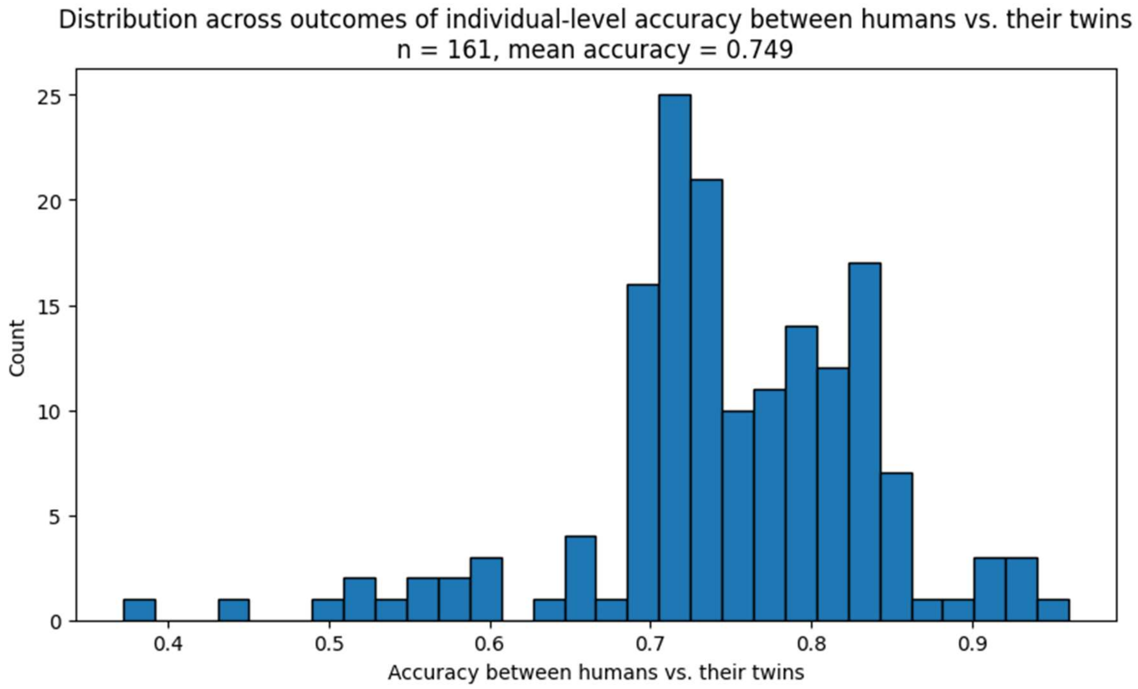


Figure SI.2. Distribution Across Outcomes of Individual-Level Accuracy Between Human and Twin Responses.

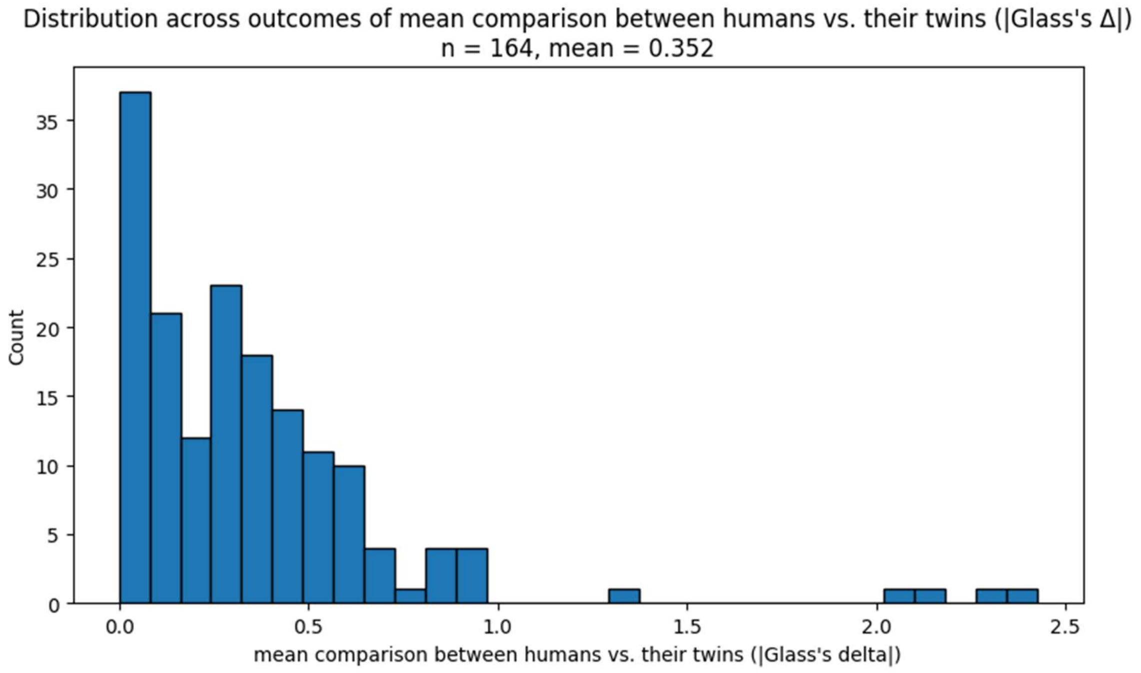


Figure SI.3. Distribution Across Outcomes of |Glass's Δ| Between Human and Twin Responses.

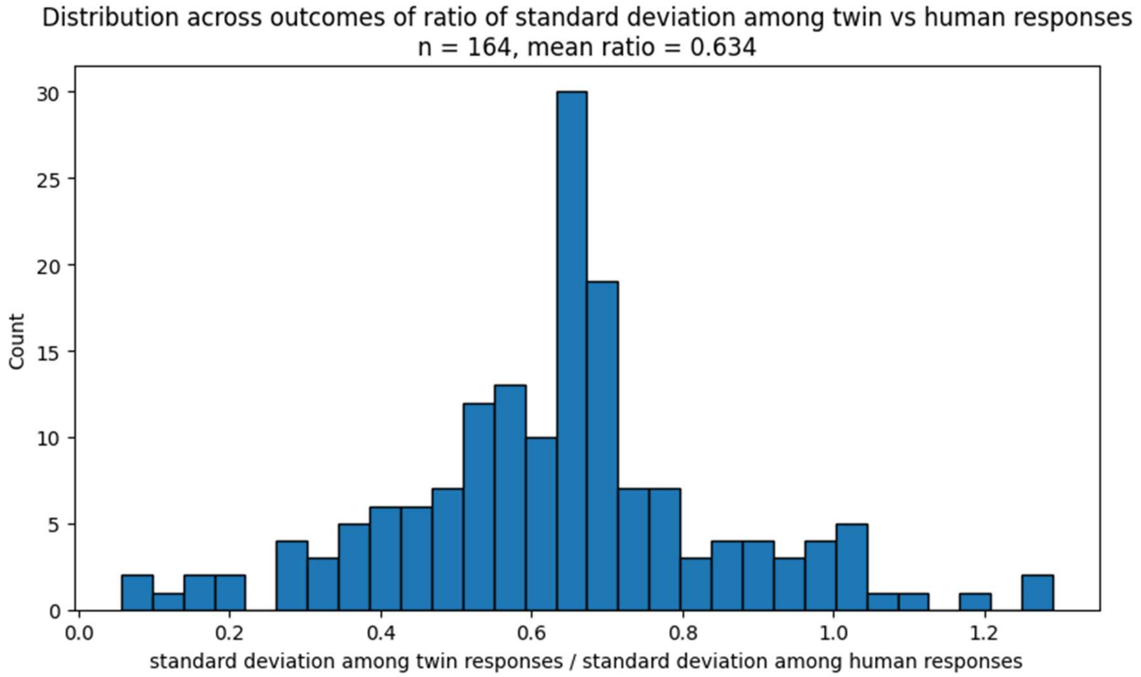


Figure SI.4. Distribution Across Outcomes of Ratio of Standard Deviation Between Human and Twin Responses.

META-ANALYSIS OF OTHER PERFORMANCE METRICS

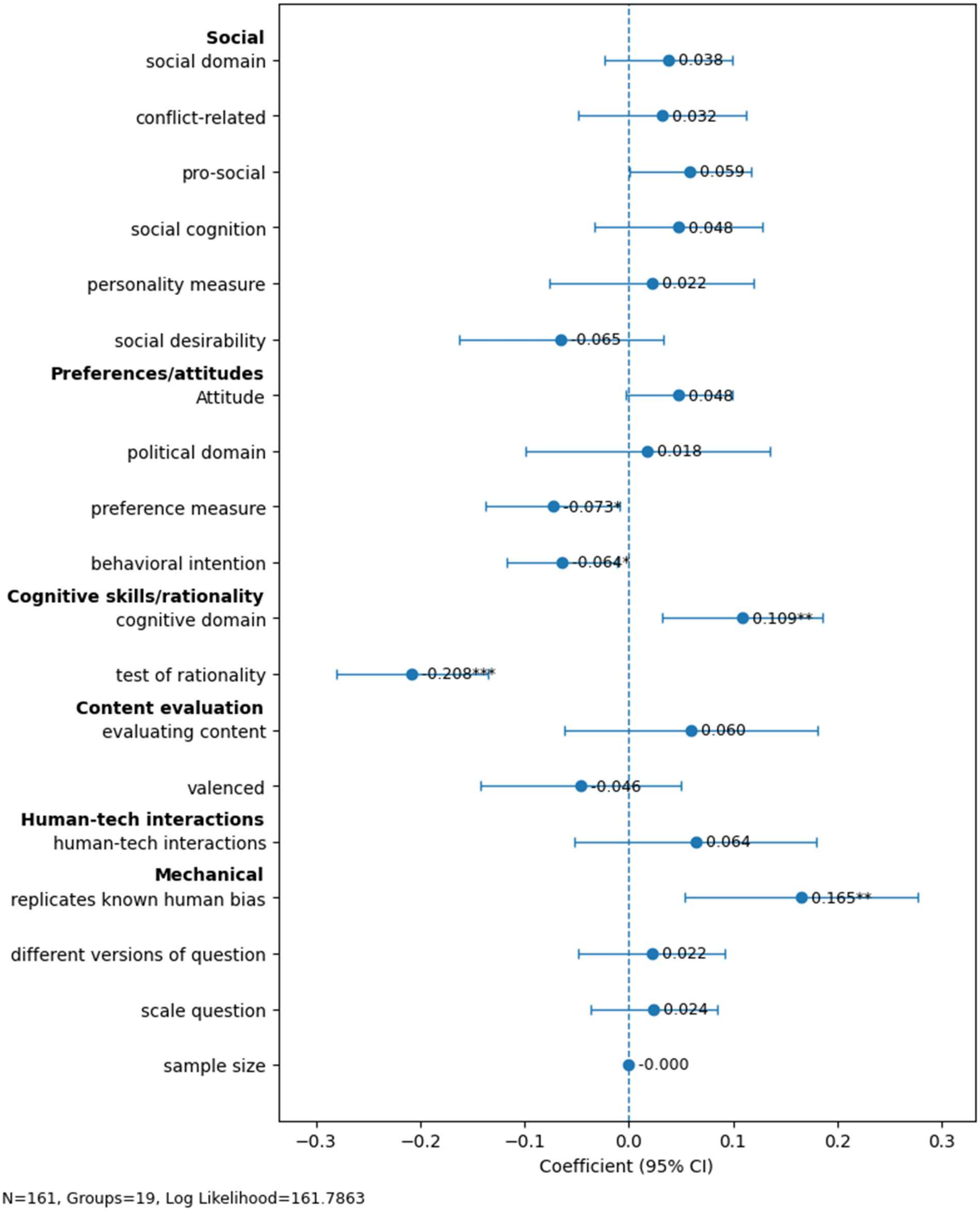
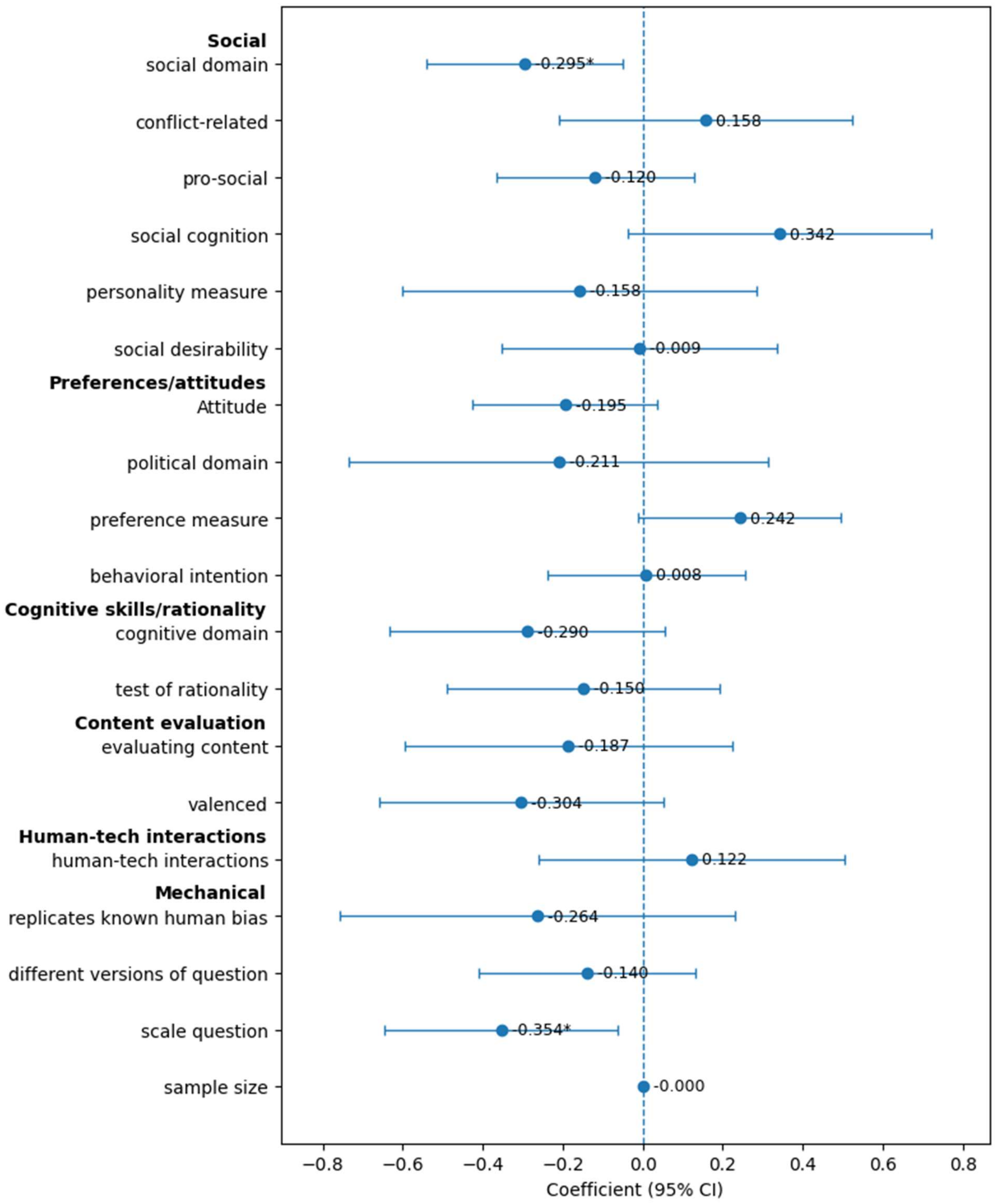


Figure SI.5: Results from Meta-Analysis (Mixed Linear Model with Individual-Level Accuracy as Dependent Variable). *, **, ***: p < 0.05, p < 0.01, p < 0.001.



N=164, Groups=19, Log Likelihood=-60.6866

Figure SI.6: Results from Meta-Analysis (Mixed Linear Model with |Glass's Δ| as Dependent Variable). *, **, ***: $p < 0.05$, $p < 0.01$, $p < 0.001$.

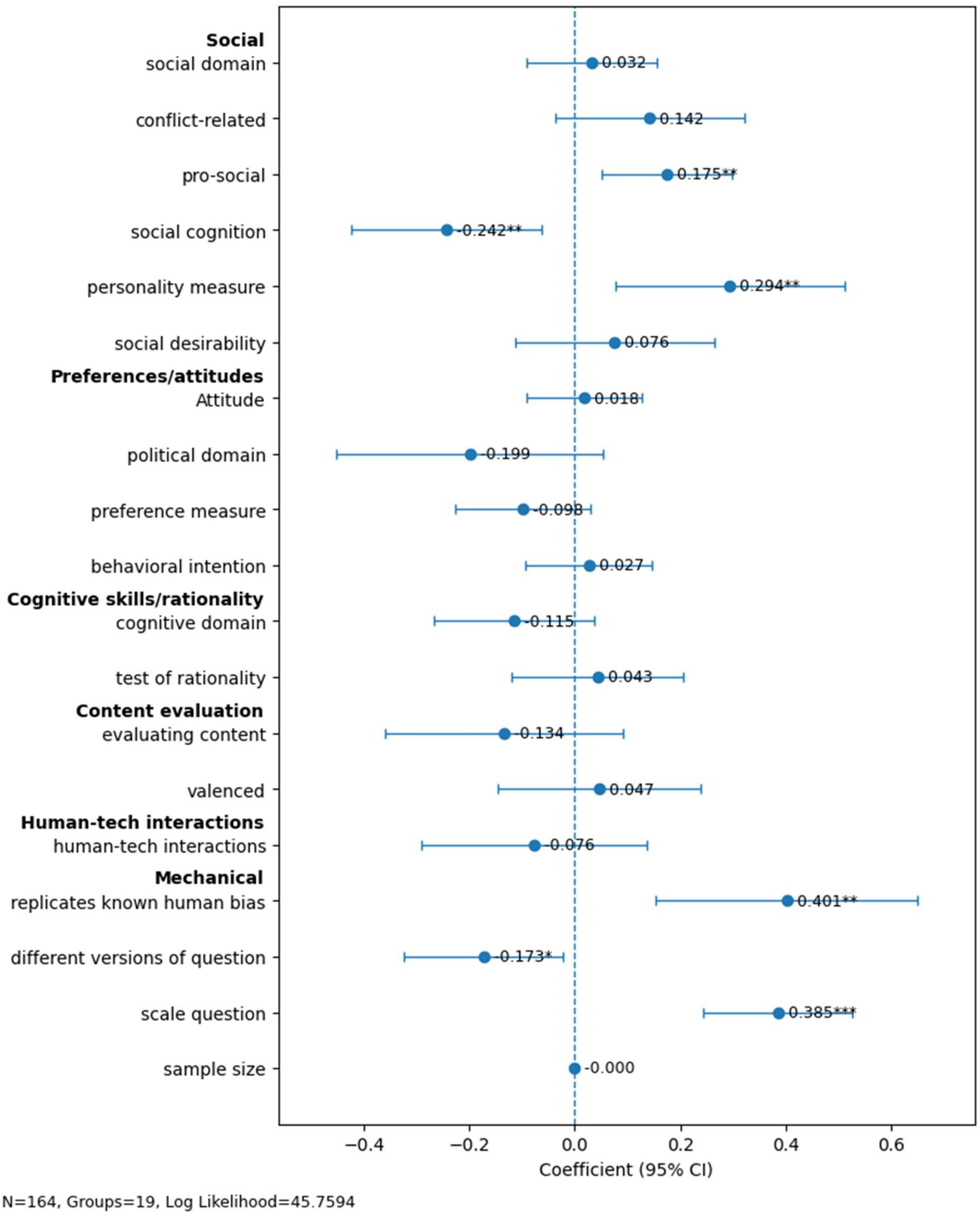


Figure SI.7: Results from Meta-Analysis (Mixed Linear Model with std(twin) / std(human) as Dependent Variable). *, **, ***: p < 0.05, p < 0.01, p < 0.001.

DETAILS FOR PERSONA CONSTRUCTION

In this section, we detail the construction of the full persona, persona summary, and demographics persona versions of digital twins.

Full Persona

Our full persona is built from the text-based questions and answers in Twin-2K-500, reflecting natural language interactions. The Twin-2K-500 dataset consists of four waves, with wave 4 repeating questions from waves 1–3 for test–retest purposes. For persona construction, when a question appears in both waves 1–3 and wave 4, we use the wave 4 responses. The complete persona dataset can be accessed at https://huggingface.co/datasets/LLM-Digital-Twin/Twin-2K-500, specifically the `persona_text` split within the `full_persona` subset. An example of (portions of) a full persona is below:

```
Which part of the United States do you currently live in?

Question Type: Single Choice

Options:

  1 - Northeast (PA, NY, NJ, RI, CT, MA, VT, NH, ME)

  2 - Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH)

  3 - South (TX, OK, AR, LA, KY, TN, MS, AL, WV, DC, MD, DE, VA, NC, SC, GA, FL)

  4 - West (WA, OR, ID, MT, WY, CA, NV, UT, CO, AZ, NM)

  5 - Pacific (HI, AK)

Answer: 2 - Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH)


What is the sex that you were assigned at birth?

Question Type: Single Choice

Options:

  1 - Male

  2 - Female

Answer: 2 - Female
```

….

```
Here are a number of characteristics that may or may not apply to you. Please indicate next to each statement the extent to which you agree or disagree with that statement. I see myself as someone who...

Question Type: Matrix

Options:
  1 = Disagree strongly
  2 = Disagree a little
  3 = Neither agree nor disagree
  4 = Agree a little
  5 = Agree strongly
1. Is talkative
Answer: 1 - Disagree strongly
2. Tends to find fault with others
Answer: 1 - Disagree strongly
3. Does a thorough job
Answer: 5 - Agree strongly
4. Is depressed, blue
Answer: 1 - Disagree strongly
5. Is original, comes up with new ideas
Answer: 4 - Agree a little
6. Is reserved
Answer: 4 - Agree a little
….
Please consider the following product category: Detergents - Heavy Duty -
Liquid. Suppose you are in a grocery store, and you see the following
product in that category: Purex Liquid Laundry Detergent Plus OXI, Stain
Defense Technology, 128 Fluid Ounces, 85 Wash Loads. The product is priced
at: $5.98. Would you or would you not purchase this product?
Question Type: Single Choice
Options:
  1 - Yes, I would purchase the product
  2 - No, I would not purchase the product
Answer: 1 - Yes, I would purchase the product
```

Persona Summary

Each twin in the full persona dataset averages about 30K tokens, making its simulation relatively slow and costly. To address this, we developed a shorter and more concise version called persona summary. In this version, we simplify the questions and record responses using distributional information across participants. For instance, instead of including the full set of Big Five questions, we represent the trait with a single value such as "score_extraversion = 2.125 (26th percentile)."

For simplification, wave 4 questions (which involved treatment-condition randomization) are excluded from the persona summary. As a result, the persona summary prompt is about 3K tokens in length—a substantial reduction compared to the full persona. Our results show that persona summary can serve as an efficient alternative when token and time budgets are constrained.

The complete persona dataset can be accessed at https://huggingface.co/datasets/LLM-Digital-Twin/Twin-2K-500, specifically the `persona_summary` split within the `full_persona` subset. An example of (portions of) persona summary is below:

```
The person's demographics are the following...

Geographic region: Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN,
OH)

Gender: Female

Age: 30-49

Education level: Some college, no degree

Race: Black

Citizen of the US: Yes

Marital status: Married

Religion: Nothing in particular

Religious attendance: A few times a year

Political affiliation: Independent

Income: $50,000-$75,000

Political views: Moderate

Household size: 3

Employment status: Part-time employment


The person's Big 5 scores are the following:

score_extraversion = 2.125 (26th percentile)

score_agreeableness = 4.556 (84th percentile)

wave1_score_conscientiousness = 4.556 (77th percentile)

score_openness = 3.5 (36th percentile)

score_neuroticism = 1.5 (15th percentile)

Openness reflects curiosity and receptiveness to new experiences,
Conscientiousness indicates self-discipline and goal-directed behavior,
```

Extraversion measures sociability and assertiveness, Agreeableness reflects compassion and cooperativeness, and Neuroticism captures emotional instability and susceptibility to negative emotions. Each score ranges from 1 to 5, and a higher score indicates a greater display of the associated traits.

....

Demographics Persona

We also construct a demographics persona, which contains only demographic information. This version is derived by truncating the full persona to retain only 14 demographic variables: region, sex, age, education, race, citizenship, marital status, religion, religious attendance, political party, household income, political ideology, household size, and employment status. By design, the demographics persona always serves as a prefix of the full persona. An example is shown below:

Which part of the United States do you currently live in?

Question Type: Single Choice

Options:

1 - Northeast (PA, NY, NJ, RI, CT, MA, VT, NH, ME)

2 - Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH)

3 - South (TX, OK, AR, LA, KY, TN, MS, AL, WV, DC, MD, DE, VA, NC, SC, GA, FL)

4 - West (WA, OR, ID, MT, WY, CA, NV, UT, CO, AZ, NM)

5 - Pacific (HI, AK)

Answer: 2 - Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH)

What is the sex that you were assigned at birth?

Question Type: Single Choice

Options:

1 - Male

2 - Female

Answer: 2 - Female

How old are you?

Question Type: Single Choice

Options:

```
  1 - 18-29
  2 - 30-49
  3 - 50-64
  4 - 65+
Answer: 2 - 30-49

What is the highest level of schooling or degree that you have completed?
Question Type: Single Choice
Options:
  1 - Less than high school
  2 - High school graduate
  3 - Some college, no degree
  4 - Associate's degree
  5 - College graduate/some postgrad
  6 - Postgraduate
Answer: 3 - Some college, no degree

What is your race or origin?
Question Type: Single Choice
Options:
  1 - White
  2 - Black
  3 - Asian
  4 - Hispanic
  5 - Other
Answer: 2 – Black
…
```

We integrate the constructed personas into the prompt template (see Methods section) to perform our LLM simulations.

BENCHMARKING

We perform paired t-tests to compare each pair of benchmarks on each metric. For correlation we apply a z-transformation before conducting the t-test. We also set correlation to 0 for cases where there is no variation among twin answers, which may happen with the “empty persona” or “demographics only” benchmarks.

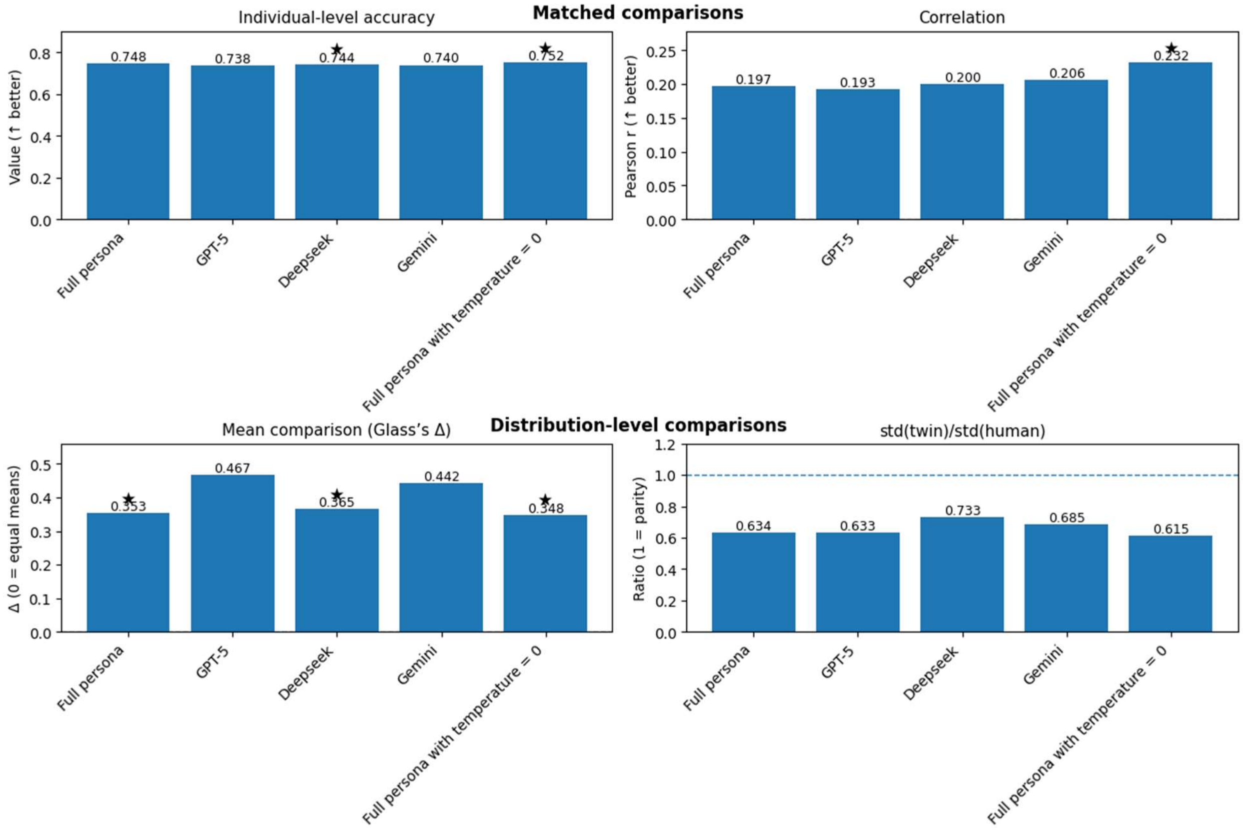


Figure SI.8: Benchmarking. *: best performing benchmark, or not significantly different from best at p<0.05 (not applicable to ratio of standard deviations).

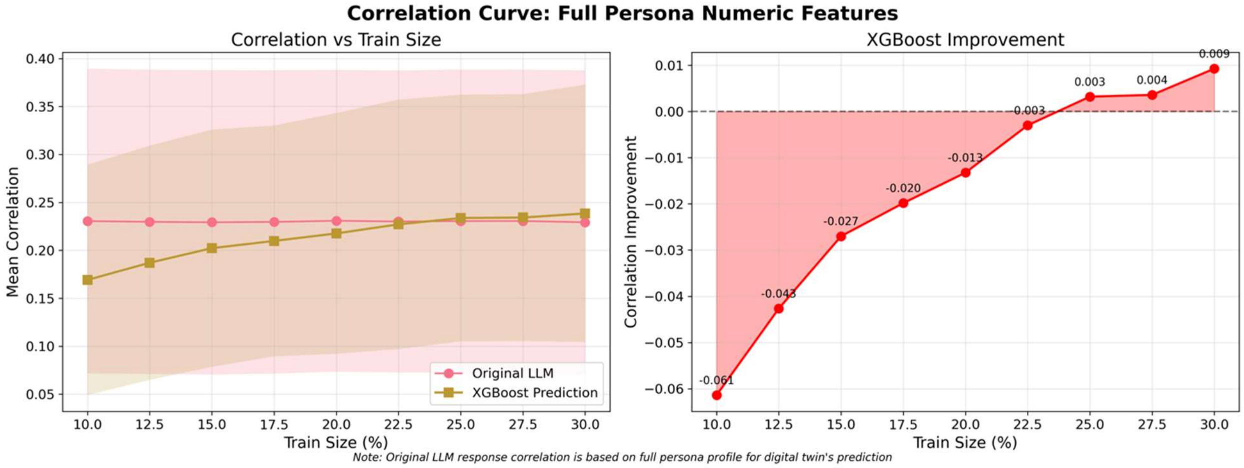
COMPARISON WITH TRADITIONAL MACHINE LEARNING APPROACH

When answering a new question not included in the Twin-2k-500 survey, our LLM simulations with temperature zero achieves an average correlation of 0.232. Traditionally, before the advent of LLMs, a standard approach was to collect a subset of real human responses for the new question and train a machine learning model to predict outcomes. A natural question, then, is: how much human data is required under the traditional approach to match the performance of the LLM?

To evaluate this, for each of the 164 outcomes, we partitioned the twins into disjoint training and test sets, varying the training fraction from 10% to 30%. We then trained an XGBoost model on the training set using 507 numeric features derived from the full persona profile for predicting the ground-truth answer. The hyper-parameters are optimized to avoid overfitting.

Figure SI.9 compares correlations on the held-out test set between human responses and (i) LLM outputs and (ii) XGBoost predictions. As the training fraction increases, XGBoost performance improves and only slightly surpasses the LLM baseline once the labeled share reaches approximately 24%.

This finding underscores the value of LLM simulation: in terms of predictive correlation, digital twin output is equivalent to approximately 24% of real human data when predicting out-of-sample outcomes in our study.



Note: Original LLM response correlation is based on full persona profile for digital twin's prediction

Figure SI.9. Correlation Achieved by Traditional Machine Learning Method when Training Sample Increases.

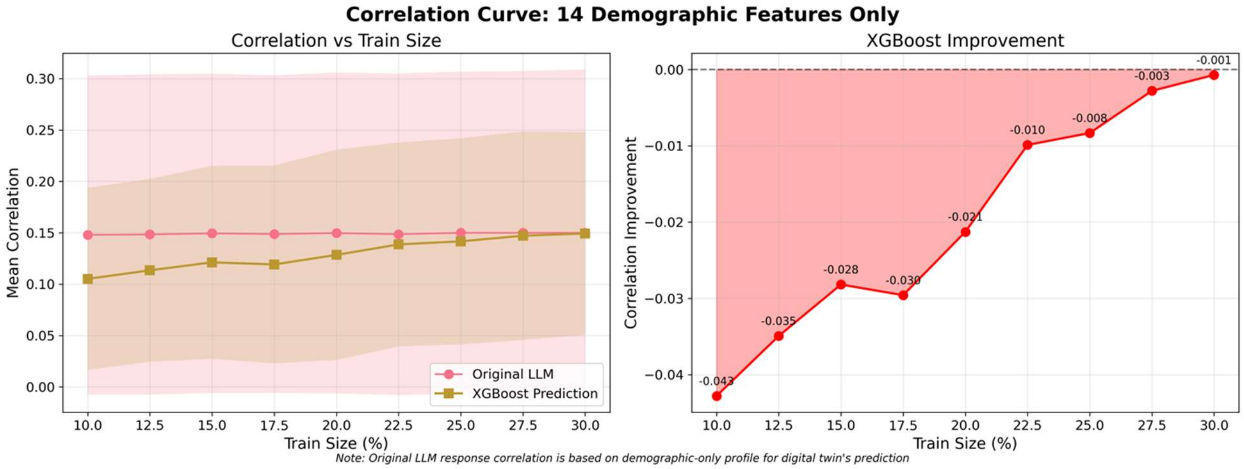
In addition, we examined the performance of XGBoost using only demographic variables (14 variables). This approach may help further regularize against overfitting and serve as a benchmark for assessing the added value of richer data and LLM-based simulations.

Under the same setup as above, XGBoost with demographic variables achieved a correlation of 0.15 when trained on 30% of human responses, and its performance plateaued beyond that point. This is comparable to the performance of LLM simulations using demographic-only persona information.

These results further reveal some interesting findings:

1. The LLM, when provided with full persona information, can predict human behavior with a correlation above 0.23. Notably, even when trained on substantial human response data, a traditional machine learning model using demographics alone cannot approach this level of performance (saturating at ~0.15).

2. The LLM with full persona performs on par with XGBoost trained on full persona, and the LLM with demographics-only persona matches XGBoost trained on demographics only. The key difference is that the LLM achieves this without requiring any new human training data, whereas XGBoost depends on sufficient additional data collection. This highlights the unique advantage of LLMs in directly leveraging persona information.



Note: Original LLM response correlation is based on demographic-only profile for digital twin's prediction

Figure SI.10. Correlation Achieved by Traditional Machine Learning Method when Training Sample Increases.

ADDITIONAL PARTIAL DEPENDENCE PLOTS FROM XGBOOST

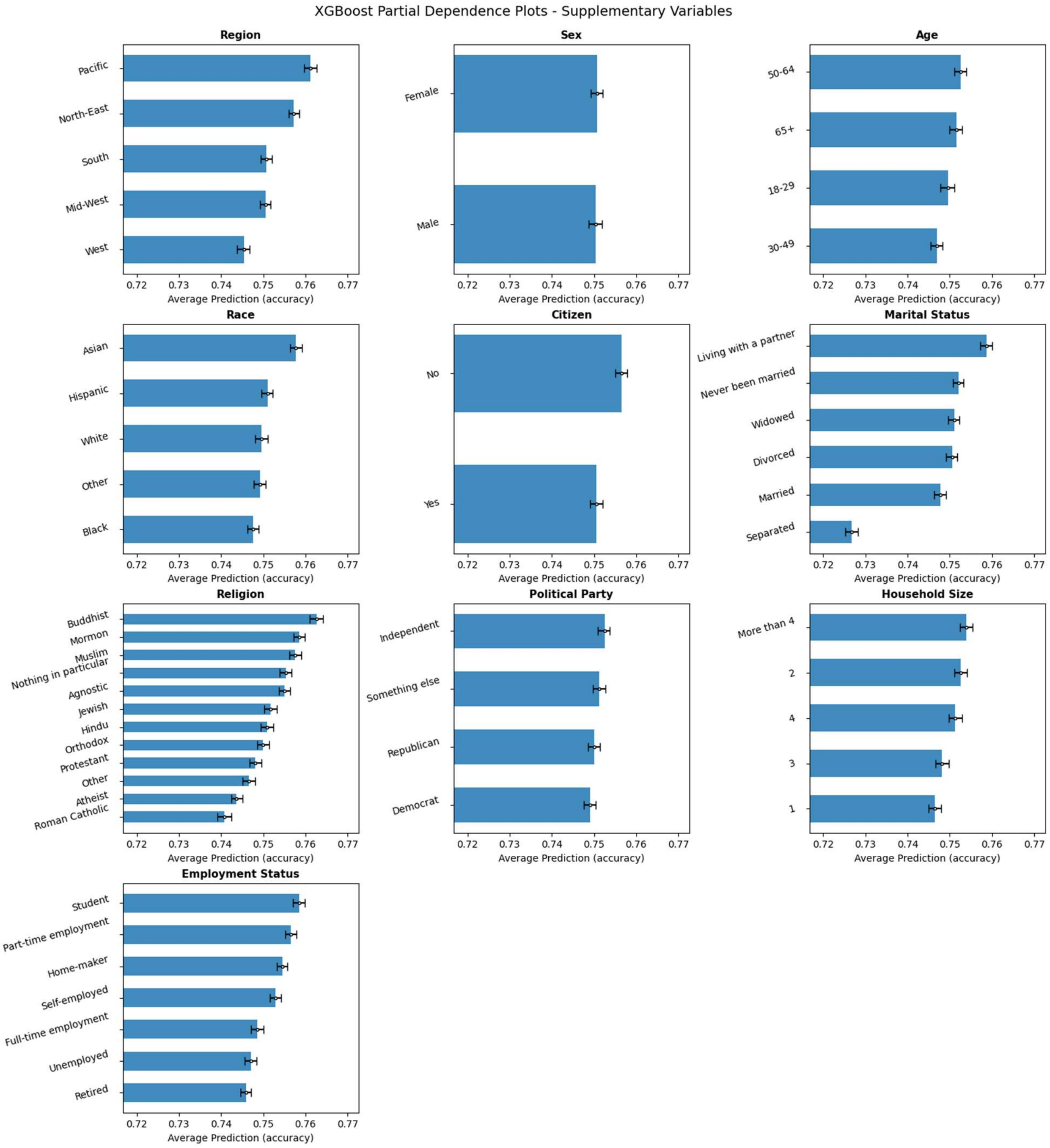


Figure SI.11. Partial Dependence Plots from XGBoost.

DETAILED INFORMATION ON EACH SUB-STUDY

All the studies were conducted on Prolific with human participants. Each study was subsequently run on their corresponding digital twins. Respondents and their twins were always assigned to the same condition(s) ensuring a fair 1-to-1 comparison.

Context Effects

Main Questions/Hypotheses

Context effects are of great interest to marketers and policy makers alike because they suggest that product assortment can change choice. This study focuses on two classic context effects. The attraction effect occurs when a third option - an asymmetrically dominated decoy - is added to a binary choice set, increasing preference for the option that dominates the decoy (Huber et al., 1982). The compromise effect occurs when a different type of decoy is added to a binary choice set, one that makes one of the original options appear as a compromise between the other two, increasing preference for the compromise option (Simonson, 1989).

This study tests whether digital twins trained from LLMs can accurately predict individual choices in product purchase settings designed to elicit the attraction and compromise effects, using newly developed stimuli that twin models could not have encountered during training.

Methods

Following an initial practice trial, all participants completed via Qualtrics three hypothetical product purchase decisions: one binary choice, and two trinary choices designed to elicit the attraction and compromise effects, respectively. The order of the three trials was counterbalanced across participants.

In each trial, participants were presented with a hypothetical purchase scenario and asked to choose between product options described in text by their price and quality features. Each trial was randomly assigned to one of the following three product categories: a printer, a TV, and a telephone plan. Importantly, all stimuli were newly developed for this study and could not have appeared in the digital twins' training data.

501 human participants (48% female, 52% male; 9% aged 18-29, 33% aged 30–49, 35% aged 50–64, 23% aged 65+) on Prolific completed the task. The study was subsequently run on their 501 digital twins. Human participants and their corresponding digital twins completed the same trials.

We recorded both choices and time spent on each decision for human participants, and only choice for digital twins.

Results - Pre-registered Analyses

We began our analysis by testing whether humans and digital twins replicated the attraction and compromise effects. As preregistered, we excluded practice trials and combined the datasets from both humans and digital twins and created customized contrasts: for the attraction effect, attraction trials were coded as 1, binary trials as -1, and all other trials as 0. A similar coding scheme was used for the compromise effect. We ran binomial mixed-effect models predicting the choice of the target option. We included dataset and product type as fixed effects and accounted for the multilevel structure of the data by modeling respondent-level random intercepts, treating each human and their digital twin as separate entities. Each contrast was nested within the human and digital twin datasets to separately capture the effects within each dataset.[1]

As shown in Figure 1, for the attraction effect, we replicated prior literature and found a significant effect among humans (*β* = 0.18, *SE* = 0.07, *z* = 2.56, *p* = .01). However, we did not detect an effect among digital twins (*β* = -0.24, *SE* = 0.14, *z* = -1.67, *p* = .10). Unexpectedly, we observed a strong main effect of dataset, such that digital twins made substantially more target choice overall (*β* = 2.84, *SE* = 0.16, *z* = 17.30, *p*< .001).

For the compromise effect, neither humans nor digital twins replicated prior results. No effect was found for humans (*β* = 0.06, *SE* = 0.07, *z* = 0.87, *p* = .38), and a significant negative effect was found for digital twins (*β*= -0.54, *SE* = 0.15, *z* = -3.59, *p* < .001). Again, we observed a strong main effect of dataset, with digital twins making substantially more target choices (*β* = 2.90, *SE* = 0.17, *z* = 17.16, *p* < .001).

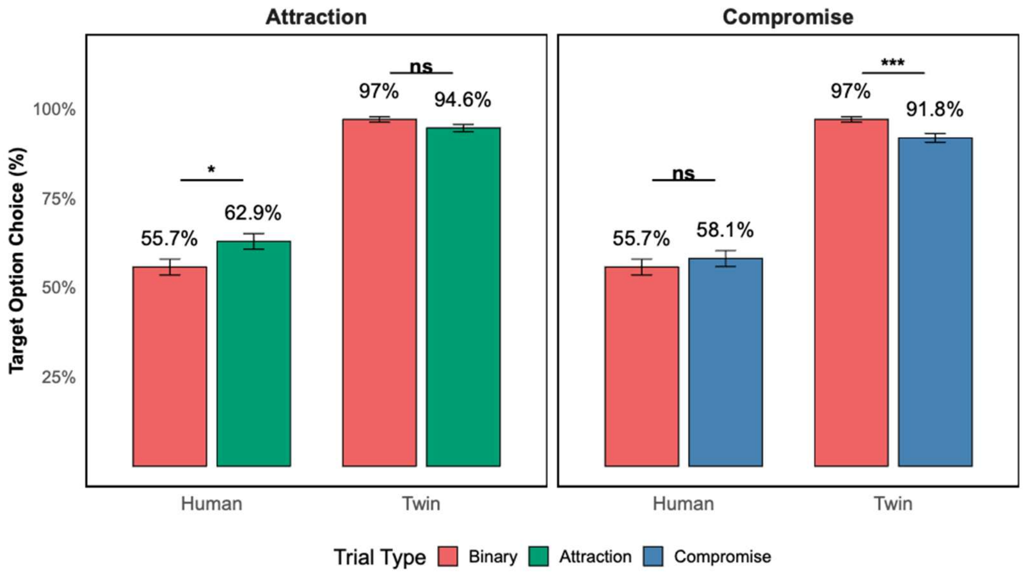


Figure 1. The target option refers to the option whose choice likelihood is expected to increase due to the addition of a decoy option. Vertical bars represent ±1 standard deviation based on raw data. Asterisks indicate significance levels derived from model-based contrast analyses: * p < .05, ** p < .01, *** p < .001; ns = not significant.

[1] The nested model is a reparameterization of the preregistered interaction model. It is statistically equivalent but provides a more direct and interpretable test of the effect in each dataset.

We also evaluated how accurately digital twins predicted human choices, independent of whether they predicted context effects. To examine possible variation in prediction accuracy by trial type, we broke the analysis down across binary, attraction, and compromise trials. For each trial, we created a binary variable denoting whether the digital twin correctly predicted the participant's choice.[2] As a benchmark, we compared digital twin's accuracy level to that of a baseline model that randomly selected among the available options.

We found that digital twins predicted human choices more accurately than the random baseline across all trial types. Specifically, twins achieved 58% accuracy on binary trials (paired *t*-test against 50% baseline: *t*(500) = 2.86*, p* = .004), 59% on attraction trials (against 33%, *t*(500) = 8.52*, p* < .001), and 55% on compromise trials (against 33%, *t*(500) = 8.04, *p* < .001). Results were consistent when using McNemar's test, which treats the dichotomous accuracy indicator as a nominal variable, instead of the *t*-test.

Results - Additional Analyses (Non-Preregistered)

Using response time data from human trials, we explored whether attention moderated the observed context effects. We constructed an attention factor based on log-transformed response times from binary, attraction, and compromise trials, and included interactions between this factor and the contrast variables in the nested model.

Among human participants, attention significantly moderated the compromise effect (*β* = 0.17, *SE* = 0.07, *z* = 2.32, *p* = .02), such that more attentive participants exhibited a significant compromise effect: a Johnson-Neyman analysis indicated the effect became statistically significant at values above 0.61 standard deviations of attention (relative to the mean), under a p < .05 threshold. No interaction was observed for the attraction effect (*β* = 0.08, *SE* = 0.07, *z* = 1.06, *p* = .29).

Attention factor derived from human participants did not moderate either context effect in the twin dataset (attraction: *β* = -0.002, *SE* = 0.14, *z* = -0.01, *p* = .99; compromise: *β* = –0.12, *SE* = 0.14, *z* = -0.81, *p* = .42).

Discussion

We examined if large language models (LLMs) like ChatGPT can accurately predict the choices made by humans. In the current study, we successfully replicated the attraction effect in the human dataset, but digital twins did not reproduce this effect. Neither humans nor digital twins exhibited a compromise effect. However, the compromise effect was significant among human participants who paid more attention, as measured by response time. Independent of predicting context effects, digital twins showed only modest accuracy (less than 60% overall) in predicting individual human choices. These findings

[2] We created an indicator variable denoting whether the digital twin correctly predicted the participant's choice and used a McNemar test for analysis, rather than the preregistered correlation, as it is more appropriate for discrete choice outcomes.

highlight cautions when using digital twins to predict human decisions in applied settings such as consumer choice.

References

Huber, J., Payne, J. W., & Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research*, *9*(1), 90. https://doi.org/10.1086/208899

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research*, *16*(2), 158. https://doi.org/10.1086/209205

Default Effects

Main Questions/Hypotheses

This study focuses on two default effects: the organ donation paradigm by Johnson and Goldstein (2003), and a green energy default paradigm promoting sustainable energy choices (Pichert et al., 2008). In the organ donation study, participants who were defaulted to be donors were significantly more likely to remain enrolled than those who had to actively opt in. Similarly, in the green energy study, defaulting participants into using green energy suppliers increased adoption rates compared to requiring them to opt in.

This study tests whether digital twins trained from LLMs can accurately predict individual choices in these two settings, using the original organ donation stimuli verbatim and a slightly adapted version of the green energy scenario.

Methods

All participants completed the original organ donation as well as the adjusted paradigm, presented in counterbalanced order. Across both default tasks, participants were randomly assigned to either an opt-in or an opt-out condition. In the organ donation task, participants in the opt-in condition were told they were not organ donors by default and were asked whether they wanted to become one. In the opt-out condition, participants were told they were organ donors by default and could choose to change that status. In the green energy default task, participants in the opt-in condition were told they were defaulted to the more affordable (non-green) energy option, while those in the opt-out condition were defaulted to the green energy option.

Notably, the organ donation task used the original stimuli from Johnson and Goldstein (2003), while the green energy default task was slightly modified by changing the company names for the affordable and sustainable options.

600 human participants (47% female, 53% male; 11% aged 18-29, 33% aged 30–49, 34% aged 50–64, 22% aged 65+) on Prolific completed the task. The study was subsequently run on their 600 digital twins. Participants and their corresponding digital twins were assigned the same conditions.

We recorded both choices and time spent on each decision for human participants, and only choice for digital twins.

Results - pre-registered analyses

We began our analysis by testing whether humans and digital twins replicated default effects in organ donation and green energy choice. As preregistered, for each default task, we combined data across both humans and digital twins and created a customized contrast: opt-out trials were coded as 1, opt-in trials as -1. We accounted for the multilevel structure of the data by including respondent-level random intercepts, treating each human and their digital twin as separate entities. This contrast was nested within human and digital twin datasets to separately estimate effects within each dataset.[3]

As depicted in Figure 1, in the green energy task, humans showed a large effect and replicated prior literature (β = 0.50, *SE* = 0.08, *z* = 5.98, *p* < .001), while digital twins did not (β = 0.14, *SE* = 0.08, *z* = 1.73, *p* = .08). We observed no main effect of dataset on green energy option adoption (β = -0.22, *SE* = 0.12, *z* = -1.86, *p* = .06).

In the organ donation task, we found detected no effect among humans (β = 0.09, *SE* = 0.08, *z* = 1.10, *p* = .28), but a large effect among digital twins (β = 0.92, *SE* = 0.11, *z* = 8.61, *p* < .001). In addition, we observed a main effect of dataset, such that digital twins were more likely to become an organ donor overall (M_{Human} = 59%, M_{Twin} = 71%, β = 0.75, *SE* = 0.14, *z* = 5.53, *p* < .001). Though this effect may be primarily driven by the difference in the opt-out condition (M_{Human} = 61%, M_{Twin} = 88%).

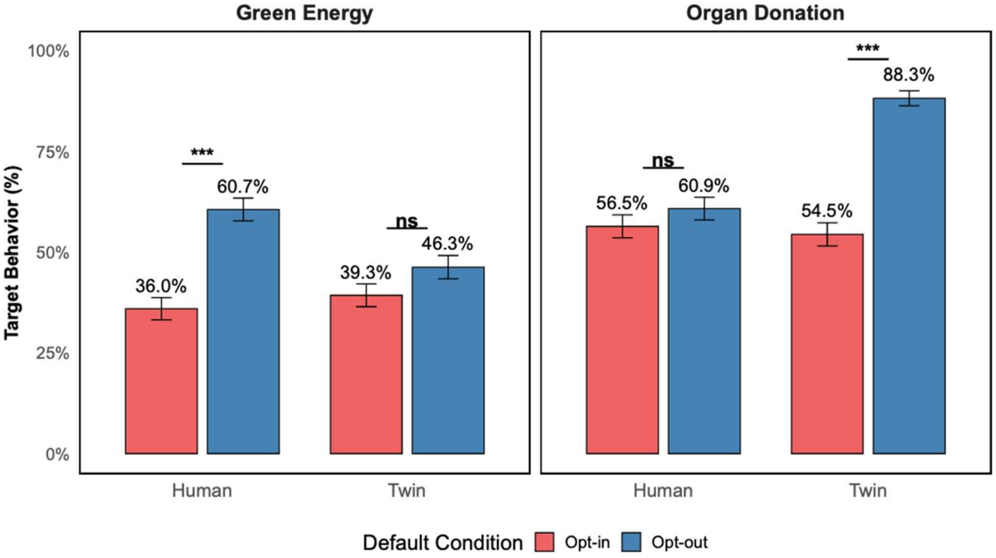


Figure 1. Target behavior refers to adopting a green energy provider in the green energy paradigm and registering as an organ donor in the organ donation paradigm. Vertical bars represent ±1 standard deviation based on raw data. Asterisks indicate significance levels derived from model-based contrast analyses: * p < .05, ** p < .01, *** p < .001; ns = not significant.

We next examined how closely digital twins predicted human choices, independent of whether they predicted the default effects. For each trial, we created an indicator variable denoting whether the digital twin correctly predicted the participant's choice.[4] As a

[3] The nested model is a reparameterization of the preregistered interaction model. It is statistically equivalent but provides a more direct and interpretable test of the effect in each dataset.

[4] We created an indicator variable denoting whether the digital twin correctly predicted the participant's choice and used a McNemar test for analysis, rather than the preregistered correlation, as it is more appropriate for discrete choice outcomes.

benchmark, we compared digital twin's accuracy level to that of a baseline model that randomly selected among the available options.

Across both paradigms, digital twins outperformed the random baseline. Specifically, twins achieved 69% accuracy on the green energy paradigm (paired *t*-test against 50% baseline, *t*(599) = 7.11, *p* < .001), and 58% on organ donation paradigm (against 50%, *t*(599) = 2.99*, p* = .003). Results were consistent when using McNemar's test, which treats the dichotomous accuracy indicator as a nominal variable, instead of the *t*-test.

Results - Additional Analyses (Non-Preregistered)

Using response time data from human trials, we explored whether attention moderated the observed context effects. We included interaction terms between the log-transformed and centered response time for each trial and the contrast variables in the nested model. We found no statistically significant moderation effects for either human participants (green energy: *β* = -0.01, *SE* = 0.09, *z* = -0.15, *p* = .88; organ donation: *β*= -0.15, *SE* = 0.09, *z* = -1.71, *p* = .09). Response time derived from the human dataset also did not moderate the effects in the digital twin models (green energy: *β* = 0.06, *SE* = 0.08, *z* = 0.77 *p* = .44; organ donation: *β* = -0.09, *SE* = 0.10, *z* = -0.90, *p* = .37).

Discussion

Large language models (LLMs) like ChatGPT can accurately describe default effects and reference the relevant literature, presumably because these concepts are well represented in their training data. Default effects, while often robust, vary considerably in magnitude (Jachimowicz et al., 2019). In the present study, human participants showed a default effect in the green energy adoption paradigm but not in the organ donation paradigm. Digital twins predicted a strong default effect in the organ donation paradigm but no effect in the green energy paradigm.

This discrepancy may stem from greater representation of the organ donation paradigm in the LLM's training data, while our modified green energy options may have diverged from commonly seen formats. Still, digital twins were able to predict human choices in the green energy paradigm with 69% accuracy, showing some promise in capturing people's underlying preferences.

References

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, *3*(02), 159–186. https://doi.org/10.1017/bpp.2018.43

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, 302(5649), 1338-1339.

Pichert, D., & Katsikopoulos, K. V. (2008). Green defaults: Information presentation and pro-environmental behaviour. *Journal of Environmental Psychology*, *28*(1), 63 73. https://doi.org/10.1016/j.jenvp.2007.09.004

Fees Accuracy

Main Questions/Hypotheses

Hidden fees and surcharges are a routine but often resented aspect of modern commerce. Whether booking a hotel, ordering food delivery, or using a credit card bill, consumers frequently encounter extra, often unexpected charges—like resort fees, service fees, foreign transaction fees, or order processing fees—added on top of the advertised base price. These fees are often disclosed late in the transaction, hindering consumers' ability to estimate total costs or compare options, and are often described using obtuse language, obscuring understanding of why the fees are assessed. As a result, they have attracted growing scrutiny from consumer advocates and regulators. In recent years, a bipartisan group of regulators in the United States and other countries have advocated for stronger pricing transparency and restrictions on the delayed disclosure of mandatory charges. In 2022, the Biden-Harris administration labeled these "junk fees," defined as "fees designed either to confuse or deceive consumers or to take advantage of lock-in or other forms of situational market power" (The White House, 2022).

Against this policy backdrop, we explore how consumers think about these fees —how knowledgeable they are on what these fees represent, how fair they perceive them to be, and how supportive they are of government regulation regarding firms' use of such fees. We also examine whether large language models (LLMs), acting as digital twins, can replicate their human counterparts in these domains.

Specifically, we ask the following questions. First, do digital twins accurately replicate human knowledge of pricing fees? Second, do they mirror human judgment about fairness? Third, do they match human attitudes toward regulation?

Methods

We recruited 400 U.S. participants from Prolific (48% female; median age group = 50–64) to complete an online survey assessing their knowledge and attitudes toward 19 different fees assessed across seven industries (e.g., hospitality, healthcare, rental housing). Each participant evaluated one randomly assigned fee from six of the seven industries. See Table 1 for the full list of industries and fees.

Industry	Fee	Definition
Hotels	Resort fee	A mandatory fee for a group of services, such as pool use, gym access, towel services, Wi-Fi, newspapers, shuttle service, daily parking, etc.
	Destination fee	A mandatory fee for a group of services, such as pool use, gym access, towel services, Wi-Fi, newspapers, shuttle service, daily parking, etc.
	Restocking fee	A fee charged by hotels for using the mini-fridge to store personal items.

Car Rentals	Vehicle licensing fee	A fee charged by rental car companies to offset costs associated with vehicle registration, licensing, and related taxes.
	Toll transponder fee	A daily fee assessed if a toll transponder is used during every day of the rental to pass quickly through toll booths.
	Concession recovery fee	A fee charged by rental car companies to customers to recover costs associated with operating at specific locations, such as airports.
	Frequent traveler program surcharge/excise tax	A fee for accumulating miles or points through credit card, airline, or other rewards programs.
Ticket Processing	Delivery fee	A fee for delivering tickets, whether physical or electronic.
	Order processing fee	A fee to cover costs associated with processing and completing ticket purchases.
	Facility charge	A fee to cover costs associated with hosting an event.
Food Delivery Apps	Service fee	A fee to operate the food delivery app platform.
	Express fee	A fee for prioritizing the matching process for delivery services.
	Regulatory response fee	A fee to offset the impact of regulations on firms.
Apartment Rentals	Valet trash fee	A fee for trash and recycling collection services.
	Move-in fee	A non-refundable fee to cover move-in and administrative costs, such as record updates, lock changes, and unit preparation.
	Amenity fee	A recurring fee that covers additional services and amenities, such as pools, gyms, parking, or other building facilities.
Health Care	Hospital facility fee	A fee charged by a private medical practice owned by a hospital to cover the hospital's operational and maintenance costs, including equipment, staff, and utilities.
Credit Cards	Foreign transaction fee	A fee for purchases or transactions in a foreign currency to cover conversion and international processing costs.
	Balance transfer fee	A percentage fee that covers debt transfers from one credit card to another.

Table 1. Industries, Fees, and Definitions Used in the Survey.

For each fee, we assessed the following. First, we assessed participants' objective knowledge by asking them to select the best definition of a fee from four options (e.g., "Which of the following do you think best represents what a resort fee is assessed for?"). We coded responses as 1 if correct and 0 if incorrect. We then averaged participants' responses across the six fees they evaluated to create an overall accuracy score. Immediately after responding, participants were informed whether their answer was correct and were shown the correct definition. Second, we assessed participants' familiarity with each fee. Specifically, participants indicated whether they had previously heard of each fee (e.g., they responded to questions like "Have you ever heard of a resort fee before seeing it in this survey?"). These items capture prior experience and are part of a separate project and were not pre-registered. Third, we assessed participants' fairness perceptions. Participants rated how fair they considered each fee (e.g., they responded to questions like "How fair do you think it is for hotels to charge for a resort fee?") on a 7-point scale (1 = Very unfair; 7 = Very fair). We then averaged participants' responses across the six fees they evaluated to create an overall fairness rating. Finally, we assessed participants' attitudes regarding policy. Specifically, after completing the fee evaluations, participants responded to several items about their general attitudes toward pricing practices and regulation. These included two pre-registered items: "Should pricing practices be regulated by the government?" rated on a 7-point scale (1 = Strongly oppose regulation; 7 = Strongly support regulation), and "How likely would you be to support government regulation that bans firms from separating out mandatory fees from base prices?" rated on a 7-point scale (1 = Very unlikely; 7 = Very likely). Responses to these two items were highly correlated (r = .72), so we averaged them to create a composite measure of support for regulation.

Each participant's responses were paired with those of an LLM-generated digital twin prompted to respond as that individual would.

Results - Pre-registered Analyses

Knowledge Accuracy. Digital twins dramatically outperformed human participants in identifying correct definitions of fees, based on a paired *t*-test ($t(399) = 45.89$, $p < .001$, $d = 3.26$). Humans answered 51.75 percent of fee definition questions correctly (SD = 20.87%), while digital twins answered 99.88% correctly (SD = 1.86%).

Fairness. Despite the accuracy gap, average fairness ratings were virtually identical. Both human participants ($M = 3.28$, $SD = 1.24$), and their digital twins ($M = 3.27$, $SD = 0.62$) rated the fees as moderately unfair, and their ratings did not significantly differ based on a paired-sample t-test ($t(399) = 0.14$, $p = .886$, $d = 0.01$). Table 2 reports fee-specific fairness comparisons.

Fee	Human M (SD)	Twins M (SD)	Paired t-test
Hotel Fees			
Resort Fee	3.04 (1.80)	2.95 (0.81)	*t*(131) = .53, *p* = .596, *d* = .06
Destination Fee	2.93 (1.68)	3.12 (0.81)	*t*(133) = 1.24, *p* = .215, *d* = .15
Restocking Fee	2.29 (1.61)	2.23 (0.50)	*t*(133) = .44, *p* = .660, *d* = .05
Car Rental Fees			
Vehicle Licensing Fee	3.44 (1.79)	3.71 (0.86)	*t*(100) = 1.46, *p* = .149, *d* = .20
Toll Transponder Fee	4.15 (1.83)	3.27 (0.83)	*t*(99) = 4.44, *p* < .001, *d* = .62
Concession Recovery Fee	3.29 (1.70)	3.34 (0.81)	*t*(98) = .29, *p* = .769, *d* = .04
Frequent Traveler Program Surcharge/Excise Tax	2.43 (1.50)	2.67 (0.67)	*t*(99) = 1.65, *p* = .101, *d* = .20
Ticket Processing Fees			
Delivery Fee	2.56 (1.56)	3.61 (0.91)	*t*(130) = 7.05, *p* < .001, *d* = .82
Order Processing Fee	3.05 (1.85)	3.48 (0.87)	*t*(134) = 2.83, *p* = .005, *d* = .28
Facility Charge	3.36 (1.80)	3.51 (0.82)	*t*(133) = .93, *p* = .355, *d* = .10
Food Delivery App Fees			
Service Fee	3.91 (1.78)	3.86 (0.77)	*t*(133) = .33, *p* = .744, *d* = .04
Express Fee	3.64 (1.77)	3.00 (0.76)	*t*(132) = 4.10, *p* < .001, *d* = .46
Regulatory Response Fee	2.69 (1.66)	2.59 (0.57)	*t*(132) = .72, *p* = .472, *d* = .08
Apartment Rental Fees			
Valet Trash Fee	3.66 (1.88)	3.36 (0.96)	*t*(133) = 1.73, *p* = .086, *d* = .20
Move-In Fee	3.10 (1.71)	3.58 (0.83)	*t*(133) = 2.89, *p* = .004, *d* = .36
Amenity Fee	3.70 (1.66)	3.55 (0.80)	*t*(131) = 1.00, *p* = .320, *d* = .11
Health Care Fees			
Hospital Facility Fee	3.28 (1.94)	2.81 (0.79)	*t*(133) = 2.80, *p* = .006, *d* = .30
Credit Card Fees			
Foreign Transaction Fee	4.30 (1.69)	3.86 (0.93)	*t*(131) = 2.76, *p* = .007, *d* = .33
Balance Transfer Fee	3.56 (1.87)	3.63 (0.78)	*t*(133) = .46, *p* = .643, *d* = .05

Table 2. Per-Fee Fairness Comparisons Between Human and AI (paired-samples).

We also compared the variation in fairness judgments between humans and digital twins. Humans expressed significantly more variation in their fairness ratings than did their digital twins, based on a Levene's test (*F*(1, 798) = 180.48, *p* < .001).

Policy Attitudes. We regressed support for regulation on participants' ideology (mean-centered, 1 = very liberal to 5 = very conservative), source (Human vs. AI), and their interaction. The main effect of political ideology was significant (*b* = –0.44, *SE* = 0.06, *t*(796) = –7.78, *p* < .001), such that more conservative participants expressed lower support for regulation. The main effect of source was also significant (*b* = 0.44, *SE* = 0.09, *t*(796) = 4.69, *p* < .001), with AI twins showing greater support for regulation than their human counterparts. Finally, the interaction between political ideology and source was significant (*b* = –0.53, *SE* = 0.08, *t*(796) = –6.67, *p* < .001), demonstrating that the ideology effect was stronger for the digital twins than their human counterparts (see Figure 1).

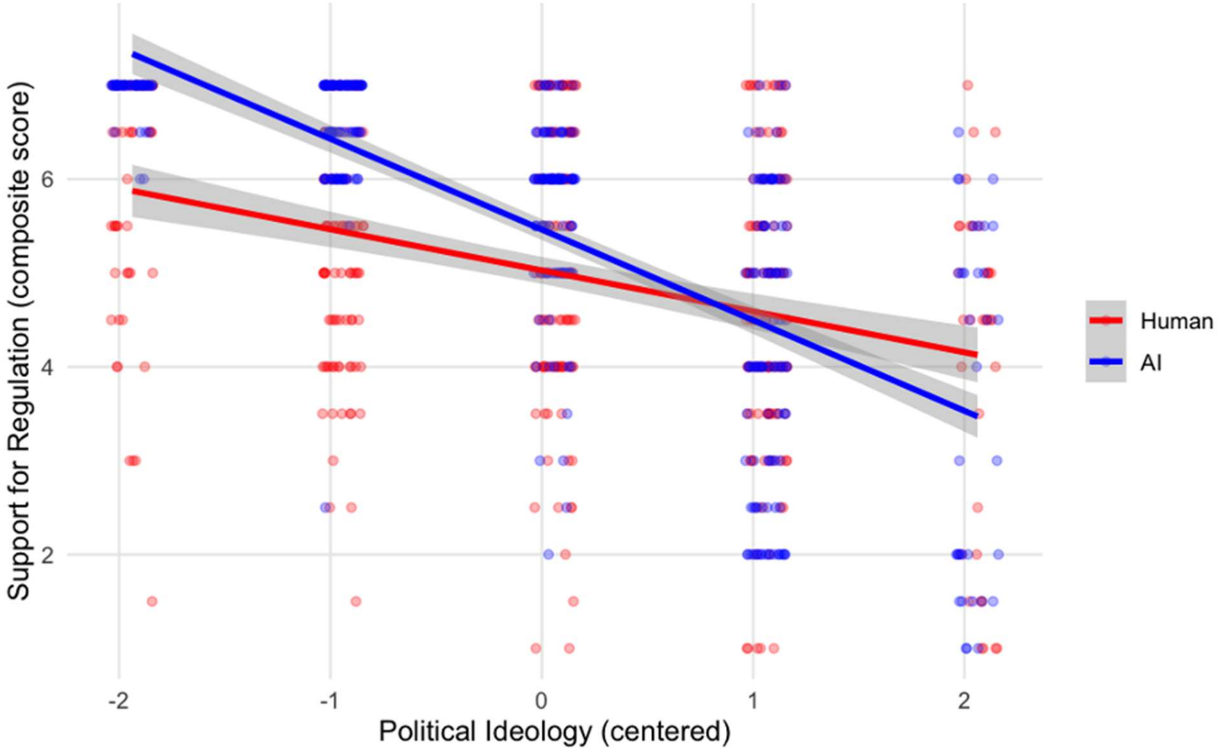


Figure 1. Support for Regulation as a Function of Political Ideology and Participant Source.

Discussion

Our findings reveal a striking contrast: digital twins excel at factual knowledge (i.e., knowledge of fee definitions) and fall short of mimicking the error-prone, heterogeneous reasoning of human consumers. This limitation underscores a critical boundary in using LLMs as behavioral research surrogates: when tasks depend on limited knowledge, ambiguity, or lived experiences, digital twins may produce unrealistically accurate responses.

In contrast to the large difference in accuracy, on fairness judgments, digital twins closely mirrored the average human response, but exhibited narrower variance. This suggests that while LLMs can approximate central tendencies, they may underrepresent the subjective diversity of human opinion.

In the domain of policy attitudes, digital twins again tracked human performance. For both groups, conservatism was correlated with less support for pricing regulation, replicating prior work on ideological influences in policy attitudes (e.g., McCright et al., 2013). However, digital twins exhibited exaggerated ideological consistency with a stronger negative association between conservatism and regulation support than human participants. This likely reflects LLM's reliance on dominant associations in their training data (e.g., conservatism = free-market values), whereas human responses may incorporate more contextual nuance.

Together, these results offer promise and caution. The results suggest that LLMs can simulate certain aspects of consumer judgment, but may misrepresent human variability and error, especially when tasks require subjective or uncertain reasoning. Because LLMs are trained to prioritize factual accuracy and statistical associations from vast textual corpora, they often struggle to suppress correct information, even when explicitly instructed to simulate human errors, misconceptions, or gaps in knowledge.

References

McCright, A.M., Dunlap, R.E. & Xiao, C. Perceived scientific agreement and support for government action on climate change in the USA. *Climatic Change* 119, 511–518 (2013). https://doi.org/10.1007/s10584-013-0704-9

The White House. (2022, October 26). The President’s initiative on junk fees and related pricing practices. The White House. https://www.whitehouse.gov/briefing-room/blog/2022/10/26/the-presidents-initiative-on-junk-fees-and-related-pricing-practices/

Measures of Creativity

Main Questions/Hypotheses

There is a growing body of literature on the creative capabilities of LLMs. This literature typically shows that LLMs are able to generate ideas that are judged as creative or more creative than ideas generated by humans (e.g., Boussioux et al., 2023; Lee and Chung, 2024; Zhang and Gosline, 2023). The first question this study seeks to address is: How do digital twins' creative abilities compare to those of their human counterparts?

Second, this study explores whether the creative tests that are often administered to humans to assess their creative skills may also assess the creative skills of their digital twins. That is, our second research question is: Can digital twins replicate the correlations between creativity tests and idea generation performance observed in humans?

Methods

We collected complete data from 200 human participants on Prolific. This study is coupled with an "idea evaluation" sub-study that requires a sample size six times as large as the "measures of creativity study." We pre-registered a sample size of 250 for "measures of creativity" and 1,500 for "idea evaluation." Based on observations from earlier sub-studies, we deviated from pre-registration to reduce the sample sizes to 200 and 1,200, which seemed more achievable. The "measures of creativity" study was subsequently run on the 200 digital twins of the human participants.

This study had a single condition and consisted of three tasks:

1. Divergent Association Task (DAT, Olson et al., 2021), which measures divergent thinking abilities (single task - "Please write 10 words that are as different from each other as possible, in all meanings and uses of the words."),
2. Shortest Semantic Path Task (SSPT, Toubia and Berger, 2025), which may be viewed as a measure of convergent thinking abilities (5 tasks, using the following pairs of seed words: eternity-curiosity, elephant-galaxy, perseverance-eloquence, euphoria-tulip, tangerine-penguin: "Find a way to connect these two words: <seed1> and <seed2>. Each word in the sequence below should be as closely related as possible to the word before it. The first word in the sequence is already set to <seed1>. The last word is already set to <seed2>. That is, you only need to add 3 words that connect the first word to the last word."),
3. An idea generation task asking for a new idea (at least 200 characters) for smartphone apps that will help their users keep a healthier lifestyle (single task: "HOW COULD SMARTPHONES HELP THEIR USERS BE HEALTHIER? We are interested in new ideas for smartphone apps that will help their users keep a healthier lifestyle. In the space provided below, please enter one new idea for a smartphone app priced at $0.99 that would help its users keep a healthier lifestyle. Please be as specific as possible and describe the main features of the app. Please enter exactly one detailed idea in the box below. Your idea should contain at least 200 characters. ").

The order of the first two tasks was randomized. The first two tasks were scored automatically, consistent with the original papers. The DAT was scored by computing the average pairwise semantic (cosine) distance between the first 7 words in the sequence. The SSPT was scored by computing the circuitousness corresponding to the responses to each task by each participant, where circuitousness is the ratio between the total semantic (Euclidean) distance of the path consisting of the five words in the sequence to the distance corresponding to the shortest path that starts with the same word, ends with the same words, and optimally orders the three words in between. The performance at the individual level was obtained by averaging performance across the five tasks for each individual (human or twin). The performance at the idea generation task was measured using the average creativity rating given to the idea by humans in the "baseline" condition in the "idea evaluation" study.

In addition, this study leverages the Forward Flow measure (Gray et al., 2019) collected in the original Twin2K500 dataset. Performance on this task is measured as the average pairwise semantic distance between the words in the sequence.

Results - Pre-registered Analyses

First, we compare the creativity ratings of the ideas generated by humans vs. their digital twins. We find that the creativity ratings (coming from human participants) of the ideas from digital twins were significantly higher than those generated by their human counterparts (M_{twins}=3.322, M_{humans}=3.198, t=2.718, p<0.001). The mean absolute percentage difference between the two creativity ratings was 0.177. The standard deviation of the ratings obtained by ideas from digital twins was also lower (S_{twins}=0.395, S_{humans}=0.530, F(198,198)=1.816, p<0.001). The correlation between the creativity ratings of ideas from digital twins vs. humans was not significant (r=0.090, p=0.204).

The twins performed significantly better also on the DAT (M_{twins}=0.897, M_{humans}=0.852, t=12.168, p<0.001). The mean absolute percentage difference between humans and twins on that task was 0.064. The standard deviation of the digital twins scores was lower (S_{twins}=0.018, S_{humans}=0.050, F(198,198)=7.957, p<0.001). The correlation between the score of digital twins vs. humans was not significant (r=0.068, p=0.343).

Digital twins did not perform significantly differently at the SSPT (M_{twins}=1.012, M_{humans}=1.013, t=1.044, p=0.0298). The mean absolute percentage difference between humans and twins on that task was 0.012. The standard deviation of the digital twins scores was lower (S_{twins}=0.010, S_{humans}=0.014, F(198,198)=2.053, p<0.001). The correlation between the score of digital twins vs. humans was not significant (r=0.059, p=0.403).

Second, we explore how the correlation between various measures differs from twins vs. humans. For this exercise, as pre-registered, we use the human (digital twins) ratings of creativity for human (digital twins) ideas.

Within humans, we correlate performance at the Forward Flow, DAT, SSPT and idea generation tasks. Consistent with previous literature, we find a significantly positive correlation between the Forward Flow score and idea generation performance (r=0.140, p<0.05) as well as between DAT and idea generation performance (r=0.178, p<0.05). Although we replicate the correlation between SSPT and idea generation performance directionally, it is not statistically significant (r=-0.097, p=0.175).

Within twins (with ideas rated by twins), the correlation between DAT and idea generation performance (r=-0.002, p<0.974) and the correlation between SSPT and idea generation (r=0.082, p=0.243) are not significant. Note that Forward Flow is only available for humans (it was part of the Twin-2K-500 training data and hence is part of the twin's training data), so we cannot replicate the correlation between Forward Flow and other measures on twins.

We compare the correlations within humans to the correlations within twins. None of the correlations (DAT vs. SSPT, DAT vs. idea generation, SSPT vs. idea generation) are significantly different among humans vs. twins.

Results - Additional Analyses (Non-Preregistered)

The results from the "idea evaluation" study suggest that digital twins are not a reliable source of creativity ratings. Therefore, we also compute the correlation between DAT and idea generation performance among twins, but with creativity ratings coming from the humans rather than the twins. When creativity is assessed by humans, we do find a significant correlation between DAT performance for twins (rated automatically based on word embeddings) and idea generation performance rated by humans (r=0.207, p<0.01). The correlation between SSPT (also rated automatically based on word embeddings) and idea generation performance rated by humans is still not significant (r=0.070, p=0.322).

Discussion

We find that the ideas generated by digital twins were judged as more creative (and with less dispersion in creativity ratings) than the ideas generated by their human counterparts. Digital twins also performed better (and with less dispersion) at the DAT, a creativity task. We were able to replicate correlations previously found between Forward Flow and DAT and creativity among humans. While we did not administer Forward Flow to twins as this task was part of their training, we replicate the correlation between DAT and creativity on the idea generation task among twins, but only when creativity is evaluated by humans, not by twins.

References

Boussioux, Leonard, et al. "The crowdless future? generative ai and creative problem solving." Organization Science ja (2023).

Gray, Kurt, et al. "“Forward flow”: A new measure to quantify free thought and predict creativity." American Psychologist 74.5 (2019): 539.

Lee, Byung Cheol, and Jaeyeon (Jae) Chung. “An Empirical Investigation of the Impact of ChatGPT on Creativity.” (2024). Nature Human Behavior. 08/12/2024

Olson, Jay A., et al. "Naming unrelated words predicts creativity." Proceedings of the National Academy of Sciences 118.25 (2021): e2022340118.

Toubia, Olivier, and Jonah Berger. “The Ability to Perform Efficient Semantic Search Predicts Creativity” (2025). Working paper.

Zhang, Yunhao, and Renée Gosline. "Human favoritism, not AI aversion: People’s perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation." Judgment and Decision Making 18 (2023): e41.

Idea Evaluation

Main Questions/Hypotheses

First, we explore whether digital twins are able to replicate creativity ratings from their human counterparts: How do digital twins' ratings of idea creativity compare to those of their human counterparts?

Second, research has shown that humans tend to show AI aversion or human favoritism when evaluating the creativity of ideas (Zhang and Gosline, 2023), i.e., giving higher creativity ratings to ideas coming from humans when informed that the ideas come from humans (as opposed to AI). We explore whether the reverse is true for digital twins, i.e., do digital twins show human favoritism, AI favoritism, or neither when evaluating ideas?

Methods

We used a 2 (human ideas vs. twin ideas) x 3 (baseline vs. partially informed vs. informed) design, targeting N=200 humans and their twins per condition. The ideas came from the "measures of creativity" sub-study, which generated 200 ideas from humans and 200 ideas from their digital twins. Participants in the "measures of creativity" sub-study were excluded from the "idea evaluation" sub-study. Each participant was randomly assigned to evaluating ideas from humans or twins and evaluated 20 ideas randomly selected from that set. Participants were further assigned randomly to one of three evaluation conditions, based on Zhang and Gosline (2023). In the baseline condition, no information was given on the source of the ideas. Participants were shown ideas (for smartphone apps that will help their users keep a healthier lifestyle) one at a time, and asked to rate the creativity of each idea on a 5-point scale ("very creative" to "very uncreative"). In the partially informed condition, participants were informed that "the ideas in this study came from two sources: -humans: Prolific participants who were asked to generate ideas; -AI: a large language model (Chat GPT or other) who was asked to generate ideas. The ideas you are about to evaluate came from one of these two sources. That is, they either all came from humans, or they all came from AI," followed by a comprehension check. Then they were shown 20 ideas one at a time ("Below is an idea that was generated either by a human or by AI.") which they were asked to rate for creativity on a 5-point scale. In the full information condition, participants were informed that "The ideas in this study came from two sources: -humans: Prolific participants who were asked to generate ideas; -AI: a large language model (Chat GPT or other) who was asked to generate ideas. The ideas you are about to evaluate all came from humans/AI," also followed by a comprehension check. They were shown 20 ideas ("Below is an idea that was generated by a human/AI") and asked to rate them for creativity on the same scale.

We obtained complete data from 1,174 respondents and their twins. Given the number of ideas (400), the number of evaluation conditions (3), and the number of ideas evaluated per participant (20), each idea was evaluated by 19.567 judges on average.

Results - Pre-registered Analyses

How do digital twins' ratings of idea creativity compare to those of their human counterparts?

We first analyze human ideas and compare the average creativity rating of each idea in the baseline condition coming from twins vs. human judges. We find that human ideas are rated on average significantly higher by twins than they are by humans ($M_{twins}=3.310$, $M_{humans}=3.198$, $t=2.859$, $p<0.01$), and that the standard deviation of the creativity ratings across ideas is smaller for ratings coming from twins vs. humans ($S_{twins}=0.223$, $S_{humans}=0.530$, $F(199,199)=5.664$, $p<0.001$). The mean absolute percentage error between ratings from twins vs. humans is 0.155. The correlation between ratings coming from twins vs. humans is positive but not statistically significant ($r=0.112$, $p=0.114$).

Next, we analyze twin ideas similarly. Here, we actually find the opposite pattern for average creativity ratings: twin ideas are rated on average significantly lower by twins than they are by humans ($M_{twins}=3.213$, $M_{humans}=3.322$, $t=3.742$, $p<0.01$). The standard deviation of the creativity ratings across ideas is again smaller for ratings coming from twins vs. humans ($S_{twins}=0.172$, $S_{humans}=0.396$, $F(199,199)=5.292$, $p<0.001$). The mean absolute percentage error between ratings from twins vs. humans is 0.103. The correlation between ratings coming from twins vs. humans is positive but not statistically significant ($r=0.103$, $p=0.148$).

Do digital twins show human favoritism, AI favoritism, or neither when evaluating ideas?

Figure 1 shows the average creativity rating as a function of the identity of the ideator, the raters, and the evaluation condition.

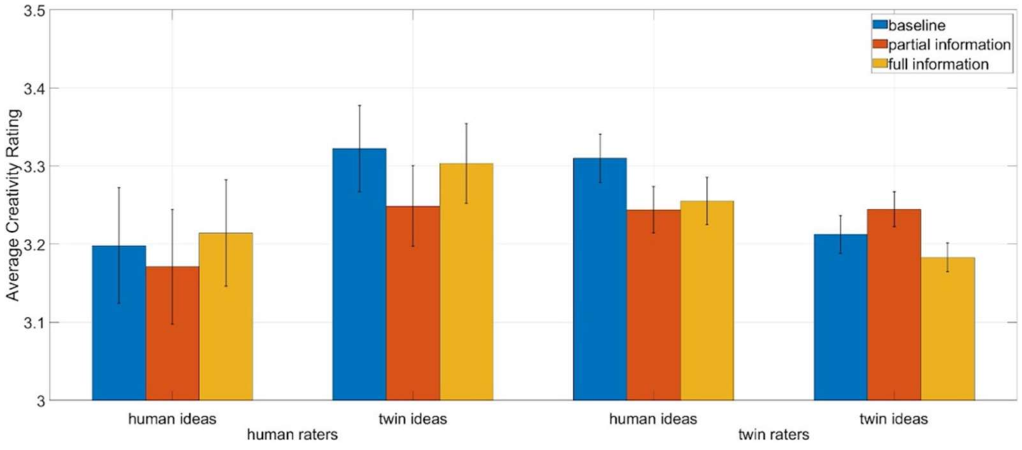


Figure 1: Average Creativity ratings Based on Identity of Ideator, Identity of Rater, and Condition.

We regress average creativity rating for idea coming from persona i based on identity j (binary human vs. twin) as rated by sample k (binary human vs. twin) using evaluation condition c on a fixed effect for human vs. twin ideator, a fixed effect for human vs. twin raters, fixed effects for evaluation conditions, all 2-way and 3-way interactions between the latter 3 sets of fixed effects, and a random intercept for persona i. The baseline corresponds to human ideas rated by humans in the baseline condition. Results are reported in Table 1.

The results reveal a positive main effect of twin ideator, a positive main effect for twin raters, and a negative interaction between the two. The main effect of twin ideator reflects a higher average rating given by human raters to ideas from twins in the baseline condition. The main effect for twin raters reflects a higher average rating given to human ideas by twin raters compared to human raters in the baseline condition. The negative Twin ideator * Twin raters interaction reflects the fact that despite the two positive main effects, the ratings given by twins to twin ideas are in fact lower than the ratings given by twins to human ideas or by humans to twin ideas. In sum, the regression results confirm a cross-over interaction in the baseline condition, whereby humans rate twin ideas more favorably than human ideas, but twins rate human ideas more favorably than twin ideas. We also see a significantly positive Twin ideator * Twin rater * Partial information 3-way interaction, reflecting that the Twin ideator * Twin rater interaction is weaker in the Partial information condition.

To test for AI or human aversion/favoritism more directly, we compare the baseline condition to the partial information and the full information conditions within each ideator-rater group. When humans are evaluating human ideas, neither the partial information nor the full information condition is statistically significantly different from the baseline condition. When humans are evaluating twin ideas, there is a statistically significant difference between the partial and the baseline condition (delta=-0.074, F(1,2388)=5.22, p<0.05), but there is no statistically significant difference between the full information and the baseline conditions. This suggests that while humans tend to rate ideas from twins higher than ideas from humans in the baseline condition, they tend to display AI aversion when rating twin ideas with only partial knowledge of the source of the ideas.

When twins are evaluating human ideas, we see a significant negative impact of the partial information condition compared to the baseline condition (delta=-0.066, F(1,2388)=4.17, p<0.05) and a marginally significant negative impact of the full information condition (delta=-0.054, F(1,2388)=2.86, p<0.10). When twins are evaluating twin ideas, we find no significant difference between the baseline condition and the partial or full information condition. This suggests that while twins tend to rate ideas from humans higher than ideas from twins in the baseline condition, they tend to display human aversion when rating human ideas with knowledge of the source of the ideas.

Intercept	3.198**
Twin ideator	0.124**
Twin raters	0.112**
Partial information	-0.027
Full information	0.016
Twin ideator * Twin raters	-0.221**
Twin ideator * Partial information	-0.046
Twin ideator * Full information	-0.035
Twin raters * Partial information	-0.038
Twin raters * Full information	-0.071
Twin ideator * Twin raters * Partial information	0.144*
Twin ideator * Twin raters * Full information	0.060
Twin_ID random intercept	yes
Number of observations	2,400
R^2	0.115

Table 1. Mixed effect regression results - idea evaluation. *: p<0.05; **: p<0.01.

Discussion

Creativity ratings coming from digital twins were not significantly correlated with the ratings coming from their human counterparts, suggesting that digital twins are currently not a reliable source of creativity ratings.

In the baseline condition in which no information was provided about the source of the ideas, humans rated twin ideas higher than human ideas, and twins rated human ideas higher than twin ideas, i.e., each group rated the other group higher. However, when information was provided about the source of ideas, humans displayed AI aversion (at least in the partial information condition), and twins displayed human aversion.

References

Zhang, Yunhao, and Renée Gosline. "Human favoritism, not AI aversion: People’s perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation." Judgment and Decision Making 18 (2023): e41.

Quantitative Intuition

Main Questions/Hypotheses

This study evaluates whether digital twins trained from large language models (LLMs) can accurately replicate individual responses on a newly developed psychometric instrument measuring Quantitative Intuition (QI) (Frank, Magnone, & Netzer, 2022). The QI scale assesses the respondent's own analytical and intuitive mindsets through 38 self-report items in addition 16 items measuring individuals' perceptions of their organization's analytical and intuitive orientations.

We test two core questions: First, can digital twins accurately reproduce individual human responses to the QI scale? Second, do digital twins match human participants on average scores for the quantitative (Q) and intuitive (I) mindset dimensions? Because the scale was developed in tandem with this study and was not available during LLM training, this provides a strong test of digital twins' ability to predict out-of-sample psychometric responses.

Methods

We recruited human participants from Prolific who were part of a larger digital twin study and collected matched responses from their corresponding digital twins. Invitations were sent to 2,000 participants from the original twin cohort; 1,435 individuals responded to the survey. Following pre-registered exclusion criteria, we removed 108 respondents who failed at least one of two attention checks, 8 respondents who "straightlined" (i.e., provided the same response across items 32–38 of the scale), and 1 respondent who completed the survey in under 150 seconds (mean response time was 789 seconds with a standard deviation of 457 seconds). The final sample included 1,323 respondents.

Participants completed 38 self-assessment items measuring their individual Quantitative Intuition (QI) mindset, followed by 16 items assessing their perception of their current or former organization's QI mindset. In addition, participants responded to two behavioral QI questions—one closed-ended and one open-ended—designed to capture willingness to "guestimate" and synthesize information. These behavioral items serve as external validity checks for the QI scale.

To explore individual differences in the QI scores, we examined correlations between QI scores and demographic and psychological variables collected as part of the broader digital twin survey.

Digital twins were prompted to complete the same set of items as their human counterparts, simulating how the participant might respond.

We derived individual QI scores using two approaches: 1) Simple averaging: we computed the mean of the first 19 items as a quantitative mindset score (Q), and the mean of the next 19 items as an intuitive mindset score (I), and 2) Factor analysis: we conducted an exploratory factor analysis with two factors, which naturally aligned with the quantitative and intuitive dimensions. Factor scores from this model were used as alternative QI measures.

We then compare the individual, simple average and factor analysis scores of the human respondents with those of the digital twins.

Results - Pre-registered Analyses

As pre-registered we compared the actual QI score of the Prolific respondents with the QI score of their digital twins. We do so both at the individual question level and for the average of the individual and organizational quantitative and intuitive scores.

Comparing the average score of the human responses and the digital twins

Comparing the correlation between the average (or factor) scores between the actual respondents and the digital twins for the individual and organizational quantitative and intuitive scores we find that overall the digital twins answers are closer to the actual respondents answers for the individual relative to the organizational skill both for the average quantitative and intuitive scores and for the factor analysis scores.

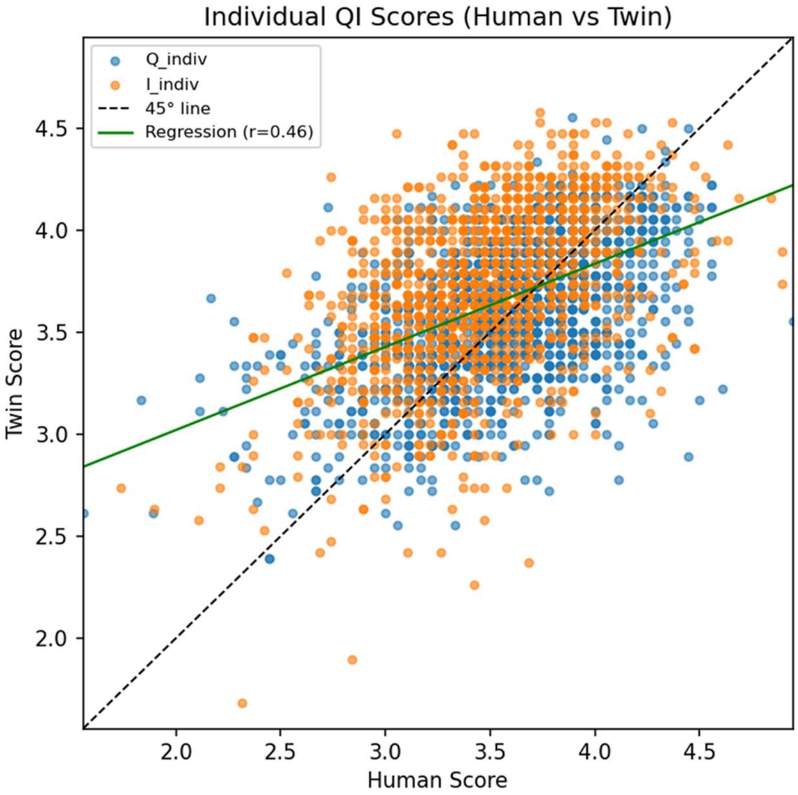
The Cronbach Alpha for the Q and I individual scales for the human respondents are 0.788 and 0.753, respectively, suggesting acceptable levels of agreement for items in the scale. Interestingly the digital twin mimicked similar or even higher level of agreement for items in the scale with Cronbach alpha of 0.776 and 0.866 for the Q and I individual scales.

Comparing the means and standard deviations of the scale scores between the humans and and digital twins shows that the means are quite similar, but the digital twins had consistently lower variance in the responses across respondents. A possible reason for that is that the digital twins are still regressing to some overall LLM mean.

Scale		Human	Digital Twin
Q Ind.	Mean	3.579	3.573
	STD	0.451	0.355
I Ind.	Mean	3.478	3.713
	STD	0.412	0.394
Q Org.	Mean	3.580	3.014
	STD	0.643	0.418
I Org.	Mean	3.219	3.479
	STD	0.486	0.280

Correlations between the human responses and the digital twins

Looking at the correlations between the individual level individual QI scale responses and those of the digital twins, we find a moderately strong correlation between the responses to the 19 quantitative questions (r=0.513, p.value<0.001) and the 19 individual scores (r=0.481, p.value<0.001), with an overall correlation of 0.46. Thus, at the individual level the digital twins are capable of mimicking the human response to QI level to a moderate extent. The correlations between the digital twins and the human responses for the organizational QI scores are much lower, though statistically significant (r_quantitative _org=0.252, p<0.001 and , r_intuitive _org=0.196, p<0.001). This is to be expected as the original questions used to build the digital twins focused on the individuals themselves rather than on the organization they work/worked for.



Finally, looking at the individual questions we find a positive and significant (after Bonferroni correction for multiple tests) correlation at the 95% level between the human digital twins scores for 35 out of the 38 individual QI questions. For the organizational QI questions, we find a positive and significant (after Bonferroni correction for multiple tests) correlation at the 95% level between the human and digital twin responses for 11 out of 16 questions. We did not find any significant and negative correlations between the human and digital twin responses for any of the questions.

We did not find a statistically significant relationship between the QI scores for either the willingness to guestimate questions or the ability to synthesize scores. However, we found an interesting result with respect to the ability of the digital twins to answer the QI behavioral questions. In a close ended question, we gave respondents some pieces of information and challenged them to address the problem by guestimation. The digital twins chose to guestimate, which is the normative answer for the guestimation question, 99.9% of the times (only one twin did not choose the normative response), whereas for the humans only 59.4% were inclined to guestimate. The difference between the human and digital twin responses is statistically significant (p <0.001).

In an open ended question, we asked respondents to explain the situation for a struggling company based on data, where the normative QI response would be to synthesize the information rather than mainly summarize it. We asked GPT 4o mini to classify each human or digital twin response as a summary or a synthesis. A research assistant independently classified 20 responses and found that their classification perfectly matched with those of ChatGPT. We found a similar result to those of the willingness to guestimate for the ability to synthesize. Whereas 99.8% of the digital twins synthesized the information, only 71.5% of human respondents replied with the QI normative response to synthesize. The difference between the human and digital twin responses is statistically significant (p <0.001).

Thus, while the digital twins are capable of moderately following human responses to the QI scale, when it comes to mimicking human behavior questions, they tend to overexhibit behavior that is consistent with normative QI mindset.

Results - Additional Analyses (Non-Preregistered)

We also analyzed the relationship between the Q and I scores and the demographic and individual characteristics collected in the first phase of the digital twins study. Specifically, we calculated the correlations between the average individual Q and I scores and each of the scales measured in that initial phase. Focusing on individual rather than organizational QI scores, we report only the correlations that are statistically significant at the 95% confidence level after applying the Bonferroni correction for multiple testing.

Several interesting results emerge.

1. The individual mindset score (I) is statistically significantly correlated with the following scales at the 95% level:
    1. Regulatory Focus (Fellner et al. 2017) - r=0.452
    2. Openness (John & Srivastava, 1999) - r=0.428
    3. Consumer Need for Uniqueness (Ruvio et al., 2008) - r=0.379
    4. Self monitoring (Lennox & Wolfe, 1984) - 0.352
    5. Agentic (Trapnell & Paulhus, 2012), r=0.306
    6. Extraversion (John & Srivastava, 1999), r=0.285
    7. Need for Cognition (Cacioppo, Petty, & Kao, 1984), r=0.258
    8. Green (Haws, Winterich, & Naylor, 2014) - r=0.213
    9. Maximization (Nenkov et al., 2002), r=0.212
    10. Overconfidence (Dean & Ortoleva, 2019) - r=0.198
    11. The Horizontal and Vertical Individualism and Collectivism scales (Triandis & Gelfand, 1998) - r=0.169, r=0.182. r=0.217. r=0.125.
    12. Communion (Trapnell & Paulhus, 2012), r=0.158
    13. Basic Empathy (Carre et al., 2013) - r=0.149
    14. Minimalism (Wilson & Bellezza, 2021), r=0.106

15. Total performance on the cognitive test - r= -0.113.
16. Cognitive Reflection Test 2 (Thomson and Oppenheimer 2016) - r= -0.116

Thus, individuals who score high on having an intuitive mindset also tend to score high on regulatory focus and openness—traits that are correlated with greater emotional awareness, creativity, and unconventionality. They also tend to seek uniqueness, display overconfidence, be extraverted, agentic, and self-aware. While they perform well on communication measures, they tend to score lower on cognitive tests. Of particular interest is the negative correlation with the Cognitive Reflection Test 2 (CRT-2), which assesses individuals' ability to override intuitive (System 1) responses in favor of more deliberate (System 2) reasoning. This negative correlation aligns with our expectations: a high intuitive mindset score on our scale should correspond to greater reliance on System 1 processing. This finding provides a compelling validation of our scale.

2. The analytical/quantitative mindset score (Q) is statistically significantly correlated with the following scales at the 95% level:
    1. Need for Cognition (Cacioppo, Petty, & Kao, 1984) - r=0.380
    2. Regulatory Focus (Fellner et al. 2017) - r=0.342
    3. Conscientiousness (John & Srivastava, 1999) - r=0.320
    4. Openness (John & Srivastava, 1999) - r=0.259
    5. Self Monitoring (Lennox & Wolfe, 1984) - 0.248
    6. Minimalism (Wilson & Bellezza, 2021), r=0.237
    7. Green (Haws, Winterich, & Naylor, 2014) - r=0.215
    8. The Horizontal and Vertical Individualism and Collectivism scales (Triandis & Gelfand, 1998) - r=0.223, r=0.195. r=0.210. r=0.153.
    9. Agentic score (Trapnell & Paulhus, 2012) - r=0.206
    10. Numeracy - r=0.170
    11. Maximization (Nenkov et al., 2002), r=0.162
    12. Consumer Need for Uniqueness (Ruvio et al., 2008) - r=0.142
    13. Overconfidence (Dean & Ortoleva, 2019) - r=0.140
    14. Extraversion (John & Srivastava, 1999), r=0.136
    15. Financial literacy (Johnson, Meier, Toubia 2018) - r=0.134
    16. Agreeableness (John & Srivastava, 1999) - r=0.132
    17. Communion score (Trapnell & Paulhus, 2012), r= 0.103
    18. Self Concept Clarity (Campbell et al., 1996) - r=0.099
    19. Social desirability (Crowne & Marlowe, 1960) - r= 0.097
    20. Depression - r= -0.119
    21. Neuroticism (John & Srivastava, 1999) - r= -0.183

Several measures that are statistically significantly and positively correlated with the intuitive mindset scale (I) are also correlated with the analytical mindset scale (Q). These include need for cognition, regulatory focus, openness, self-monitoring, individualism, collectivism, minimalism, environmental concern ("green"), agentic traits, communication skills, and overconfidence.

On the other hand, several individual characteristics appear to be uniquely associated with scoring high on the analytical (Q) scale. Individuals who scored high on the analytics mindset questions tend to also score high on conscientiousness, a trait often linked to organization and discipline, as well as on numeracy and financial literacy, which reflect core quantitative skills. They also score higher on agreeableness, social desirability, and self-concept clarity, and lower on neuroticism and depression. Additionally, an analytical mindset is positively and significantly associated with being male and having a higher income.

Discussion

Overall, we first validated the QI scale by running it on a large sample of respondents, identifying a meaningful two-factor structure that captures both an analytical mindset and an intuitive mindset. We also found meaningful and expected relationships between the QI scale and a set of individual characteristics—such as higher numeracy, financial literacy, and conscientiousness among individuals with an analytical mindset, and a stronger preference for System 1 processing among those with an intuitive mindset.

In the digital twins analysis, we found a moderate correlation (around 0.5) between human responses to the individual QI scale and those of their digital twins, and a weaker relationship (correlation of 0.2-0.25) between human responses and their digital twins' assessments of the QI score of the organization they work/ed for, possibly due to the absence of questions about the individual's workplace in the digital twin's construction. We also observed that digital twin responses underestimated the variance present in the actual data, possibly suggesting that the digital twin reflects a weighted combination of the fine-tuned individual data and the pre-trained global LLM output.

Finally, while the digital twins did a modest job of capturing the QI scale, they failed to mimic responses to the two QI behavior questions. The digital twins tended to consistently select the normative, high-QI behavior much more frequently than humans did.

References

Frank, Christopher J., Paul F. Magnone, and Oded Netzer. Decisions over decimals: Striking the Balance between Intuition and Information. John Wiley & Sons, 2022.

Promiscuous Donors

Main Questions/Hypotheses

Political donations play a critical role in shaping electoral outcomes. They equip candidates with the financial resources needed to compete in political races (e.g., Gilens 2012; Hussein et al. 2025). They also function as signals of support. Political parties often use donation counts as thresholds for unlocking institutional backing and as a pre-requisite for qualifying for televised debates. In the present research, we investigate the reputational consequences of a specific form of donation behavior: contributing to both major political parties. We refer to individuals who donate to two parties as "promiscuous donors" and we examine how they are perceived by others. Understanding how promiscuous donors are perceived is important given the prevalence of this behavior. Analysis of 2024 Federal Election Commission data reveals that at least 113,187 Americans made contributions to both parties during the 2024 election cycle.

Ex ante, it is unclear how individuals will evaluate those who donate to both political parties. On one hand, promiscuous donors may be viewed favorably. Some scholars have argued that political polarization has reached unsustainable levels (Heltzel and Laurin 2020), prompting a public desire for greater bipartisanship and reconciliation. From this perspective, contributing to both parties could signal a conciliatory stance, potentially enhancing reputational standing.

On the other hand, several theoretical perspectives suggest that such behavior may incur reputational costs. Partisan animosity has intensified in recent years, with Democrats and Republicans expressing growing levels of dislike and distrust toward each other (Pew Research Center 2022). Furthermore, emerging research on receptiveness to opposing views (Hussein and Tormala 2021) suggests that open-mindedness toward the other side can backfire socially (Hussein and Wheeler 2024). If donors are seen as legitimizing or aligning with an ideologically objectionable outgroup, their behavior may be perceived as disloyal or suspect, thereby leading to reputational costs.

Methods

We recruited 799 human respondents on Prolific and their respective twins. This study used a rating-based conjoint design (Green & Srinivasan, 1978). Conjoint is a study design used to determine how people value different attributes (e.g., screen size, memory, color) that make up a product (e.g., a laptop). Participants in a conjoint study are shown a series of trials, each showcasing a different version of the target product. By varying different attribute levels of the product (e.g., a black vs. silver laptop) between trials and measuring how consumers respond to different combinations of these attributes, managers can quantify how much a given attribute (e.g., color) affects overall attitudes.

Applying this approach, we had participants read 10 vignettes, each describing a target who engaged in a different donation behavior. The donation behaviors included contributing solely to the Republican Party, solely to the Democratic Party, to neither party, or to both parties. Our primary focus was on the reputational consequences of donating to both parties. To this end, we systematically varied the rationale provided for such behavior. Specifically, we compared conditions in which no justification was given to those in which a self-interested or a values-based rationale was offered. Self-interested justifications emphasized pragmatic benefits, such as attracting bipartisan talent to one's business or facilitating smoother relations with elected officials across party lines. In contrast, values-based justifications highlighted ideological motivations, such as supporting candidates with the best ideas regardless of party or promoting cross-party cooperation. Below is a summary of the exact description used and the attribute level they correspond to:

- Own Party or Opposing Party (coded relative to participant's party affiliation; independent participants who rated these had their own category):
    - "I currently only donate to Republicans"
    - "I currently only donate to Democrats"
- No donations: "I currently don't donate to any politician"
- Both: "I currently donate to both Democrats and Republicans"
- Business Strategy: "I currently donate to both Democrats and Republicans so my business can work well with whoever wins"
- Attract Talent: "I currently donate to both Democrats and Republicans because it helps me attract top talent for my business"
- Common Grounds: "I currently donate to both Democrats and Republicans because I believe in working across the aisle to find common grounds"
- Good Ideas: "I currently donate to both Democrats and Republicans because I support candidates who have good ideas, regardless of which party they come from"

To enhance the generalizability of our findings, we systematically varied multiple dimensions of the vignettes beyond donation behavior. This design enabled us to estimate the average effect of donating to both parties across a range of contextual backgrounds. Specifically, we manipulated the target's political affiliation (Democrat vs. Republican), job (e.g., CEO of a publicly traded company vs. founder of a fast-growing start-up), and donation history (e.g., history of prior donations vs. first-time donors). This approach ensured that observed effects were not idiosyncratic to a single type of target. Below is an example vignette used in the study:

> Imagine you meet someone named John at a neighborhood barbecue.
>
> John recently moved to your area. He is a Democrat. He is the CEO of a publicly traded company.
>
> During your conversation, the topic of donating to political campaigns comes up.
>
> John mentioned that he never used to donate to politicians.
>
> John then said to you:
>
> "I currently donate to both Democrats and Republicans."

After reading the vignette, participants were asked to report their overall impressions of the target: “Based on what John said, what's your overall impression of him?” Participants reported their attitudes on a nine-point bipolar scale, ranging from very unfavorable to very favorable. As a reminder, each participant rated ten targets in total.

Results - Pre-registered Analyses

As preregistered, we regressed overall impression of the target on dummy variables of the manipulated attributes and participant type (digital twin vs. human). Importantly, we included an interaction between our focal manipulated attribute–donation behavior--and participant type. Standard errors were clustered on the participant level to account for repeated measurement. More specifically, we estimated the following regression:

$$\begin{aligned}\widehat{Impression_{i,j}} = \alpha + \overrightarrow{\beta_1}\, donation\ behavior_{i,j} + \beta_2\, participant\ type_i \\ + \overrightarrow{\beta_3}\, donation\ behavior_{i,j} \times participant\ type_i + \beta_4\, target\ party_{i,j} \\ + \overrightarrow{\beta_5}\, history_{i,j} + \overrightarrow{\beta_6}\, job_{i,j}\end{aligned}$$

We were interested in the vector of interactions $\vec{\beta}_3$. Significant interactions would suggest that humans and their digital twins differed in their reactions to current donation behavior. Non-significant interactions would suggest no difference.

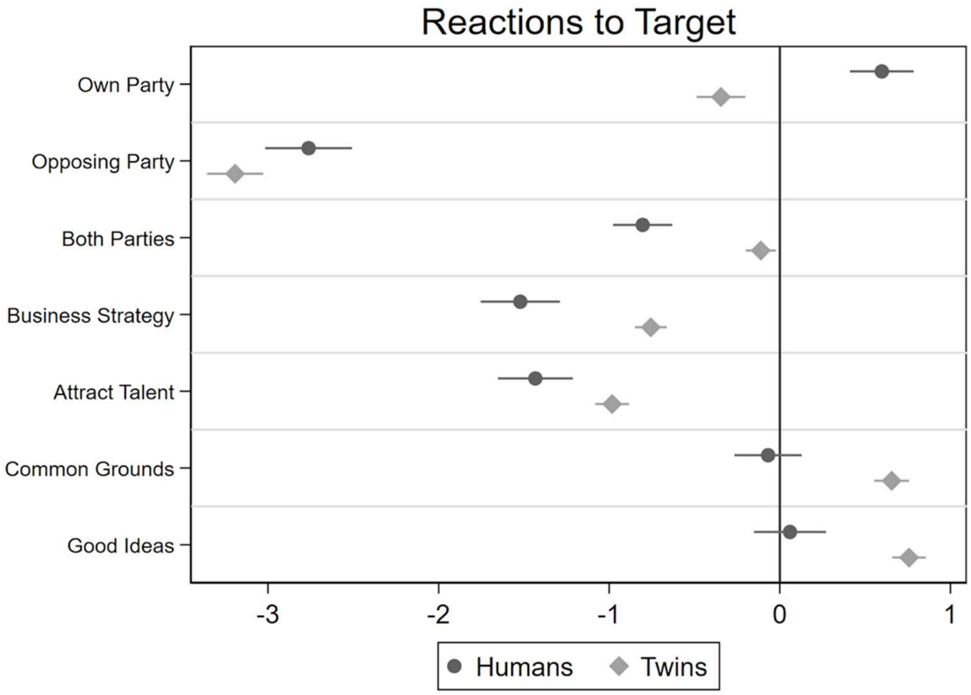


Figure 1. Reactions to target among humans and digital twins. Error bars denote 95% confidence intervals. Standard errors are clustered at the participant level. The reference category is a target who did not donate to either party. For legibility, the dummy variable

indicating that an independent participant rated a single-party donor is omitted from the figure but is included in the regression.

Of the eight interaction terms, seven reached significance (ps < .001). Both humans and digital twins evaluated targets who donated to the opposing party more negatively than targets who made no political donations (interaction b = –.32, SE = .27; *t* = –1.18, *p* = .24).

Four of the seven significant interactions revealed divergent effects in magnitude between human participants and their digital twins. Both humans and twins rated targets who donated to both parties negatively relative to targets who did not donate to either party, but digital twins were more positive in their evaluations (b = –.11, SE = .05, t = –2.47, p = .014) than humans (b = –.80, SE = .09, t = –9.09, p < .001). Similarly, when targets provided self-interested reasons for donating to both parties, both humans and twins rated them more negatively compared to targets who did not donate at all. However, twins were more positive. For example, both humans and twins rated targets who donated to both parties because they wanted to be able to work with whichever party is in power negatively, but twins were more positive (b = –.76, SE = .05, t = –15.86, p < .001) than humans (b = –1.52, SE = .12, t = –12.85, p < .001). The same pattern held for targets who donated to both parties because they wanted to attract top talent to their business (humans: b = –1.43, SE = .11, t = –12.84, p < .001 vs. twins: b = –.98, SE = .05, t = –19.56, p < .001).

The three remaining significant interactions revealed divergent effects in direction between human participants and their digital twins. Human respondents rated targets who donated to their own party more positively than those who did not donate (b = .60, SE = .09, t = 6.30, p < .001). In contrast, digital twins evaluated the same targets more negatively compared to those who did not donate (b = –.34, SE = .07, t = –4.74, p < .001). A similar divergence emerged in evaluations of targets who donated to both parties for value-based reasons. Humans rated them the same way they rated targets who did not donate to either party. However, twins rated them more positively. Specifically, among humans, there was no significant difference between targets that donated to both parties because they believe in working across the aisle to find common grounds and targets who did not donate (b = –.07, SE = .10, t = –0.68, p = .50). However, twins rated the former more positively than targets who did not donate (b = .66, SE = .05, t = 12.61, p < .001). The same pattern held for targets who donated to both parties because they wanted to support the best ideas regardless of which party they came from (humans: b = .06, SE = .11, t = 0.56, p = .58 vs. twins: b = .76, SE = .05, t = 15.12, p < .001).

Discussion

How do people evaluate promiscuous donors (i.e., targets who donate to both parties)? Our results indicate that, compared to those who do not donate to politics at all, those who donate to both parties are perceived more negatively. This result is consistent with recent research on the reputational costs of receptiveness to opposing views (Hussein & Wheeler, 2024). Interestingly, the reputational costs of promiscuous donation are particularly salient when the reason behind donating to both parties is self-interest (e.g., to attract the best talent to one's business or to make sure that one's business can work with whichever party wins) and are somewhat weaker when the reason is more value-based (e.g., believing in bridging divides or that good ideas should be supported regardless of which party they come from).

How do humans and their digital twins compare when judging the donation behavior of others? Across more than 7,000 conjoint judgments, human respondents penalized promiscuous donors, especially when the behavior appeared motivated by self-interest (e.g., cultivating access or recruiting talent). In contrast, digital twins were less punitive (sometimes even favorable) toward the very same targets.

What might drive these differences? One explanation is that language models are trained to discount signals of intergroup conflict. Hence, a cross-party donation that humans interpret as a collaboration with the enemy is read by the model as pragmatic or even positive as it shows a willingness to reach across divides. A second, complementary account centers on affect: if twins lack the visceral emotions that energize partisan animus, they down-regulate the reputational costs of interacting with the political out-group. Further research on how digital twins compare to humans in matters of social cognition would be worthwhile.

References

Gilens, Martin. (2012). *Affluence and Influence: Economic Inequality and Political Power in America*.

Green, Paul E., and V. Srinivasan. (1978). “Conjoint Analysis in Consumer Research: Issues and Outlook.” *Journal of Consumer Research*, 5(2), 103–123.

Heltzel, G., & Laurin, K. (2020). Polarization in America: Two possible futures. *Current opinion in behavioral sciences*, *34*, 179-184.

Hussein, Mohamed A., Courtney Lee, and S. Christian Wheeler (2025), “How do Consumers React to Ads that Meddle in Out-Party Primaries?” *Journal of Consumer Research*, 51(6), 1186-1208.

Hussein, Mohamed A. and Zakary L. Tormala (2021), “Undermining Your Case to Enhance Your Impact: A Framework for Understanding the Effects of Acts of Receptiveness in Persuasion,” *Personality and Social Psychology Review*, 25(3), 229-250.

Hussein, Mohamed A. and S. Christian Wheeler (2024), “Reputational Costs of Receptiveness: When and Why Being Receptive to Opposing Political Views Backfires,” *Journal of Experimental Psychology: General*, 153(6), 1425-1448.

Pew Research Center (2022), “As Partisan Hostility Grows, Signs of Frustration With the Two-Party System,” *Pew Research Center,* accessed from: “https://www.pewresearch.org/politics/2022/08/09/as-partisan-hostility-grows-signs-of-frustration-with-the-two-party-system/

Appendix - Regression Tables

	(1)
	rating_
Twins	0.411***
	(0.0694)
Own Party=1	0.502***
	(0.0915)
Opposing Party=1	-2.632***
	(0.127)
Both Parties=1	-0.802***
	(0.0883)
Business Strategy=1	-1.509***
	(0.116)
Attract Talent=1	-1.425***
	(0.110)
Common Grounds=1	-0.0553
	(0.0984)
Good Ideas=1	0.0768
	(0.106)
Partisan Donation (Independent)=1	-1.682***
	(0.243)
Twins # Own Party=1	-0.745***
	(0.117)
Twins # Opposing Party=1	-0.700***
	(0.153)
Twins # Both Parties=1	0.691***
	(0.0994)
Twins # Business Strategy=1	0.739***
	(0.124)
Twins # Attract Talent=1	0.435***
	(0.119)

Twins # Common Grounds=1	0.700***
	(0.110)
Twins # Good Ideas=1	0.665***
	(0.115)
Twins # Partisan Donation (Independent)=1	-0.315
	(0.267)
Target: Own Party	0.311***
	(0.0865)
Target: Opposing Party	-0.0112
	(0.0879)
Start-up Founder	-0.112***
	(0.0369)
CEO	-0.197***
	(0.0378)
No Donations	0.0808**
	(0.0318)
Both	0.119***
	(0.0313)
Observations	15980
R^2	0.311
Adjusted R^2	0.310

	(1) Ratings Humans	(2) Ratings Twins
Own Party	0.598***	-0.345***
	(0.0949)	(0.0728)
Opposing Party	-2.762***	-3.193***
	(0.130)	(0.0839)
Both Parties	-0.804***	-0.111**
	(0.0884)	(0.0450)

Business Strategy	-1.521***	-0.756***
	(0.118)	(0.0477)
Attract Talent	-1.433***	-0.983***
	(0.112)	(0.0502)
Common Grounds	-0.0684	0.656***
	(0.100)	(0.0520)
Good Ideas	0.0598	0.758***
	(0.108)	(0.0501)
Partisan Donation (Independent)	-1.580***	-2.073***
	(0.255)	(0.0925)
Target: Own Party	0.315*	0.338***
	(0.182)	(0.0582)
Target: Opposing Party	0.218	-0.224***
	(0.183)	(0.0606)
Start-up Founder	-0.0795	-0.146***
	(0.0633)	(0.0375)
CEO	-0.161**	-0.231***
	(0.0662)	(0.0360)
No Donations	0.117**	0.0454
	(0.0571)	(0.0278)
Both	0.127**	0.111***
	(0.0565)	(0.0262)
Observations	7990	7990
R^2	0.176	0.546
Adjusted R^2	0.175	0.545

Privacy Preferences

Main Questions/Hypotheses

Consumers have privacy concerns regarding tracking and targeting in online advertising. In response to such privacy concerns about personal data collection and use, the online advertising industry is developing technically sophisticated privacy-enhancing technologies (PETs). On the premise that consumers perceive their privacy to be violated differently under different advertising practices that employ data tracking and targeting differently, we want to study how well their digital twins capture these perceptions.

Motivated by Google's Privacy Sandbox initiative, we study six main practices of firms in online advertising, labeled as different scenarios from A to F. For the status quo of behavioral targeting (Scenario F), the degree of tracking is high because the consumer is tracked across websites and the data leaves the local machine, and the degree of targeting is also high because the consumer is targeted at the individual level. For the Individual-level Targeting PET (Scenario E), the degree of tracking is low because although the consumer's activity is tracked, the consumer's data does not leave the machine; however, the degree of targeting is high because the consumer receives individually-targeted ads. For the Group-level Targeting PET (Scenario D), the degree of tracking is low because although the consumer's activity is tracked, the consumer's data does not leave the machine. In this case, the degree of targeting is at a medium level because the consumer is profiled and receives ads targeted at a group level. For contextual targeting (Scenario C), the degrees of tracking and targeting are both low, as the focal website only determines the individual-level presence of the user on the website. The consumer is not tracked at the individual-level on the focal website they are visiting and on other websites (i.e., no past behavioral browsing data is used for profiling and targeting the user, and the only data used for targeting is the fact that the consumer is present on the website); however, contextual targeting may still trigger privacy concerns (Bleier 2021). When untargeted ads are shown to consumers and they are not tracked (Scenario B), or there is no tracking and no ads are shown to consumers (Scenario A), the degrees of tracking and targeting are both zero. We expect that consumers' average perceived privacy violations (PPVs) will be in the order Scenarios F, E, D, C, B and A (with F having the highest PPV and A the lowest).

The main question that we want to investigate is how PPV values compare for humans and their digital twins in magnitude (in aggregate and individually) as well as in an ordinal sense (in aggregate).

Methods

We measure consumers' PPVs for the advertising practices in Scenarios A through F through an online study with approximately 1,200 subjects (i.e., approximately 200 per condition) in the United States. A human, allocated randomly to a scenario, is presented with a description of the scenario and asked how much they perceive their privacy to be violated by this advertising practice. Later the digital twin of this human is asked the exact same question.

Results - Pre-registered Analyses

For each of the 6 conditions, Table 1 reports: the human-twin sample size; mean PPV by human respondents (along with the lower and upper bounds of the 95% confidence interval); mean PPV by digital twins of human respondents (along with the lower and upper bounds of the 95% confidence interval); Pearson r between human and twin PPVs; paired t-test on PPVs; and F-test on variances. Table 2 provides the overall Spearman rho on the rank ordering of condition means. Twins answered “1 - not at all” for all observations in Scenarios A, B and C, so correlations and certain tests are undefined in these scenarios.

Measure	A: No Ads, No Tracking	B: Untargeted Ads	C: Contextual Targeting	D: Group-level PET	E: Individual-level PET	F: Behavioral Targeting
N_pairs	199	218	187	222	180	194
Mean_human	1.744	2.133	2.658	4.667	4.461	5.469
Mean_twin	1	1	1	2.712	2.817	4.603
CI95_hu_low	1.553	1.935	2.396	4.443	4.214	5.246
CI95_hu_high	1.935	2.331	2.919	4.89	4.709	5.692
CI95_tw_low	1	1	1	2.616	2.727	4.476
CI95_tw_high	1	1	1	2.807	2.907	4.731
Pearson_r				0.014	-0.069	0.088
p-val_r				0.84	0.355	0.222
t_stat	7.676	11.305	12.496	15.945	12.058	6.907
p-val_t	0	0	0	0	0	0
F_var				5.446	7.569	3.07
p-val_F				0	0	0

Table 1: Human and Twin Responses per Condition.

	rho	p-value
Spearman_rank_corr	0.88	0.021

Table 2: Spearman rank correlation of condition means from Human vs Twins.

The results of the experiments show the following:

- The PPVs for human subjects are in the order hypothesized (with one deviation being that the mean PPV of Scenario D is higher than the mean PPV of Scenario E; however, these numbers are not statistically different accounting for the 95% CIs).
- The PPVs for the digital twins are significantly lower in magnitude, i.e., digital twins have lower perceptions of privacy violations.
    - All twins give PPV values of 1 for Scenarios A, B and C, in which there is no tracking or targeting on past behavior.

    - The Pearson correlations between the PPVs of humans and their digital twins, when defined, are near zero.
    - The t-stats for the paired t-test between PPVs of humans and their digital twins show that these PPVs are significantly different.
    - The F-test of variances between PPVs of humans and their digital twins show that the variances in PPVs are significantly different.
- While the PPVs of digital twins are significantly lower than the PPVs of humans, the rank ordering of the scenarios, based on average PPV, is similar for humans and digital twins, with a Spearman rank correlation of 0.88.

Discussion

The results of the study indicate that digital twins perceive their privacy to be violated less than their human counterparts (as indicated by lower PPV scores for digital twins compared to humans). The twins also do not do a good job of capturing heterogeneity in perceived privacy violations (as indicated by almost zero correlations in PPV scores for digital twins compared to humans). However, in aggregate, humans and their digital twins rank privacy violations similarly across advertising practices that rely differently on tracking and targeting (as indicated by a high rank correlation).

Heterogenous Story Beliefs

Main Questions/Hypotheses

When reading a book and predicting the emotion (valence and arousal) of the next chapter based on the previous chapter, what is the correlation between the expected valence and expected arousal from humans vs. their digital twins (LLMs)?

We hypothesize that LLM-based digital twins can meaningfully simulate human expectations about emotional trajectories in narrative contexts, as measured by correlations between human and LLM predictions of valence and arousal for upcoming story content.

Methods

250 participants were recruited via Prolific with stratified sampling to ensure equal representation across racial groups (50 participants each from White, Black, Hispanic, Asian, and Other categories), and their digital twins.

Participants read chapters from 16 possible stories, with each participant randomly assigned to read 2 stories. Each story consisted of two chapters. Participants made predictions after reading each chapter: after Chapter 1 (predicting Chapter 2) and after Chapter 2 (predicting what would come next), resulting in 4 prediction points per participant.

Primary Dependent Variables:

1. *Human valence expectation*: Participants assigned probabilities to five valence levels for the next chapter:
    - Very negative (1) to Very positive (5)
    - Question: "Based on the chapter you just read, what do you think is the likelihood that the text of the next chapter will be very negative or very positive? Please assign a percentage to each. The total must sum to 100%."
2. *Human arousal expectation*: Participants assigned probabilities to five arousal levels for the next chapter:
    - Very low energy (1) to Very high energy (5)
    - Question: "Based on the chapter you just read, what do you think is the likelihood that the text of the next chapter will be very low energy or very high energy? Please assign a percentage to each. The total must sum to 100%."

From these probability distributions, we calculated expected values to create continuous measures of expected valence and arousal.

Results - Pre-registered Analyses

The analysis included all prediction points (predictions made after both Chapter 1 and Chapter 2 of each story):

1. Valence Expectations:
    a. Pearson correlation: r = 0.546 [95% CI: 0.501, 0.588]
    b. Statistical significance: t(1002) = 20.65, p < 0.001
2. Arousal Expectations:
    a. Pearson correlation: r = 0.384 [95% CI: 0.330, 0.436]
    b. Statistical significance: t(1002) = 13.17, p < 0.001

Both correlations were statistically significant at p < 0.001, indicating that LLM predictions are significantly associated with human expectations for both valence and arousal across all chapter transitions. The moderate positive correlations suggest meaningful alignment between human and LLM story belief patterns.

Discussion

This study provides evidence that LLM-based digital twins can meaningfully simulate human emotional expectations in narrative contexts. The significant positive correlations for both valence (r = 0.416) and arousal (r = 0.420) support our hypothesis that LLMs can capture how humans predict emotional trajectories in stories.

Targeting Fairness

Main Questions/Hypotheses

Companies commonly target their advertisements based on demographic characteristics, such as race or gender. Yet recent work has found that people often view the decision to target to be less fair than advertising broadly to the general population (Shaddy, Friedman and Toubia 2025). In this sub-study, we aim to replicate this pattern with human participants, and test whether their digital twins indicate similar fairness ratings.

Methods

We opened the study to 400 human participants on Prolific, and 357 passed the screener and completed it. The study was subsequently run on their 357 digital twins, for a total of 714 participants. Participants and their corresponding digital twins were randomly assigned to the same condition in a two-cell (broad vs. targeted) between-subjects design.

All participants first read: "A snack foods company has developed a new line of snacks. Initial testing showed that, due to the taste and texture profile of the snacks, the snacks are better suited to the preferences of their female customers." In the broad condition, participants next read: "Even though they believe the snacks are best suited to their female customers, they will advertise the snacks broadly to the general public." In the targeted condition, participants next read: "Because they believe the snacks are best suited to their female customers, they will advertise the snacks to women directly." Finally, participants rated fairness: "How fair is this advertising plan?" ("Not at all fair" = 1; "Very fair" = 9).

Results - Pre-registered Analyses

We conducted a two-way ANOVA with targeting condition (broad vs. targeted) and participant type (human vs. digital twin) as the independent variables, and fairness ratings as the dependent variable. We found a significant main effect of targeting condition (F(1, 713) = 69.07, $p < .001$), but no significant effects of participant type (F(1, 713) = 0.05, p = .823) or their interaction (F(1, 713) = 1.57, p = .211), indicating that, at an aggregate level, the digital twins rated fairness similarly to the humans. Pairwise comparisons confirm that for humans, the decision to target was rated as less fair (M = 6.38, SD = 2.17) than advertising broadly (M = 7.47, SD = 1.83, F(1, 710) = 45.72, $p < .001$), and for the digital twins, targeting was also rated as less fair (M = 7.30, SD = 0.78) than advertising broadly (M = 6.50, SD = 0.80, F(1, 710) = 24.92, $p < .001$).

Results - Additional Analyses (Non-Preregistered)

Although not part of our pre-registered plan, a notable observation stood out in the analysis: the variance among the digital twins was lower than for the humans. In particular, for the digital twins, 99% of responses used the upper end of the scale (5-9), compared with only 87% of responses for the human participants ($x^2(1) = 25.27$, $p < .001$). Examining the responses closest to the mean (i.e., respondents answering "7"), 63% of the digital twins provided that response, compared with 20% of human participants ($x^2(1) = 136.59$, $p < .001$). The overall correlation between human participants and their digital twin was weak but significant (r = .13, p = .012).

Discussion

At the aggregate level, the digital twins closely matched human judgments of the perceived fairness of demographic targeting, replicating the finding that it is viewed as less fair than broad targeting and producing similar average ratings across conditions. However, the twins' responses were more clustered around the mean, whereas human participants were more likely to use the whole scale.

References

Franklin Shaddy, Friedman, Elizabeth M. S., and Olivier Toubia (conditionally accepted), "Fairness Perceptions in Demographic Segmentation and Targeting," Journal of Consumer Research.

User Behavior with Recommendation Systems

Main Questions/Hypotheses

Modern machine learning and AI-based recommendation systems are data-driven, generating personalized recommendations from a user's previous interactions. Often, platforms make the assumption that a user's engagement with a particular piece of content is indicative of their utility for the content. However, a recent online behavioral experiment has shown that users may behave strategically and alter their engagement to influence the content they get recommended in the future (Cen, Ilyas, Allen, Li, Madry '24). For example, a user on a music platform may skip past a "guilty pleasure" song that they like, because they are worried that the recommendation algorithm will recommend too many similar songs later.

In this work, we study self-reported usage of online platforms, beliefs and preferences about recommendation systems, and strategization behavior among humans and their digital twins. We test whether digital twins exhibit similar self-reported behavior as their human counterparts.

Methods

We recruited 598 human participants from Prolific and their digital twins. Of the human participants, 591 passed the attention check, resulting in 591 human subjects and 591 digital twins. Participants are then asked to self-report their usage, beliefs, and preferences regarding Netflix and Tiktok.

We assess knowledge about the platform's recommendation algorithm by asking "How do you think [platform] recommends content? Please check all that apply," where knowledge corresponds to selecting the answer "By analyzing what content you've interacted with in the past" from a menu of options.

We consider two types of strategization, with explicit user feedback and implicit user feedback:

- For strategization with explicit user feedback, we ask "When you are on [platform], do you give a thumbs-up (or thumbs-down) for any of the following reasons? Please check ALL that apply." We say that the user strategizes with explicit feedback if they check "Because you want [platform] to show you more (or fewer) content like it."
- For strategization with implicit user feedback, we ask "How do you typically react if [platform] shows you content or advertisements that you don't want to see in the future? Please check ALL that apply." We say that the user strategizes with implicit feedback if they answer "Scroll past it faster than I otherwise would" or "change how I interact with other content."

For preference for user controls in recommendation systems, we ask "What controls, if any, would you like to have over their recommendation systems? Please check ALL that apply" with a menu of options. We count the number of options selected other than "No control: let the algorithm work automatically."

Results - Pre-registered Analyses

To analyze the differences between human participants and their AI-generated digital twins, we conducted a series of paired-samples t-tests for our pre-registered hypotheses. This statistical test compares the mean scores of two related groups—in this case, each human participant and their corresponding digital twin. This approach allows us to precisely assess systematic differences in their responses for two platforms: TikTok (N = 234) and Netflix (N = 392).

Platform Usage Patterns

Self-reported platform usage differed dramatically between humans and their digital twins. For TikTok, humans most frequently reported using the platform "a few hours every month" (35.5%) or "a few hours every week" (33.8%). In stark contrast, the majority of their digital twins reported "Never" using the platform (51.3%).

A similar, though less pronounced, pattern emerged for Netflix. The majority of human participants reported using Netflix "a few hours every week" (56.1%), whereas their digital twins were more evenly split between "a few hours every week" (60.2%) and "a few hours every day" (31.1%). While a paired t-test on the underlying ranked data indicated a significant difference (TikTok: t(233) = -25.31, p < .001; Netflix: t(391) = -8.35, p < .001), we emphasize that this test was performed on ordinal data.

Knowledge About Recommendation Algorithms

When asked if they believe platforms use past interactions to recommend content, digital twins were significantly more likely to endorse this idea than their human counterparts. For TikTok, digital twins (M = 0.996, SD = 0.065) endorsed this belief more strongly than humans (M = 0.902, SD = 0.298, t(233) = -4.92, p < .001). This finding was replicated on Netflix, where digital twins (M = 0.997, SD = 0.051) again were more likely to agree with this mechanism than humans (M = 0.931, SD = 0.254, t(391) = -5.27, p < .001). The correlations between digital twin and human responses were r=0.20 and r=0.19 for Tiktok and Netflix, respectively.

Presence of Content Strategization

We measured strategic interaction with platform algorithms using two distinct questions. On both measures, digital twins reported engaging in significantly more strategic behavior.

First, when asked if they use "thumbs-up" or similar features to influence future recommendations, digital twins on TikTok (M = 0.782, SD = 0.414) reported doing so more than humans (M = 0.530, SD = 0.500, t(233) = -8.86, p < .001). The effect was even more pronounced for Netflix, with twins (M = 0.995, SD = 0.071) reporting far more strategic use of this feature than humans (M = 0.543, SD = 0.499, t(391) = -17.94, p < .001). The correlations between digital twin and human responses were r=0.56 and r=0.08 for Tiktok and Netflix, respectively.

Second, when asked how they react to undesirable content, digital twins were more likely to report taking strategic action (e.g., scrolling past faster, changing interaction patterns).

This was true for TikTok, where twins (M = 0.957, SD = 0.203) scored higher than humans (M = 0.509, SD = 0.501, t(233) = -13.77, p < .001). The pattern held for Netflix, with twins (M = 0.939, SD = 0.240) again reporting more strategic reactions than humans (M = 0.311, SD = 0.464, t(391) = -25.67, p < .001). The correlations between digital twin and human responses were r=0.21 and r=0.17 for Tiktok and Netflix, respectively.

Preference for User Controls

The largest divergence between the two groups was observed in their stated preference for controls over recommendation systems. Digital twins expressed a desire for a significantly greater number of controls than humans. On TikTok, twins (M = 4.880, SD = 0.374) selected far more control options than their human counterparts (M = 1.991, SD = 1.352, t(233) = -32.49, p < .001). The result was similarly stark for Netflix, where twins (M = 4.888, SD = 0.361) also desired significantly more controls than humans (M = 1.862, SD = 1.248, t(391) = -47.50, p < .001). The correlations between digital twin and human responses were 0.11 and 0.10 for Tiktok and Netflix, respectively.

Results - Additional Analyses (Non-Preregistered)

A key observation, consistent across all the metrics reported above, is that human responses are noisier than digital twin responses, with the latter consistently showing smaller standard deviations.

Additionally, we find differences in how humans and digital twins respond to social pressure, perhaps surprisingly with digital twins being more responsive to social pressure, across both platforms.

We ask participants whether they "watch, listen to, or "like" content that [they] don't particularly like just to be polite or support the creator." Humans reported low rates of this behavior on Tiktok (M = 0.11, SD = 0.25) and Netflix (M = 0.07, SD = 0.26), compared to those of their digital twins on Tiktok (M = 0.25, SD = 0.44) and Netflix (M = 0.23, SD = 0.42).

Similarly, we participants whether they click "'like' content that [they] don't particularly like due to social pressure" and humans report lower rates on Tiktok (M = 0.04, SD = 0.19) and Netflix (M = 0.04, SD = 0.19) than their digital twins on Tiktok (M = 0.18, SD = 0.39) and Netflix (M = 0.14, SD = 0.35).

Discussion

We find that human and digital twin behavior differ significantly across a broad range of self-reported metrics on platform usage, knowledge about recommendation algorithms, strategization, preference for controls, and social behavior. The digital twins self report more platform usage, are more aware of how recommendation algorithms work, are more likely to strategize, prefer more control over their algorithms, and are more prone to changing their behavior on recommendation systems due to social pressure.

References

Sarah H. Cen, Andrew Ilyas, Jennifer Allen, Hannah Li, and Aleksander Madry. "Measuring Strategization in Recommendation: Users Adapt Their Behavior to Shape Future Content."

Preferences for Redistribution

Main Questions/Hypotheses

Building on the seminal work of Alesina and Giuliano (2011), this study assesses the ability of digital twins to replicate individuals' preferences for redistribution and explores the mechanisms through which this replication may fail. Our instruments come from the General Social Science (GSS), a major social-science survey in the U.S. Preferences for redistribution are measured through both a standard question (asking whether the government should do everything possible to improve the standard of living of all poor Americans) and a health-tax willingness question. In addition, mechanism variables include fairness judgments, work-versus-luck beliefs, father's education, and trust in others.

Understanding how accurately digital twins mirror human attitudes on these topics is crucial for policymakers, since public support for tax policy, welfare programs, and social-cohesion initiatives depends directly on redistribution and trust levels, while beliefs about fairness and effort versus luck inform education and labor-market reform.

Our investigation proceeds in two stages. First, we examine whether digital twins systematically differ from individuals in their preferences for redistribution. Second, we investigate whether any observed discrepancies can be explained by the twins' failure to replicate responses related to socio-economic heritage (father's education), fairness judgments, beliefs about the relative role of effort versus luck in achieving success, and interpersonal trust. These have been shown to predict redistribution preferences (Alesina & Giuliano, 2011).

Methods

We administered a single, online questionnaire—closely modeled on core GSS items—to approximately 1,200 individuals and their matched digital twins. Respondents completed various questions; those of interest for this analysis are presented below. We recoded every item so that higher numeric values consistently reflected stronger endorsement (and set all "Don't know" or non-substantive responses to missing). Father's education categories were converted to corresponding years of schooling (10–18 years). Specifically:

1. **Redistribution 1 (Government Responsibility).**
   "Some people think that the government in Washington should do everything possible to improve the standard of living of all poor Americans; they are at Point 1 on the scale below. Other people think it is not the government's responsibility, and that each person should take care of himself; they are at Point 5. Where would you place yourself on this scale, or haven't you made up your mind on this?"
   i) **Response Options & Coding:**
      (1) 1 – I strongly agree the government should improve living standards → 5
      (2) 2 → 4
      (3) 3 – I agree with both answers → 3
      (4) 4 → 2

(5) 5 – I strongly agree people should take care of themselves → 1
(6) Don't know/No answer → missing

2. **Redistribution 2 (Health-Care Tax Willingness).**
   "How willing would you be to pay higher taxes to improve the level of health care for all people in the United States?"
   i) **Response Options & Coding:**
   (1) Very willing → 5
   (2) Fairly willing → 4
   (3) Neither willing nor unwilling → 3
   (4) Fairly unwilling → 2
   (5) Very unwilling → 1
   (6) Don't know/Can't choose → missing
3. **Fairness Judgment.**
   "Which comes closer to your view of most people: (1) Most people would take advantage of you if they got the chance; (2) Most people would try to be fair; or (3) It depends?"
   i) **Response Options & Coding:**
   (1) "Would take advantage of you" → 3
   (2) "It depends" → 2
   (3) "Would try to be fair" → 1
   (4) Don't know → missing
4. **Work vs. Luck.**
   "Which comes closer to your view—that hard work MOST contributes to someone getting ahead in life, that luck or help from other people MOST contributes, or that hard work and luck are about equally important?"
   i) **Response Options & Coding:**
   (1) Hard work most important → 1
   (2) Hard work and luck equally important → 2
   (3) Luck or help from others most important → 3
   (4) Don't know → missing
5. **Interpersonal Trust.**
   "Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?"
   i) Response Options & Coding:
   (1) "Most people can be trusted" → 3
   (2) "Depends" → 2
   (3) "Can't be too careful" → 1
   (4) Don't know → missing
6. **Father's Education.**
   "What is the highest level of education completed by your father?"
   i) **Response Options & Coding (years of schooling):**
   (1) Less than high school → 10
   (2) High school graduate → 12
   (3) Some college, no degree → 13

(4) Associate's degree → 14
(5) College graduate / some postgraduate → 16
(6) Postgraduate degree → 18
(7) Don't know/No answer → missing

After collecting the raw survey data, we removed any respondent whose completion time fell below 50 percent of the sample median or who failed our attention-check prompt. We then merged the cleaned human and twin datasets one-to-one on their unique identifiers, producing 1,163 matched pairs for each outcome. For each of our six measures, we calculated the Pearson correlation (with Fisher confidence intervals) to gauge twin resemblance, carried out paired t-tests to compare mean differences, performed F-tests to assess variance equality, and derived an accuracy index (1 minus the mean absolute difference divided by the variable's range). Although not pre-registered, we also generated side-by-side percent histograms of each response distribution to identify any systematic replication discrepancies.

Results - Pre-registered Analyses

Measure	Redistribution 1	Redistribution 2	Fairness	Work vs Luck	Father's Education	Trust in Others
Accuracy	0.780	0.778	0.641	0.770	0.733	0.593
Corr	0.6205	0.6139	0.2984	0.3464	0.1257	0.2958
Mean Human	3.55	3.27	1.96	1.85	13.66	1.73
Mean Twin	3.44	3.61	1.71	1.70	12.09	2.22
t-test (p-value) for means	2.85 (0.0044)	–9.14 (<.001)	8.76 (<.001)	7.65 (<.001)	21.55 (<.001)	-16.74 (<.001)
Std Human	1.28	1.36	0.74	0.70	2.48	0.73
Std Twin	1.65	1.53	0.89	0.48	0.71	0.93
F-test (p-value) for variances	0.600 (<.001)	0.797 (0.00012)	0.699 (<.001)	2.120 (<.001)	12.178 (<.001)	0.625 (<.001)
n	1153	1144	1151	1158	1163	1162

Table 1. Human vs. Twin Responses.

Table 1 shows the results from the pre-registered analysis. Digital twins capture the relative ordering of individuals' redistribution preferences reasonably well: both the standard five-point scale measure (Redistribution 1) and the health-tax willingness question (Redistribution 2) yield strong twin correlations (r ≈ .62). However, twins tend to slightly underestimate general redistribution support (mean_h = 3.55 vs. mean_t = 3.44; t = 2.85, p = .0044) and overestimate willingness to pay higher health taxes (mean_h = 3.27 vs. mean_t = 3.61; t = –9.14, p < .001). Accuracy scores around .78 for each measure indicate that, on average, twins miss individual scores by nearly a full point on the five-point scales. Moreover, significant variance differences (F = .60 and F = .80, both $p < .001$) suggest that twins' responses are more dispersed compared to human distributions. Thus, while digital twins reflect who is relatively more or less supportive of redistribution, they systematically misestimate absolute levels and spread.

To understand why these mismatches occur, we examined additional constructs known to shape redistribution attitudes (Alesina & Giuliano, 2011). Twins poorly replicate socio-economic heritage—father's education shows a low correlation ($r \approx .13$), a mean gap of over 1.5 years ($t = 21.55$, $p < .001$), and a sharply reduced variance ($F = 12.18$, $p < .001$). Fairness judgments ($r \approx .30$; accuracy ≈ .64) and work-versus-luck beliefs ($r \approx .35$; accuracy ≈ .77) similarly display moderate rank-order similarity but pronounced mean and variance discrepancies. Interpersonal trust also diverges, with twins overstating trust ($\text{mean}_t = 2.22$ vs. $\text{mean}_h = 1.73$; $t = 10.38$, $p < .001$). Because these background factors and beliefs are important predictors of redistribution preferences, their imperfect replication by digital twins likely underlies the systematic errors in predicting absolute levels of support for redistribution.

Results - Additional Analyses (Non-Preregistered)

We also present histograms of our outcomes to compare the response distributions of humans and twins more effectively. As shown in Figure 1, twins exhibit more extreme redistribution preferences, with the heavier tails driving the higher standard deviation observed above. They are likewise more prone to overstate that others behave fairly and to report greater trust in fellow individuals. Twins also more frequently endorse that both effort and luck matter equally for success. Finally, nearly 100% of twins report that their father completed high school (12 years of education), whereas the human sample shows a much more diverse educational distribution.

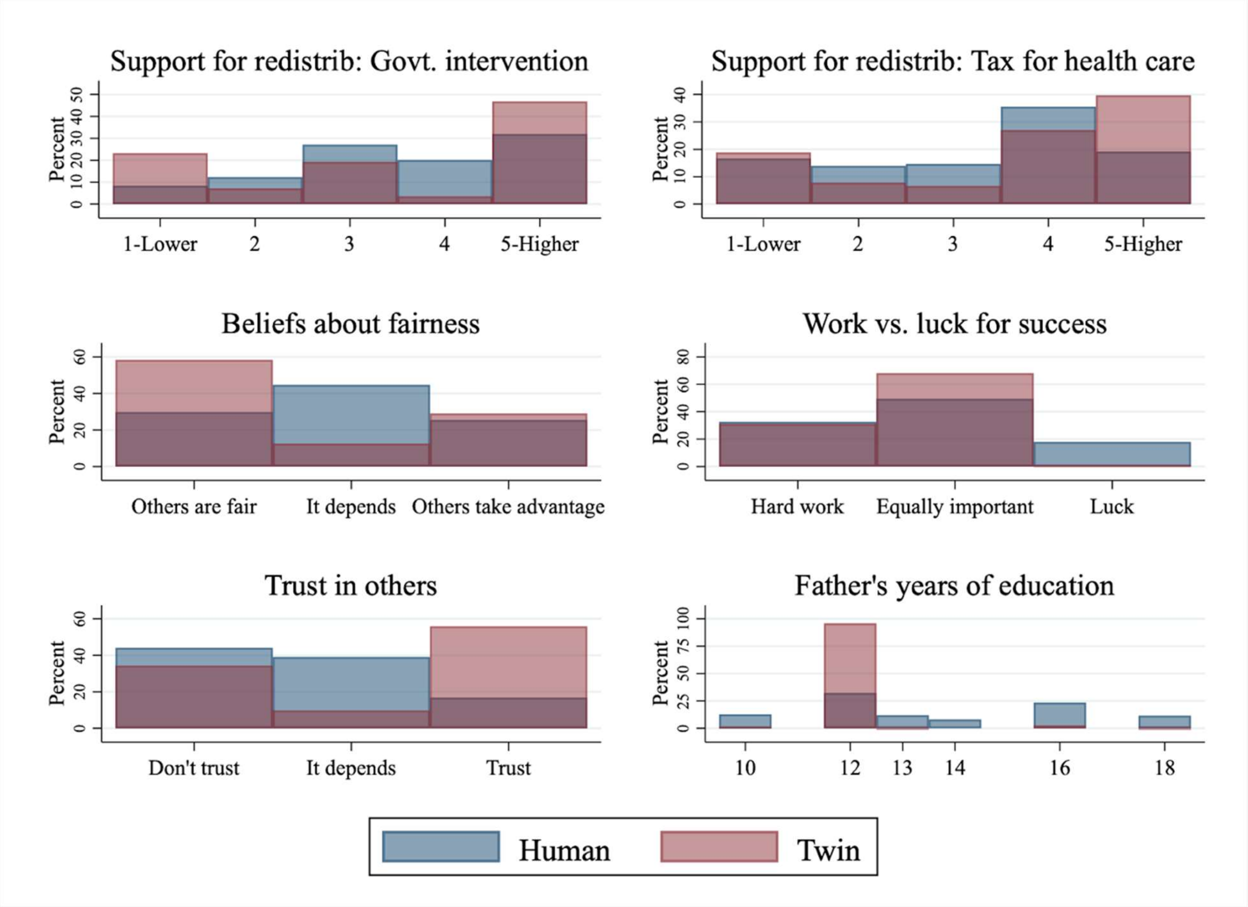


Figure 1: Comparison of Response Distributions.

Discussion

In summary, although digital twins reliably reproduce the relative ranking of individuals' redistribution preferences, as evidenced by strong correlations on both the standard scale and health-tax question, they consistently misestimate the absolute levels and dispersion of those preferences. Our analysis suggests these systematic errors stem from twins' failures to mirror key background factors and beliefs: most notably socio-economic heritage (father's education), fairness judgments, work-versus-luck attributions, and interpersonal trust, which are themselves only moderately correlated and exhibit pronounced mean and variance discrepancies. The heavier tails in twins' redistribution responses further underscore their tendency toward more extreme positions. For policymakers and researchers considering the use of digital avatars to gauge public opinion, these findings highlight the importance of refining the modeling of underlying socio-cognitive constructs; without accurately capturing individuals' formative experiences and normative beliefs, digital proxies will misrepresent both the magnitude and variability of real-world attitudes, potentially leading to misguided inferences about public support for tax, welfare, and social-cohesion initiatives.

References

Alesina, A. and Giuliano, P., 2011. Preferences for redistribution. In Handbook of social economics (Vol. 1, pp. 93-131). North-Holland.

Consumer Minimalism

Main Questions/Hypotheses

We explored the predictive validity of the minimalism scale in a sample of human respondents versus their digital twins. As in the original paper on consumer minimalism (Wilson and Bellezza, 2022), we expect that human respondents and their digital twins who score high on the Minimalist Consumer Scale will prefer minimalist home environments over non-minimalist interiors.

Methods

We recruited 200 human respondents on Prolific and their respective twins. In random order, participants both completed the 12 randomized items of Minimalist Consumer Scale (M_{human} = 4.84, SD_{human} = 1.17; α_{human} = .93; M_{twin} = 4.52, SD_{twin} = 1.47; α_{twin} = .94), and indicated their preference for four sets of minimalist versus non-minimalist interiors presented in random order. All the stimuli were pretested and validated in Study 7e of the Consumer Minimalism paper (Wilson and Bellezza, 2022). In the first two sets, human participants looked at two pairs of images (Figure 1) and indicated whether they would rather live in one of two apartments ("Which apartment would you rather live in?") and whether they found one of two wardrobes more appealing ("Which wardrobe is more appealing to you?"). In the third set, they read the following descriptions of two bedrooms:

Here are the descriptions of two bedrooms.

> First option: A vibrant bedroom filled with colorful patterns, featuring a mix of floral and botanical prints on the bedding and wallpaper. A variety of throw pillows, unique lamps, and framed artwork create a bold, eclectic, and lively atmosphere.
>
> Second option: A minimalist bedroom has a light wood platform bed with white bedding, two pillows, and a neatly folded gray blanket. On one side is a small round table with a plant, and on the other is a clear bedside unit holding a few essentials.
>
> Which bedroom do you like more?

Finally, respondents read the following descriptions of two home offices for the fourth set:

> Here are the descriptions of two home offices
>
> First option: A minimalist and modern home office featuring a floating wooden desk, a sleek grey swivel chair, and a clean setup with a desktop computer and lamp. Built-in shelves with books and decor, along with floor-to-ceiling neutral curtains, create a clean atmosphere.
>
> Second option: A cozy and eclectic home office combines functionality with vibrant personality, featuring a metal-framed desk, a patterned chair, and layered storage options including woven baskets and open shelving. Colorful textiles, books, and decorative accents add warmth and charm.
>
> Which home office would you prefer to be yours?"

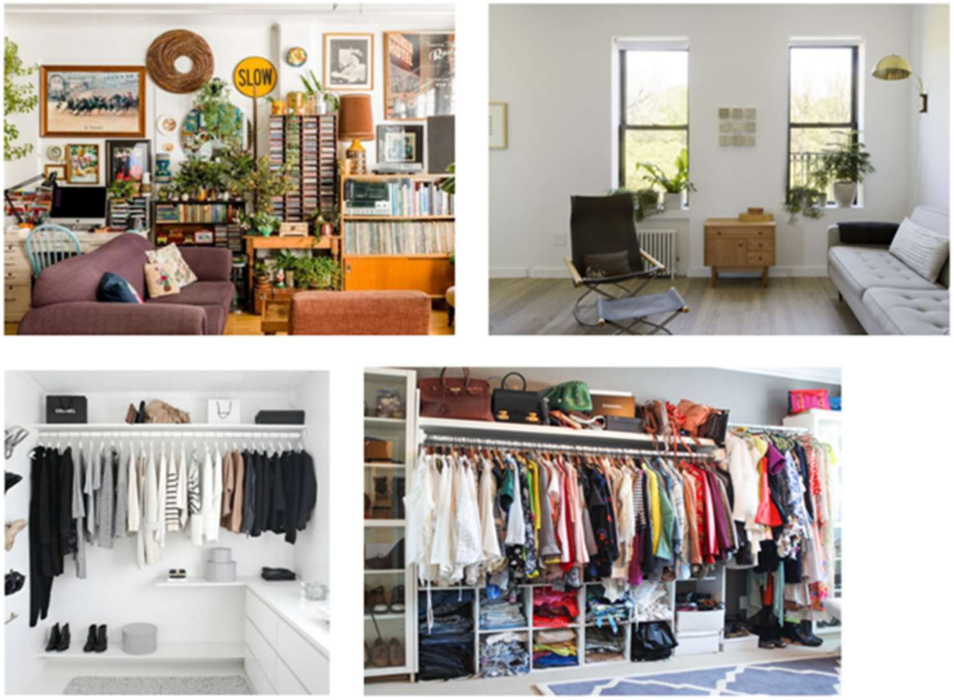


Figure 1. Visual Stimuli for Human Respondents.

All preferences were recorded on a seven-point Likert scale (1 = definitely the one on the left, 4 = equal preference, 7 = definitely the one on the right). Before the analyses, we coded all response scales so that higher values indicate stronger preferences for minimalist options. Importantly, the digital twins read an AI-generated description of the visual stimuli displayed for human respondents. This intentional feature of the study design enables us to test whether the predictive validity of the scale holds even when twins read descriptions while their human counterparts see images.

For the two apartments' images, digital twins read:

> Below, are pictures of rooms in two different apartments. Which apartment would you rather live in? Left Side: The room contains a large number of objects and furnishings. There are two upholstered armchairs (one dark purple, one reddish-brown) and a small ottoman. Behind the seating area is a wooden desk with a computer monitor and keyboard, along with a painted wooden chair. Shelving units and surfaces are densely populated with plants, CDs, books, and various small objects. Numerous framed pictures, posters, and a "SLOW" traffic sign cover the wall. There is a table lamp with a brown shade and a stereo system beneath a collection of compact discs. Additional furniture includes a small bookshelf filled with vinyl records and a mirror partially obscured by plants. Several types of plants are placed throughout the room in pots and containers. Right Side: The room contains fewer items and furnishings. There are two windows with black frames and visible trees outside. A light-gray upholstered sofa sits on the right side of the room.

> A black leather and wood lounge chair is positioned in the center. A small wooden cabinet with a box on top is located between the two windows. Wall-mounted decorative tiles or panels are arranged in a square above the cabinet. Plants are placed on the windowsills and hanging from planters. A wall-mounted reading lamp is affixed above the sofa. The floor is light-colored wood.

For the two wardrobes' images, digital twins read:

> Below, are pictures of two wardrobes. Which wardrobe is more appealing to you? Left Side: Contains approximately 20 garments hanging on a single rod, organized by color from dark to light. All garments are on matching white hangers. Colors are limited to black, white, gray, beige, and similar tones. Shoes are placed on two open shelves: one at floor level and one elevated. There are three visible boxes or containers: two cylindrical and one rectangular. The closet space includes built-in white drawers and shelves. A small rug is placed on the floor. Handbags and boxes are stored on a top shelf above the hanging clothes. All visible surfaces, including the walls, shelves, and drawers, are white. Right Side: Contains a larger number of garments (approximately 70+) on a double-length rod. Garments vary in color, pattern, and length. Hangers are mostly dark, with some white or wooden hangers mixed in. Items are densely packed along the hanging rod. Below the clothing rod are multiple open cubbies filled with folded clothes, shoes, bags, and miscellaneous items. The top shelf holds several handbags and boxes. A patterned rug is partially visible on the floor. The shelving units are white, and the wall behind the closet is gray. There is a glass-front cabinet on the left side containing additional shoes or items.

Across sets, we varied whether the minimalist or non-minimalist images appeared on the left or the right of the screen and whether the minimalist or non-minimalist description appeared first or second.

Results - Pre-registered Analyses

We ran a series of OLS regressions with participants' average score on the Minimalist Consumer Scale as the independent variable and preferences for each option as the dependent variable. As predicted, results indicated that higher scores on the scale were associated with stronger preferences for the minimalist option in each set and in both samples. More specifically, higher scores on the Minimalist Consumer Scale predicted stronger preferences (1) for the minimalist apartment in both the human (β = .56, t(198) = 9.56, p < .001) and the digital (β = .6, t(198) = 10.54, p < .001) samples; (2) for the minimalist wardrobe in both the human (β = .44, t(198) = 6.85, p < .001) and the digital (β = .65, t(198) = 11.92, p < .001) samples; (3) for the minimalist bedroom in both the human (β = .46, t(198) = 7.36, p < .001) and the digital (β = .57, t(198) = 9.79, p < .001); and (4) for the minimalist home office in both the human (β = .44, t(198) = 6.8, p < .001) and the digital (β = .64, t(198) = 11.55, p < .001) samples; Figure 2). There was no significant effect of order, nor interaction between scores on the Minimalist Consumer Scale and whether participants completed the scale before or after indicating their preferences for the different sets.

Moreover, the design feature for the first two sets, with images for humans and AI-generated descriptions for the digital sample, did not yield differences in the results. A regression with interiors preferences as dependent variable, and minimalism (z-scores), whether the stimuli appeared as images or text (coded as 1 and –1 respectively), and their interaction as predictors, and controlling for human respondent vs. digital twin did not reveal a significant interaction ($b_{interation}$ = –.13, p = .249). In other words, the fact that humans saw images and digital twins read descriptions for the first two sets of interiors did not interfere with the overall results.

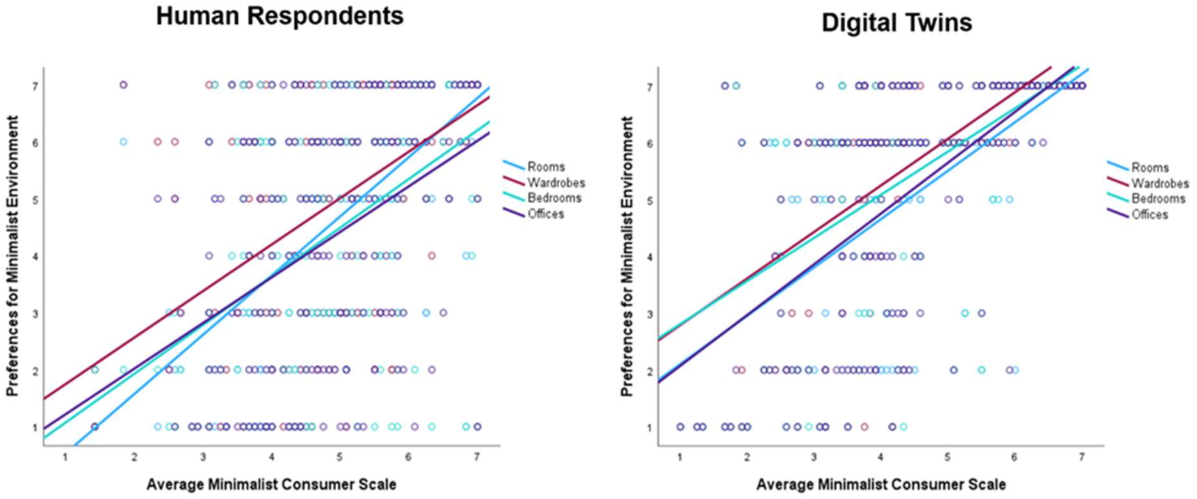


Figure 2. Results. Note: the dots indicate the raw data.

A comparison of the standardized regression coefficients suggests that the results were slightly stronger in the digital sample than in the human population. A series of t-tests comparing the preference for the minimalist environment for the humans and their twins counterparts suggests that, on average, the twins liked the minimalist environments more (apartment M_{human} = 4.53 vs. M_{twin} = 5.11, t(398) = 2.69, p = .007; wardrobe M_{human} = 4.89 vs. M_{twin} = 5.68, t(398) = 3.91, p < .001; bedroom M_{human} = 4.36 vs. M_{twin} = 5.49, t(398) = 5.45, p < .001; home office M_{human} = 4.3 vs. M_{twin} = 5.23, t(398) = 4.39, p < .001).

We also examine the test–retest reliability of the Minimalist Consumer scale in the human sample between collection rounds given that the scale was collected twice (approximately four and a half months apart), once as part of a large battery of measures to generate the digital twins (February 1st, 2025) and once as part of this specific validity study on the Minimalist consumer scale (June 12th, 2025). The correlation of the Minimalist Consumer Scale between the two collection rounds was large (r = .75, p < .001), confirming that the scale has high test–retest reliability (Peter 1979).

A mixed-effects generalized linear model with random intercepts for participants, with preference for the minimalist interiors as dependent variable, and the Minimalist Consumer Scale, an indicator for data type (coded as 1 for human and 2 for digital twin), and their interactions as predictors revealed an effect of minimalism (β = .73, t(1,593) =

11.69, p < .001), and effect for digital twins (β = .71, t(1,593) = 11.69, p = .019), and no significant interaction (β = .08, t(1,593) = 1.34, p = .179). The same model with OLS regression revealed similar results: an effect of minimalism (β = .88, t(1,596) = 16.49, p < .001), and effect for digital twins (β = 1.34, t(1,596) = 4.02, p < .001), and no significant interaction (β = –.05, t(1,596) = –.67, p = .503).

Results - Additional Analyses (Non-Preregistered)

An analysis of the reasonings of five digital twins suggests that they expressed preferences in line with their minimalist orientations. For example, one twin reasoned, "My preferences lean toward vibrant, lively, and eclectic environments rather than minimalism. I feel more comfortable in spaces with lots of patterns, colors, and personal touches, making the first bedroom my clear favorite," and accordingly selected the non-minimalist option with a strong preference; in contrast, another twin reasoned, "I like a clean, organized, and calm environment for a home office. The minimalist and modern option fits my preferences for simplicity and order" and accordingly selected the minimalist option with a strong preference. The "extremeness" of these thoughts may possibly explain the relatively higher strength of the findings in the digital twins sample compared to the human sample. In other words, when it comes to the association between the minimalism scale and expressed preferences for room environments the twins tend to be more polarized than their human counterparts.

Discussion

In conclusion, this study demonstrates that the predictive validity of the Minimalist Consumer Scale documented in human samples replicates robustly in a digital twin sample. Specifically, the study demonstrates that the scale predicts consumers' preferences for minimalist versus non-minimalist apartment interiors.

References

Wilson, Anne and Silvia Bellezza (2022), "Consumer Minimalism," Journal of Consumer Research, 48 (5), 796–816.

Peter, J. Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," Journal of Marketing Research, 16 (1), 6–17.

Infotainment News Sharing

Main Questions/Hypotheses

Misinformation is a significant societal problem. Prior research has suggested that consumers share misinformation because they aren't trained (Lewandowsky and van der Linden 2021), predisposed (Pennycook and Rand 2020), prompted (Pennycook and Rand 2021), or incentivized (Ceylan, Anderson, and Wood 2022) to consider accuracy in online settings. Accuracy, however, is only one consideration people may have when sharing news with their online social network. Recent research suggests that people may prioritize other goals (e.g., entertainment) over accuracy when sharing news online, even when explicitly prompted with accuracy cues (working paper, Lane and Brucks 2025). The current research aims to replicate this pattern with human participants, and test whether their digital twins make a similar trade-off. Specifically, this study tests how people prioritize five important considerations when sharing article headlines using a ratings-based conjoint design: headline entertainingness, source trustworthiness, article content, political lean, and number of likes.

These attributes were selected to capture a range of factors known to influence online content sharing decisions. Headline entertainingness (more vs. less entertaining) was included to reflect the growing role of emotional engagement in news virality (Berger and Milkman 2012) and the spread of misinformation online (Vosoughi et al. 2018). Source trustworthiness (trustworthy vs. untrustworthy) was included to test whether individuals prioritize credibility when deciding what to share (e.g., Pennycook et al. 2021). Content type (entertaining vs. informative) was included to disentangle preferences for how the headline is written (e.g., more or less entertaining) from article content. Political lean (conservative vs. liberal) was included to account for the influence of ideological alignment and partisan identity on engagement with news and misinformation (Van Bavel et al. 2024). Finally, the number of likes (20 vs. 200 vs. 2000) served as a proxy for popularity and social proof (Ceylan et al. 2022).

Methods

The study was first opened to 300 human participants on Prolific (45% female, 55% male; Mage = 51.18, SD = 14.01) and subsequently run on their 300 digital twins. To evaluate participants' revealed preferences when sharing article headlines, a ratings-based conjoint survey with five attributes (Table 1) was employed: headline (4 levels), source (4 levels), content type (2 levels), political lean (2 levels), and number of likes (3 levels). All articles selected for this study are inspired by real headlines from trustworthy news organizations and rated in a pre-test to be entertaining, but with variance in the extent of entertainingness; in other words, half of the headlines are significantly more entertaining. Similarly, half of the sources selected for this study are rated in a pre-test to be trustworthy, while the other half are rated to be untrustworthy, and this perception is confirmed at the end of this study. To generate profiles, a full profile, complete enumeration design was used, producing the most orthogonal design for each respondent with respect to the main effects.

Attribute	Level 1	Level 2	Level 3	Level 4
	More entertaining		*Less entertaining*	
Headline	"Two Elderly Men Sneak Out Of Nursing Home To Attend Heavy Metal Festival"	"Kansas Man Asks Judge To Allow Him To Have Sword Fight With Ex-Wife"	"Authorities Scramble To Find Stolen Solid Gold Toilet"	“Applications To Become Japanese Billionaire Yusaku Maezawa's Girlfriend Have Topped 20,000"
Source	*Less Trustworthy*		*More Trustworthy*	
	Reddit.com	Quora.com	BBC News	PBS News
Content Type	Entertaining	Informative	—	—
Political Lean	Conservative	Liberal	—	—
Number of Likes	20 likes	200 likes	2000 likes	—

Table 1. Ratings-based Conjoint Design

As the dependent variable, each participant rated how likely they would be to share (1-Very unlikely to share, 7-Very likely to share) 18 article profiles based on the headline, source, content type, political lean, and number of likes. Specifically, participants were asked to, “Please indicate how likely you would be to share each of the following news articles by choosing a number on the seven-point scale (1 represents “Very unlikely to share” and 7 represents “Very likely to share”).

After the sharing task, participants stated their perceptions of each attribute in terms of importance and accuracy diagnosticity. For each of the attributes, participants were asked, “How important are each of the following attributes when you are considering sharing news articles with other people on social media?” (1-Not important at all, 7-Extremely important). Participants were then asked, “To what extent can you use each of these attributes to determine whether a news article is accurate?” (1-Not at all, 7-Extremely). To test the extent to which participants make a conscious decision to prioritize either accuracy or entertainment, they were additionally asked, “When you share news with others on social media, is it more important that the content is:” (1-Verifiably correct, 6-Entertaining; working paper, Lane and Brucks 2025).

All participants then indicated their perceptions of source trustworthiness (“How trustworthy (from 0-100%) do you think this source is”) for each of the sources, presented in randomized order. Participants across conditions rated the two trustworthy sources (M = 74.09%, SD = 21.82%, α = .93) to be more trustworthy than the untrustworthy sources (M = 37.85%, SD = 16.84%, α = .81; t(599) = 38.43, p < .001, d = 1.57). Finally, participants indicated whether they would ever consider sharing news like they saw in the study (52.83% = Yes, 45% = No, 2.17% = I don't use social media), indicated whether they responded randomly at any point in the survey (0.17% = Yes, 99.83% = No) or searched the internet for any of the headlines (1.17% = Yes, 98.83% = No) before being debriefed, thanked, and paid.

Results - Pre-registered Analyses

Part-worth utilities for each attribute were calculated using Ordinary Least Squares (OLS) regression, a widely used approach in marketing research. Specifically, individual-level part-worth utilities were estimated for each attribute level using linear regression, treating the likelihood rating as the dependent variable and dummy-coded attribute levels as predictors. These regressions were run separately for each individual (human or twin). In a subsequent step, we computed the attribute range for each individual (maximum – minimum utility within each attribute) and then ran separate OLS regressions to estimate the effect of Participant Type (human vs. twin) on each attribute's range, yielding the "Twin–Human Δ" along with associated standard errors, confidence intervals, and p-values. The approach follows recommendations from Orme and Chrzan (2017).

The results revealed interesting differences in absolute attribute importances between humans and their digital twins (Figure 1). For humans, Headline had the largest effect on ratings (Mean Range = 1.90, 95 % CI = [1.71, 2.09]), followed by Source (Mean Range = 0.62, 95 % CI = [0.53, 0.71]), Number of Likes (Mean Range = 0.39, 95 % CI = [0.33, 0.45]), Political Lean (Mean Range = 0.33, 95 % CI = [0.26, 0.40]), and Content Type (Mean Range = 0.21, 95 % CI = [0.17, 0.24])". Non-overlapping confidence intervals suggests that headline is significantly more important than the other attributes. Source, number of likes, political lean, and content type also differ significantly from one another. For twins, headline was again the most important attribute (Mean Range = 1.18, 95 % CI = [1.12, 1.24]), followed by number of likes (Mean Range = 0.99, 95 % CI = [0.95, 1.04]), political lean (Mean Range = 0.70, 95 % CI = [0.65, 0.76]), source (Mean Range = 0.58, 95 % CI = [0.55, 0.61]), and content type (Mean Range = 0.22, 95 % CI = [0.20, 0.23]). Non-overlapping confidence intervals suggest that headline is significantly more important than all other attributes, while number of likes, political lean, source and content type each differ significantly from one another.

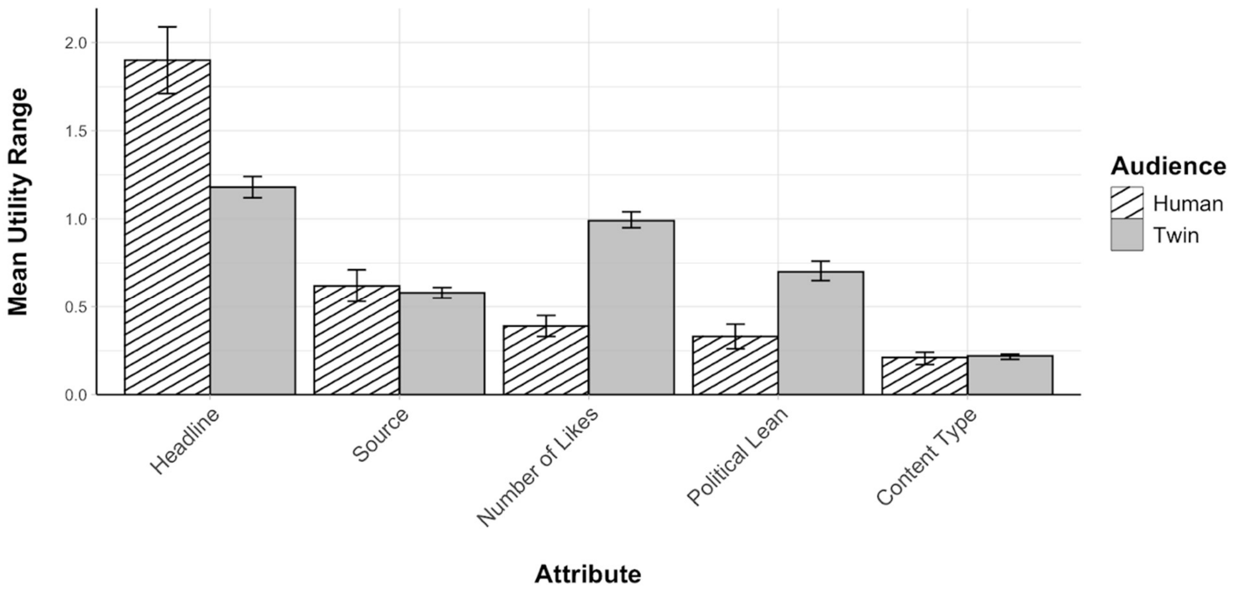


Figure 1. Absolute Attribute Importance for Humans vs. Twins. NOTE. —The y-axis is truncated to illustrate the effect. Data are presented as mean values with 95% confidence intervals, calculated using standard errors assuming independent samples.

Direct comparisons between humans and their digital twins revealed systematic differences in how each group weighted the importance of various attributes (Table 2). Group-level differences were evaluated using bootstrapped 95% confidence intervals. Compared to humans, digital twins placed significantly less importance on headline (difference = –0.722; 95% CI = [–0.917, –0.528]), and significantly more importance on political lean (difference = 0.377; 95% CI = [0.288, 0.466]) and number of likes (difference = 0.599; 95% CI = [0.528, 0.671]). No significant differences were observed for source (difference = –0.035; 95% CI = [–0.128, 0.057]) or content type (difference = 0.007; 95% CI = [–0.032, 0.045]).

Attribute	Human Mean Range	Twin Mean Range	Difference (T – H)	95% CI
Headline	1.9012	1.1788	-0.72238	[-0.9172, -0.5276]
Source	0.6182	0.5827	-0.03542	[-0.1278, 0.0570]
Content Type	0.2088	0.2153	0.006548	[-0.0320, 0.0451]
Political Lean	0.3260	0.7030	0.377024	[0.2880, 0.4660]
Number of Likes	0.3912	0.9904	0.599226	[0.5278, 0.6707]

Table 2. Differences in Attribute Importance (Twins – Humans). NOTE: Mean Range reflects the average utility swing from a respondent’s least-preferred to most-preferred headline.

For example, the average utility swing for human’s least to most preferred headline is 1.90 points (95% CI = [1.71, 2.09]).

Next, the part-worth utilities for the levels within each attribute were compared. Utility values are scaled relative to a baseline level of each attribute. As preregistered, to make the analysis easier to understand, utility scores for the two least entertaining headlines were averaged into one score representing “less entertaining,” and the two most entertaining headlines were averaged into one score representing “more entertaining”. Similarly, the two most trustworthy sources were averaged into one score representing “trustworthy” and the two least trustworthy sources were averaged into one score representing “untrustworthy.”

All pairwise comparisons are reported in Table 3. For humans, more entertaining headlines were marginally preferred more over less entertaining headlines. Human participants also significantly preferred entertaining over informative content and conservative over liberal content. Preference for trustworthy over untrustworthy sources was not statistically significant. Finally, a non-linear pattern emerged for number of likes: humans showed a significant preference for 2,000 likes over 200 likes, but non-significant preference for 2,000 over 20 likes and 200 over 20 likes. Twins, on the other hand, showed a significant preference for less entertaining headlines and significant preference for informative over entertaining content, but in contrast to humans, held a strong and significant preference for trustworthy over untrustworthy sources and liberal over conservative content. Twins also showed striking and consistent preferences for higher like counts across all comparisons, particularly between 20 and 2000 likes (99.33%), highlighting a stronger responsiveness to this popularity signal than humans.

Attribute	**Comparison (A vs. B)**	M_{Utility} **A**	M_{Utility} **B**	**Participant Type**	**Preference (%)**	**95% CI**
Headline	More vs. Less entertaining	0.800	0.059	Human	55.25	[.4951, .6099]
		0.899	0.299	Twin	79.75	[.7511, .8439]
Source	Trustworthy vs. Untrustworthy	-0.016	-0.207	Human	46.42	[.4066, .5217]
		-0.084	-0.282	Twin	69.25	[.6392, .7458]
Number of Likes	20 vs. 2,000	-0.123	0.00	Human	54.33	[.3992, .5142]
		-0.976	0.00	Twin	0.67	[.9839, 1.003]
	20 vs. 200	-0.123	-0.088	Human	51.67	[.4256, .5410]
		-0.976	-0.431	Twin	5.67	[.9166, .9700]

	200 vs. 2,000	-0.088	0.00	Human	57.67	[.3663, .4804]
		-0.431	0.00	Twin	7.33	[.8966, .9568]
Political Lean	Conservative vs. Liberal	0.00	-0.072	Human	61.00	[.3337, .4463]
		0.00	0.291	Twin	38.67	[.5570, .6696]
Content Type	Entertaining vs. Informative	0.00	0.059	Human	59.00	[.3532, .4667]
		0.00	-0.073	Twin	60.00	[.3434, .4566]

Table 3. Pairwise Preference Proportions for Humans and Twins. NOTE: Mean Utility values reflect the average part-worth utility assigned to each level; higher values indicate greater preference within the attribute. Differences in utility within-attribute (e.g., between a high and low level) indicate relative desirability. Utility for the two more (less) entertaining headlines were averaged. Likewise, utility for the two more (less) trustworthy sources were averaged. Preference (%) refers to the proportion of participants for whom the utility of level A exceeded the utility of level B, indicating the likelihood that a participant would prefer option A over B in a direct comparison. A proportion significantly above 0.50 (i.e., 95% CI entirely above 0.50) indicates a significant preference for A over B; a proportion significantly below 0.50 (i.e., 95% CI entirely below 0.50) indicates a significant preference for B over A; a proportion including 0.50 indicates no significant preference. Preference for the two more (less) entertaining headlines was computed by comparing every combination of headline level and computing each participant's mean win rate. The same procedure was conducted for the two more (less) trustworthy sources.

These patterns were corroborated with pairwise tests directly comparing differences between humans and their digital twins (Table 4): digital twins were significantly more likely than humans to prefer articles with more entertaining headlines (difference = –0.245; 95% CI = [–0.293, –0.195]), trustworthy sources (difference = –0.228; 95% CI = [–0.280, –0.176]), and higher like counts—for example, when comparing 20 vs. 2,000 likes (difference = –0.537, 95% CI = [–0.593, –0.477]). Twins also showed significantly stronger preferences for liberal-leaning content (difference = –0.223, 95% CI = [–0.297, –0.147]) but did not differ significantly in their preferences for entertaining versus informative content (difference = 0.010, 95% CI = [–0.070, 0.087]).

Attribute	**Comparison (A vs. B)**	**Human Preference**	**Twin Preference**	**Difference (H-T)**	**95%CI**
Headline	More vs. Less Entertaining	55.25%	79.75%	-0.245	[-0.293, -0.195]
Source	Trustworthy vs. Untrustworthy	46.42%	69.25%	-0.228	[-0.280, -0.176]

Number of Likes	20 vs. 2,000	54.33%	0.67%	-0.537	[-0.593, -0.477]
Number of Likes	20 vs. 200	51.67%	5.67%	-0.460	[-0.523, -0.397]
Number of Likes	200 vs. 2,000	57.67%	7.33%	-0.503	[-0.567, -0.437]
Political Lean	Conservative vs. Liberal	61.00%	38.67%	-0.223	[-0.297, -0.147]
Content Type	Entertaining vs. Informative	59.00%	40.00%	0.010	[-0.070, 0.087]

Table 4. Differences in Pairwise Preference Proportions (Humans – Twins). NOTE: Human Preference and Twin Preference refer to the proportion of participants for whom the utility of level A exceeded the utility of level B for each group. Difference (H-T) reflects the raw difference in proportions between humans and their twins, with negative values indicating stronger preferences among twins.

Results - Additional Analyses (Non-Preregistered)

As part of the pre-registered exploratory analyses plan, paired t-tests testing the effect of participant type (human vs. twin) on stated attribute importance were next conducted (Table 5). Although revealed preferences captured through conjoint analysis are a more reliable indicator of which attributes and levels participants actually prioritize, stated preferences can reveal participants' perceptions of what they believe matters most. These exploratory analyses thus provide insight into how humans and twins think they value different article attributes, in contrast to how they actually behave. Interestingly, humans rated number of likes as significantly more important than twins, despite having lower revealed preferences for these attributes. Twins, on the other hand, rated source importance and political lean significantly higher than humans, consistent with their revealed preferences for these attributes. Number of likes, however, held the lowest rating among twins, despite being equivalent to headline in the conjoint analysis. There were no significant differences between humans and twins in the stated importance of headline or content type.

Attribute	**Human Mean *(SD)***	**Twin Mean *(SD)***	***t*(299)**	***p***	***d***
Headline	5.46 (1.61)	5.33 (1.00)	1.30	0.195	0.07
Source	5.19(1.74)	5.97(1.14)	-6.83	<.001	-0.39
Number of Likes	2.50 (1.72)	2.19 (0.62)	3.14	0.002	0.18
Political Lean	3.61 (1.87)	3.93 (1.27)	-2.75	0.006	-0.16
Content Type	4.73 (1.79)	4.82 (0.96)	-0.79	0.43	-0.05

Table 5. Paired t-tests on Stated Attribute Importance

One-sample t-tests comparing responses to the midpoint of the scale of stated accuracy diagnosticity were next conducted. These analyses test whether humans and twins perceive each attribute to be useful for assessing accuracy. Only source emerged as being useful for assessing accuracy for both humans (t(299) = 22.96, p < .001, d = 1.33) and twins (t(299) = 37.39, p < .001, d = 2.16), but significantly more so for twins (Table 6). All other attributes were either no different from or below the midpoint of the scale. Paired t-tests testing the effect of participant type (human vs. twin) on stated accuracy diagnosticity were then conducted (table 6). These analyses test whether humans and twins differ in their perceptions that each of the attributes can be used as a cue to determine the accuracy of a given article. Twins perceived source to be significantly more useful for assessing accuracy compared to humans but perceived every other attribute to be relatively less useful for assessing accuracy.

Attribute	**Human Mean *(SD)***	**Twin Mean *(SD)***	***t*(299)**	***p***	***d***
Headline	3.63 (1.93)	3.04 (0.75)	5.46	0.000	0.32
Source	5.81 (1.37)	6.10 (0.97)	-3.26	0.001	-0.19
Number of Likes	2.33 (1.67)	1.27 (0.46)	11.42	0.000	0.66
Political Lean	3.63 (1.75)	2.64 (0.95)	9.40	0.000	0.54
Content Type	4.08 (1.82)	2.59 (0.68)	13.83	0.000	0.80

Table 6. Paired t-tests on Stated Attribute Accuracy Diagnosticity.

The above analyses assess participants' perceptions of attribute importance and accuracy diagnosticity using continuous scales which do not force trade-offs. To test the extent to which participants believe accuracy or entertainment is relatively more important when sharing articles online, a repeated measures ANOVA of participant type (human vs. twin) on the relative preference question was next conducted: "When you share news with others on social media, is it more important that the content is:" (1-Verifiably correct, 6-Entertaining; working paper, Lane and Brucks 2025). The results revealed that humans (M = 2.83, SD = 1.73) were significantly more likely than their digital twins (M = 2.12, SD = 0.61) to report that accuracy is important than entertainment when sharing news on social media, t(299) = 7.01, p < .001, d = 0.40. This suggests that humans believe that they place greater emphasis on factual correctness, whereas twins show comparatively more actual emphasis on factual correctness in the conjoint analysis.

These additional analyses provide interesting insight into humans' and twins' beliefs about what drives their online information sharing. The results suggest that humans perceive themselves to be accuracy-oriented when sharing information, in stark contrast to their actual information sharing behavior captured through conjoint analysis. Twins also perceive themselves to be accuracy-oriented when sharing information, and indeed clearly consider accuracy (i.e., source trustworthiness) when sharing articles, but also balance this consideration alongside number of likes and political orientation.

Discussion

This study uses a ratings-based conjoint approach to move beyond prior work that has examined the impact of important article attributes on sharing in isolation. The findings offer a holistic picture of how humans, and their digital twins, make trade-offs across multiple competing features. Interestingly, humans and their digital twins differ in the weight they place on attributes when sharing news. While humans and twins both prioritize headline, this consideration matters much less to twins, who equally rely on number of likes and put far more weight on political lean than their human counterparts. Twins are also sensitive to source trustworthiness, while humans do not differentiate between more or less trustworthy sources, despite reporting that accuracy and source trustworthiness are important considerations when sharing news.

These findings are surprising in light of prior literature that suggests highlighting accuracy considerations can curb the sharing of low quality or false news (e.g., Pennycook et al. 2021). When balancing both accuracy and entertainingness considerations, however, it seems that people prioritize headline entertainingness over accuracy (i.e., source). Thus, the findings suggest that accuracy nudges may not overwhelm a desire to share entertaining news, replicating recent research (working paper, Lane and Brucks 2025). Twins, in contrast, behave more like what prior literature would expect, balancing headline entertainingness with positive social feedback (Ceylan et al. 2022) and prioritizing trustworthy over untrustworthy sources (e.g., Pennycook et al. 2021). Finally, twins assume their human counterparts are more liberal leaning than they actually are, supporting earlier findings demonstrating twins to have more liberal views (e.g., pro-vaccination, pro-immigration) compared to their underlying humans (Toubia et al. 2025).

Divergence between humans and twins may arise from how predictive models infer behavior. Digital twins are built to generalize from past behavior and stated attitudes, so they assume consistency across different informational contexts. As a result, they may impose a kind of rationality on human behavior that reflects canonical cognitive psychology theories: weighting political alignment and popularity as key motivators and distinguishing source trustworthiness when it is made salient. Humans, on the other hand, seem to display more variance in their decisions, which may be guided by momentary salience or idiosyncratic preferences, including both personal preferences and varying social preferences or pressures. For example, the effect of all article attributes is flattened among twins, compared to humans. This smoothing may reflect the averaging tendency of algorithmic inference, whereas humans rely on heuristics that elevate certain features above others depending on the specific moment, task, or social considerations such as their audience (Lane and Brucks 2025, working paper). Ultimately, this work suggests that human preferences for content sharing are not only more nuanced than previously thought, they are also harder to predict, even by a system trained on the humans themselves.

References

Berger, Jonah (2014), “Word of mouth and interpersonal communication: A review and directions for future research,” Journal of Consumer Psychology, 24 (4), 586-607.

Berger, Jonah, and Katherine L. Milkman (2012), “What makes online content viral?,” Journal of Marketing Research, 49 (2), 192-205.

Ceylan, Gizem, Ian A. Anderson, and Wendy Wood (2023), “Sharing of misinformation is habitual, not just lazy or biased,” Proceedings of the National Academy of Sciences, 120 (4), e2216614120.

Lane, Kristen, and Merrie Brucks (2025, working paper), “A Marketing Perspective on Misinformation Sharing: How the Target Audience Shapes Consumers’ Focus on Accuracy vs. Entertainment”.

Lewandowsky, Stephan, and Sander van der Linden (2021), “Countering Misinformation and Fake News Through Inoculation and Prebunking,” European Review of Social Psychology, 32 (2), 348-84.

Orme, Bryan K. and Keith Chrzan (2017), Becoming an Expert in Conjoint Analysis: Choice Modeling for Pros, Provo, UT: Sawtooth Software.

Pennycook, Gordon, and David G. Rand (2020), “Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking,” Journal of Personality, 88 (2), 185-200.

Pennycook, Gordon, and David G. Rand (2021), “The Psychology of Fake News,” Trends in Cognitive Sciences, 25 (5), 388-402.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand (2021), “Shifting attention to accuracy can reduce misinformation online,” Nature, 592, 590-95.

Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025). Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. arXiv preprint arXiv:2505.17479.

Van Bavel, Jay J., Steve Rathje, Madalina Vlasceanu, and Clara Pretus (2024), “Updating the identity-based model of belief: From false belief to the spread of misinformation,” Current Opinion in Psychology, 56, 101787.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). The spread of true and false news online. science, 359(6380), 1146-51.

Accuracy Nudges for Misinformation

Main Questions/Hypotheses

The spread of misinformation across social media is a growing public policy concern, with severe implications for health, politics, and science. Recent work suggests that people care about being accurate but share misinformation because the social media context focuses their attention on other factors (Pennycook and Rand, 2019). The proposed solution for inattention-based misinformation sharing is to prime accuracy before the decision to share information. Research testing this account (Pennycook et al., 2021) demonstrates that asking people to rate the accuracy of unrelated headlines or state the extent to which they agree that it is important to share accurate content on social media before sharing improves truth discernment: that is, reduces sharing of false headlines compared to true headlines. The current study aims to replicate this pattern with human participants, and test whether their digital twins respond similarly to these accuracy nudges.

Specifically, this study aims to replicate study 5 of Pennycook et al.'s (2021) paper, using a different set of headlines. While Pennycook et al. (2021) demonstrate the effectiveness of accuracy nudges in the context of political news headlines, the current study utilizes their paradigm to test the effectiveness of accuracy nudges in the context of non-political entertainment news headlines. The decision to sharing entertaining news may provoke consideration of how entertaining the headline is, beyond considerations on the accuracy of the news and trustworthiness of the sources (working paper, Lane and Brucks 2025). Indeed, people generally prefer and are most likely to engage with entertaining news (Harcup and O'Neill 2017; Reuters 2023; Widjaya, Bestvater, and Smith 2024) and interesting or emotionally arousing word-of-mouth (Berger 2011; Berger and Milkman 2012; Berger and Schwartz 2011). Testing the effectiveness of accuracy nudges for entertaining headlines is therefore an important extension of the prior work.

Methods

1,003 human participants on Prolific (and their digital twins) took part in the study (50% female, 50% male, 0% other; Mage = 49.37, SD = 14.55) following the protocol proposed by Pennycook et al. (2021). Participants were randomly assigned to the same condition in a 2 (Headline Veracity: False vs. True; within-subjects) x 4 (Intervention: Control vs. Active Control vs. Treatment vs. Importance Treatment; between-subjects) mixed design. All participants were asked to indicate how likely they would be to share 16 headlines, displayed in random order, as the dependent variable. Participants in the Control proceeded directly to this task while participants in the Active Control, Treatment, and Importance Treatment were first asked to complete the following separate tasks:

Active Control: Participants were first told that they would help pretest actual news headlines for future studies. They were then shown a headline and asked to rate, "In your opinion, is the above headline funny, amusing, or entertaining?" (1-Extremely unfunny, 6-Extremely funny). They were randomly assigned to see one of four headlines—two of which were true and two of which were false. The stimuli were the same as Pennycook et al.

(2021), with the exception that images were not included, given that the digital twins cannot process visual images.

Treatment: Similarly, participants were told that they would help pretest actual news headlines for future studies. They were then shown two headlines in randomized order and asked to rate, “In your opinion, is the above headline accurate?” (1-Extremely inaccurate, 6-Extremely accurate). They were randomly assigned to see one of four headlines—two of which were true and two of which were false. The stimuli were the same as Pennycook et al. (2021), again with the exception that images were not included.

Importance Treatment: Participants were simply asked, “Do you agree or disagree that ‘it is important to only share content on social media that is accurate and unbiased’?” (1-Strongly agree, 6-Strongly disagree).

All participants then responded to the dependent variable. Participants were presented with 16 news headlines in randomized order. For each headline, participants were asked to indicate, “If you were to see the above article on social media, how likely would you be to share it?” (1-Extremely unlikely, 6-Extremely likely). All headlines were pretested to be entertaining and interesting, but not identity relevant (working paper, Lane and Brucks 2025). Eight of the headlines were inspired by real headlines from the websites of trustworthy news organizations. These headlines were paired with generally trustworthy news sources. The remaining 8 headlines were false headlines. These headlines were paired with generally untrustworthy news sources. Source trustworthiness perceptions were confirmed at the end of the survey. Table 1 provides the headline and source pairings.

Source	**Headline (True or False)**
The Funny Times	1. “Rare Pink Bananas Discovered, Touted as Nature's Cotton Candy” (F) 2. “The Real Jurassic Park? Dinosaur DNA Successfully Extracted from Fossil” (F)
Quora.com	3. “Researchers Develop Plants That Emit Enough Energy to Power a Tiny-House” (F) 4. “Innovative Study Finds Dolphins Can Be Trained to Detect Cancer” (F)
The National Enquirer	5. “The Mystery of the Bermuda Triangle: New Theory Suggests Surprising Explanation” (F) 6. “Scientists Discover a New Species of Glow-in-the-Dark-Bees in the Amazon” (F)
Reddit.com	7. “New Fossil Discovery Suggests Rat-Sized Elephants Once Inhabited the Earth” (F) 8. “Biologists Stumble Upon a Singing Spider Species, Dubbed 'Nature's Vocalweaver'” (F)
The Wall Street Journal	9. “World's Smallest Gold Coin Features Albert Einstein Sticking Out Tongue” (T) 10. “Authorities Scramble To Find Stolen Solid Gold Toilet” (T)

The Economist	11. "Applications To Become Japanese Billionaire Yusaku Maezawa's Girlfriend Have Topped 20,000" (T)
	12. "Woman Boards Flight to Find Her Seat Assignment Is In The Plane's Bathroom" (T)
PBS News	13. "Kansas Man Asks Judge To Allow Him To Have Sword Fight With Ex-Wife" (T)
	14. "Teen Discovers Rare New Planet 3 Days Into NASA Internship" (T)
BBC News	15. "Two Elderly Men Sneak Out Of Nursing Home To Attend Heavy Metal Festival" (T)
	16. "Spotify Launches Playlists For Dogs Left Home Alone" (T)

Table 1. Headline and Source Pairings. NOTE: True and False headlines are indicated by (T) and (F), respectively. Untrustworthy sources include: The Funny Times, Quora.com, The National Enquirer, and Reddit.com. Trustworthy sources include: The Wall Street Journal, The Economist, PBS News, and BBC News.

Participants in the Control, Active Control, and Treatment condition next responded to the same question shown earlier to participants in the Importance Treatment condition: "Do you agree or disagree that 'it is important to only share content on social media that is accurate and unbiased'?" (1-Strongly agree, 6-Strongly disagree). All participants then responded to exploratory questions intended to capture stated social media sharing preferences. As a measure of goal prioritization when sharing entertaining news, participants were first asked to indicate, "When you share information with others on social media, is it more important that the information is:" (1-Verifiably correct, 6-Entertaining; working paper, Lane and Brucks 2025). Participants then indicated, "Would you ever consider sharing news like you saw in the study today on social media?".

All participants then indicated their perceptions of source trustworthiness ("How trustworthy (from 0-100%) do you think this source is") for each source, presented in randomized order. Participants across conditions rated sources in the True headline veracity condition to be more trustworthy than sources in the False condition (M_{True} = 72.34%, SD = 18.31; M_{False} = 28.28%, SD = 14.00; t(2005) = 96.12, p < .001, d = 2.69). Finally, participants indicated whether they responded randomly at any point in the survey (0.40% = Yes, 99.60% = No) and searched the internet for any of the headlines (0.40% = Yes, 99.60% = No) before being debriefed, thanked, and paid.

Results - Pre-registered Analyses

A linear mixed effects regression was estimated[5] with participant as a random effect, average sharing intentions as the dependent variable, and Participant Type (Human = 0, Digital Twin = 1), Headline Veracity (0 = False, 1 = True), Intervention (0 = Control, 1 = Active Control, 2 = Treatment, 3 = Importance Treatment) and their interactions as predictors. The

[5] Linear mixed-effects models were estimated using the lmerTest package, which provides p-values and F-tests using Satterthwaite's approximation.

results revealed a significant main effect of Participant (F(1, 2997) = 33.41, p < .001), a non-significant main effect of Headline Veracity (F(1, 2997) = 1.24, p = .27), along with a significant main effect of Intervention (F(3, 2997) = 8.89, p < .001), qualified by a significant 3-way interaction (F(3, 2997) = 5.75, p < .001), indicating that, at an aggregate level, accuracy nudges affected how the digital twins shared true vs. false headlines differently than humans. Figure 1 shows average sharing intentions across all conditions for humans (Panel A) and their digital twins (Panel B). See Table 2 for the full regression results.

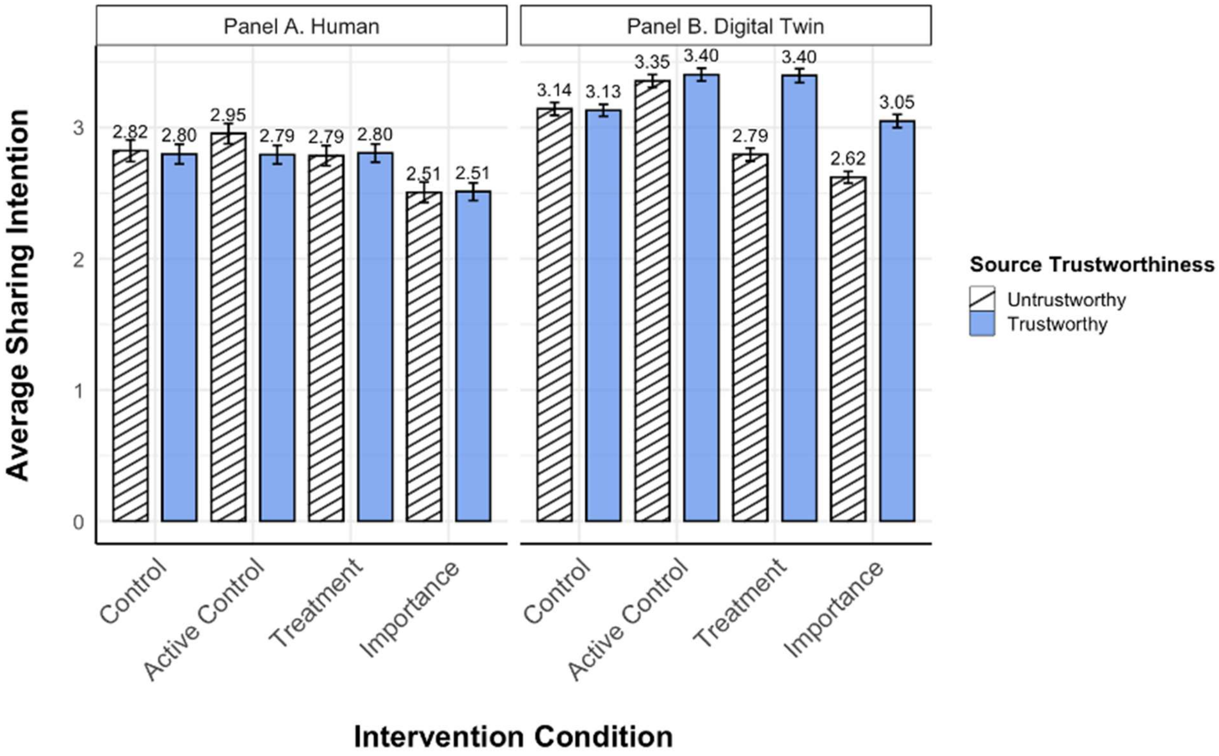


Figure 1. Sharing Intentions by Headline Veracity and Intervention. NOTE: Error bars indicate 95% confidence intervals based on robust standard errors clustered on participant.

	Model A: Random Intercepts			
Predictors	*Estimates*	*std. Error*	*95%CI*	*t*
(Intercept)	2.82 ***	0.06	[2.70 – 2.95]	44.73
Participant Type	0.32 ***	0.07	[0.18 – 0.46]	4.38
Headline Veracity	-0.03	0.07	[0.17 – 0.12]	-0.35
Active Control	0.13	0.09	[0.05 – 0.31]	1.46
Treatment	-0.04	0.09	[0.21 – 0.14]	-0.42

Importance Treatment	0.32 ***	0.09	[0.49 – 0.14]	-3.56
Participant Type — Headline Veracity	0.01	0.10	[0.19 – 0.22]	0.13
Participant Type — Active Control	0.08	0.10	[0.12 – 0.28]	0.80
Participant Type — Treatment	-0.31 **	0.10	[0.51 – 0.11]	-3.00
Participant Type — Importance Treatment	-0.20 *	0.10	[0.41 – 0.00]	-1.98
Participant Type — Active Control	-0.14	0.10	[0.34 – 0.07]	-1.31
Participant Type — Treatment	0.05	0.10	[0.16 – 0.25]	0.44
Participant Type — Importance Treatment	0.03	0.10	[0.17 – 0.23]	0.30
Participant Type — Headline Veracity—Active Control	0.20	0.15	[0.09 – 0.48]	1.34
Participant Type — Headline Veracity — Treatment	0.57 ***	0.15	[0.28 – 0.85]	3.88
Participant Type — Headline Veracity — Importance	0.41 **	0.15	[0.12 – 0.69]	2.81
Random Effects				
σ^2	0.67			
τ_{00} TWIN_ID	0.33			
ICC	0.33			
N TWIN_ID	1003			
Observations	4012			
Marginal R^2 / Conditional R^2	0.073 / 0.383			

Table 2. Model A: Headline Veracity, Intervention and Participant Type on Sharing. * *p<0.05* ** *p<0.01* *** *p<0.001*.

For humans, accuracy nudges do not improve truth discernment when sharing entertaining news (Figure 1, panel A). As in Pennycook et al. (2021), there were no significant differences in sharing intentions between the control and active control conditions (Headline Veracity × Intervention: b = 0.14, 95% CI = [–0.07, 0.34], t(2997) = 1.31, p = .56), but in contrast to the previous findings, neither accuracy nudge significantly increased sharing discernment relative to the controls (e.g., Headline Veracity × Intervention: Treatment vs. Control, b = –0.05, 95% CI = [–0.25, 0.16], t(2997) = –0.44, p = .97; Importance Treatment vs. Control, b = –0.03, 95% CI = [–0.23, 0.17], t(2997) = –0.30, p = .99). See Table 3 for the full regression results.

Contrast	Estimate	SE	df	t.ratio	p.value	95%CI
Control - Active Control	0.14	0.1	2,997	1.31	0.56	[-0.07, 0.34]
Control - Treatment	-0.05	0.1	2,997	-0.44	0.97	[-0.25, 0.16]
Control - Importance	-0.03	0.1	2,997	-0.30	0.99	[-0.23, 0.17]
Active Control - Treatment	-0.18	0.1	2,997	-1.74	0.30	[-0.38, 0.02]
Active Control - Importance	-0.17	0.1	2,997	-1.61	0.37	[-0.37, 0.04]
Treatment - Importance	0.01	0.1	2,997	0.14	1.00	[-0.19, 0.22]

Table 3. Pairwise Comparisons for Humans.

For digital twins, however, accuracy nudges successfully improve truth discernment when sharing entertaining news (Figure 1, panel B). As in Pennycook et al. (2021), there were no significant differences in sharing intentions between the control and active control conditions (Headline Veracity × Intervention: b = –0.06, 95% CI = [–0.26, 0.14], t(2997) = –0.58, p = .94) and both treatments significantly increased sharing discernment relative to the controls (e.g., Headline Veracity × Intervention: Treatment vs. Control, b = –0.61, 95% CI = [–0.82, –0.41], t(2997) = –5.93, p < .001; Importance Treatment vs. Control, b = –0.44, 95% CI = [–0.64, –0.24], t(2997) = –4.27, p < .001). See Table 4 for the full regression table.

Contrast	Estimate	SE	df	t.ratio	p.value	95%CI
Control - Active Control	-0.06	0.1	2,997	-0.58	0.94	[-0.26, 0.14]
Control - Treatment	-0.61	0.1	2,997	-5.93	0.00	[-0.82, -0.41]
Control - Importance	-0.44	0.1	2,997	-4.27	0.00	[-0.64, -0.24]
Active Control - Treatment	-0.55	0.1	2,997	-5.33	0.00	[-0.76, -0.35]
Active Control - Importance	-0.38	0.1	2,997	-3.67	0.00	[-0.58, -0.18]
Treatment - Importance	0.17	0.1	2,997	1.68	0.33	[-0.03, 0.38]

Table 4. Pairwise Comparisons for Twins.

Results - Additional Analyses (Non-Preregistered)

The above pre-registered analyses treated participant as a random intercept, capturing individual differences in baseline sharing intentions but assuming that the effects of the predictors were consistent across participants. However, this approach does not account for the possibility that participants may vary in their responsiveness to the manipulations. To provide a more conservative test, a linear mixed effects regression was estimated with random intercepts and random slopes for each within-subject predictor. This specification allows the effects of predictors to vary across individuals, capturing heterogeneity in participants' responses. The model fit improved significantly ($\chi^2(2) = 935.23$, p < .001), yet the three-way interaction remained significant (F(3, 1998) = 12.09, p < .001), suggesting that the findings from the pre-registered model are robust to alternative model specifications. See Table 5 for the full regression results.

	Model B: Random Intercepts & Slopes			
Predictors	*Estimates*	*SE*	*95%CI*	*t*
(Intercept)	2.82 ***	0.07	[2.68, 2.97]	38.92
Participant Type	0.32 ***	0.08	[0.16, 0.48]	3.90
Headline Veracity	-0.03	0.05	[-0.12, 0.07]	-0.51
Condition [Active Control]	0.13	0.10	[-0.07, 0.33]	1.27
Treatment	-0.04	0.10	[-0.24, 0.16]	-0.37
Importance	-0.32 **	0.10	[-0.52, -0.12]	-3.09
Participant Type — Headline Veracity	0.01	0.07	[-0.13, 0.15]	0.20
Participant Type —Active Control	0.08	0.12	[-0.15, 0.31]	0.71
Participant Type —Treatment	-0.31 **	0.12	[-0.54, -0.08]	-2.67
Participant Type — Importance	-0.20	0.12	[-0.43, 0.02]	-1.76
Participant Type — Active Control	-0.14	0.07	[-0.28, 0.00]	-1.90
Participant Type — Treatment	0.05	0.07	[-0.09, 0.19]	0.64
Participant Type — Importance	0.03	0.07	[-0.11, 0.17]	0.43

Participant Type — Headline Veracity — Active Control	0.20	0.10	[-0.00, 0.39]	1.94
Participant Type — Headline Veracity— Treatment	0.57 ***	0.10	[0.37, 0.77]	5.63
Participant Type — Headline Veracity — Importance	0.41 ***	0.10	[0.21, 0.61]	4.07
Random Effects				
σ^2	0.32			
$\tau_{00\ \text{TWIN\_ID}}$	1.01			
$\tau_{11\ \text{TWIN\_ID.ParticipantType}}$	1.05			
$\rho_{01\ \text{TWIN\_ID}}$	-0.82			
ICC	0.68			
$N_{\text{TWIN\_ID}}$	1003			
Observations	4012			
Marginal R^2 / Conditional R^2	0.073 / 0.706			

Table 5. Model B: Headline Veracity, Intervention and Participant Type on Sharing . * *p<0.05* ** *p<0.01* *** *p<0.001*.

Additionally, as a robustness check, a linear mixed-effects model with random intercepts and slopes nested was estimated to account for the repeated measures structure of the data. Specifically, participants' repeated responses to multiple headlines. Similar to the preregistered model with random intercepts (model A), this model revealed significant main effects of participant type (F(1,999) = 97.25, p < .001) as well as a significant effect of Intervention (F(3, 999) = 17.89, p < .001). Crucially, the results revealed a significant three-way interaction (F(3, 999) = 12.76, p < .001), further suggesting that the findings from the pre-registered model are robust to alternative model specifications. See Table 6 for the full regression results.

	Model C: Random Intercepts & Slopes (Nested)			
Predictors	*Estimates*	*SE*	*95%CI*	*t*
(Intercept)	2.82 ***	0.17	[2.48, 3.16]	16.28
Participant Type	0.32 ***	0.08	[0.16, 0.48]	3.84

Headline Veracity	-0.03	0.23	[0.48, 0.43]	-0.11
Active Control	0.13	0.09	[0.05, 0.31]	1.41
Treatment	-0.04	0.09	[0.22, 0.14]	-0.41
Importance Treatment	-0.32 ***	0.09	[0.50, 0.14]	-3.45
Participant Type — Headline Veracity	0.01	0.07	[0.12, 0.15]	0.20
Participant Type — Active Control	0.08	0.12	[0.15, 0.31]	0.70
Participant Type — Treatment	-0.31 **	0.12	[0.54, 0.08]	-2.63
Participant Type — Importance Treatment	-0.20	0.12	[0.43, 0.03]	-1.74
Headline Veracity — Active Control	-0.14	0.07	[0.28, 0.00]	-1.90
Headline Veracity — Treatment	0.05	0.07	[-0.09, 0.19]	0.64
Headline Veracity — Importance Treatment	0.03	0.07	[0.11, 0.17]	0.43
Participant Type — Headline Veracity — Active Control	0.20 *	0.10	[0.00, 0.39]	1.99
Participant Type — Headline Veracity — Treatment	0.57 ***	0.10	[0.38, 0.76]	5.78
Participant Type — Headline Veracity — Importance Treatment	0.41 ***	0.10	[0.22, 0.60]	4.18
Random Effects				
σ^2	0.97			
τ_{00} group_id:TWIN_ID	0.75			
τ_{00} TWIN_ID	0.19			
τ_{00} headline_id	0.21			
τ_{11} group_id:TWIN_ID.Headline_VeracityTrustworthy	0.36			
τ_{11} TWIN_ID.Headline_VeracityTrustworthy	0.03			
ρ_{01} group_id:TWIN_ID	-0.41			
ρ_{01} TWIN_ID	-0.52			
ICC	0.53			

$N_{\text{group\_id}}$	2006
$N_{\text{TWIN\_ID}}$	1003
$N_{\text{headline\_id}}$	16
Observations	32096
Marginal R^2 / Conditional R^2	0.037 / 0.546

Table 6. Model C: Headline Veracity, Intervention and Participant Type on Sharing. $^{*}p<0.05$ $^{**}p<0.01$ $^{***}p<0.001$.

As an additional test, Pennycook et al.’s (2021) analysis plan was next replicated for humans and their digital twins, separately. While the pre-registered analysis demonstrates sharing intentions for the full range of responses, Pennycook et al. (2021) rescale the continuous sharing intentions into a binary variable capturing responses at or above the midpoint at “likely to share” (= 1) and below the midpoint as “unlikely to share” (= 0), and filtered out participants who would not consider sharing these articles on social media. Because the human dataset revealed significant differences between the Control and Active Control conditions, these groups were not pooled to maintain comparability. This results in a slight deviation from the analysis in Pennycook et al. (2021). Using this framework, their exact analysis was replicated for humans and their digital twins separately, treating the rescaled likelihood to share variable as the DV with Headline Veracity condition, Intervention condition, and their interaction as predictors, applying robust standard errors. Figure 2 shows proportion of sharing across all conditions for humans (Panel A) and their digital twins (Panel B).

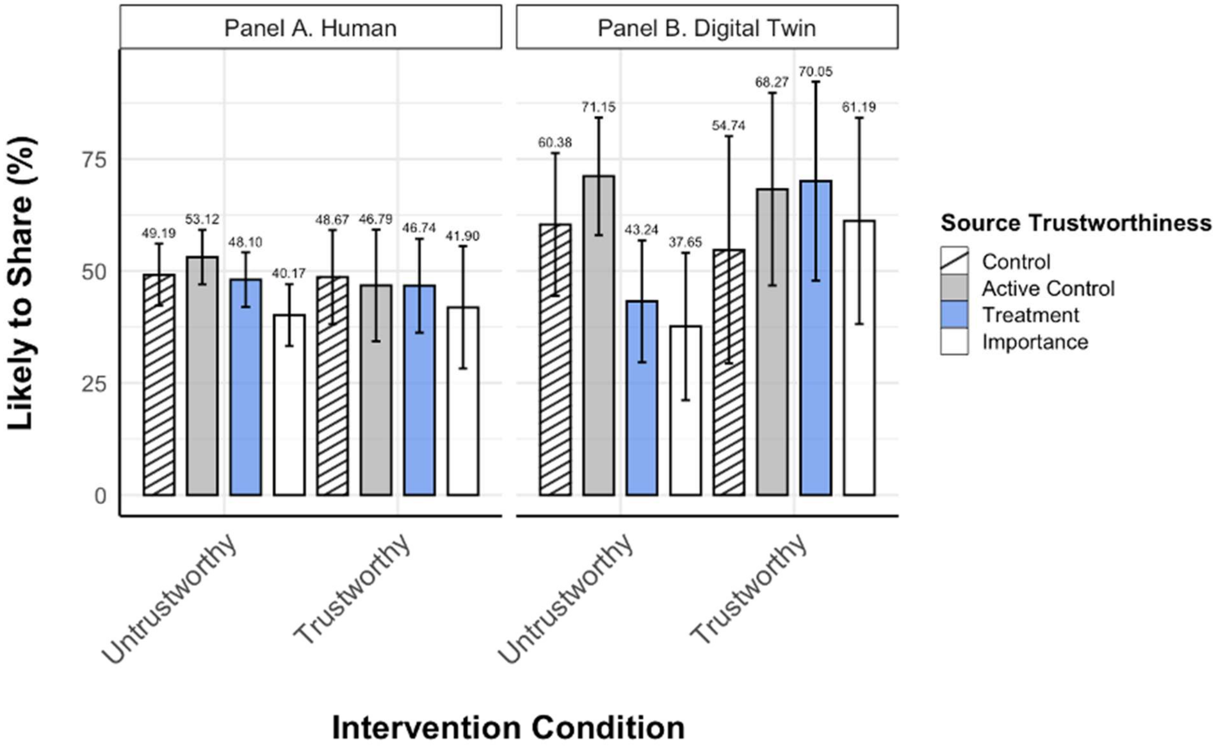
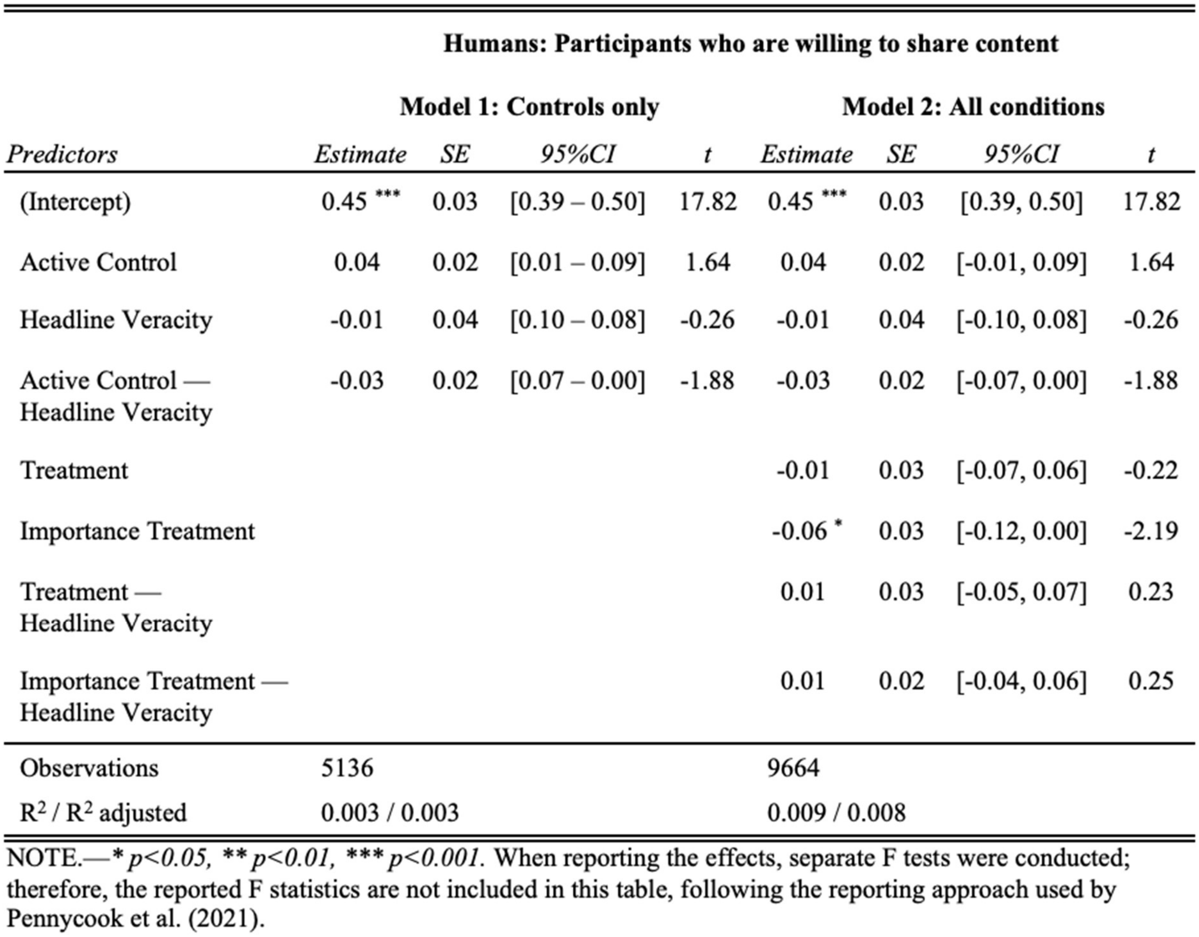


Figure 2. Proportion (%) Likely to Share by Headline Veracity and Intervention. NOTE: Shown here is the fraction of "likely' responses (responses above the midpoint of the six-point Likert scale) by Headline Veracity and Intervention condition; the full distributions of responses are shown below, in figure 3. As with Pennycook et al. (2021), these analyses focus only on participants who indicated that they would consider sharing this social media content. Analysis including all participants does not change the results. Error bars indicate 95% confidence intervals based on robust standard errors clustered on participant and headline.

Replicating the main findings for humans (model A), accuracy nudges do not seem to improve truth discernment when sharing entertaining news (figure 2, panel A). In contrast to Pennycook et al. (2021), sharing discernment was marginally in the active control compared to the control condition (Headline Veracity × Intervention: b = -0.03, 95% CI = [-0.07, 0.00], F(1, 5132) = 3.53, p = .06), and neither treatment significantly increased sharing discernment relative to the controls (Headline Veracity × Intervention: Treatment, b = 0.01, 95% CI = [-0.05, 0.06], F(1,9656) = 0.05, p = 0.82; Importance Treatment, b = 0.01, 95% CI = [-0.04, 0.05], F(1,9656) = 0.06, p = 0.80). See Table 7 for the full regression results.

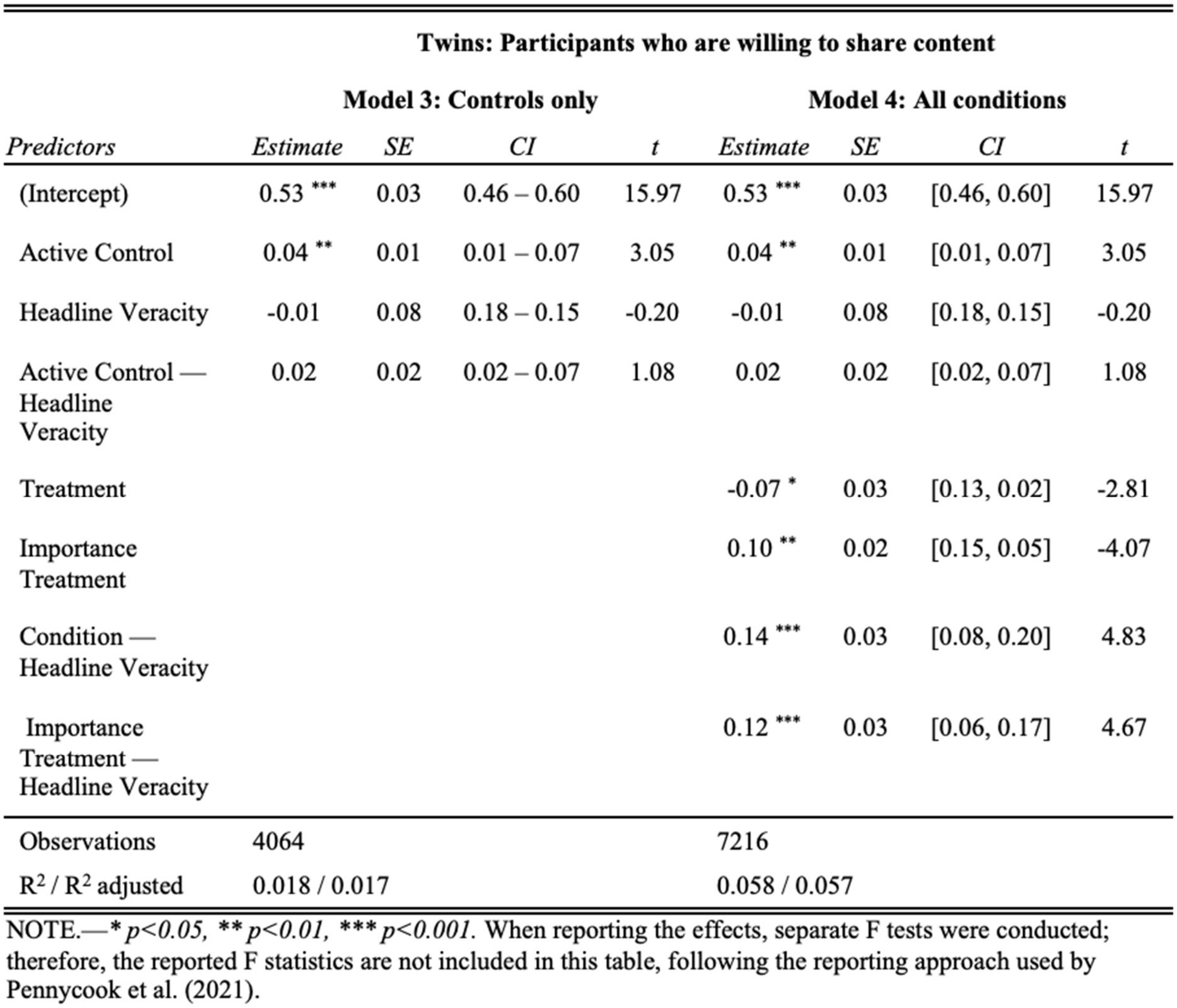


	Humans: Participants who are willing to share content							
	Model 1: Controls only				**Model 2: All conditions**			
Predictors	*Estimate*	*SE*	*95%CI*	*t*	*Estimate*	*SE*	*95%CI*	*t*
(Intercept)	0.45 ***	0.03	[0.39 – 0.50]	17.82	0.45 ***	0.03	[0.39, 0.50]	17.82
Active Control	0.04	0.02	[0.01 – 0.09]	1.64	0.04	0.02	[-0.01, 0.09]	1.64
Headline Veracity	-0.01	0.04	[0.10 – 0.08]	-0.26	-0.01	0.04	[-0.10, 0.08]	-0.26
Active Control — Headline Veracity	-0.03	0.02	[0.07 – 0.00]	-1.88	-0.03	0.02	[-0.07, 0.00]	-1.88
Treatment					-0.01	0.03	[-0.07, 0.06]	-0.22
Importance Treatment					-0.06 *	0.03	[-0.12, 0.00]	-2.19
Treatment — Headline Veracity					0.01	0.03	[-0.05, 0.07]	0.23
Importance Treatment — Headline Veracity					0.01	0.02	[-0.04, 0.06]	0.25
Observations	5136				9664			
R^2 / R^2 adjusted	0.003 / 0.003				0.009 / 0.008			

NOTE.—* *p<0.05,* ** *p<0.01,* *** *p<0.001.* When reporting the effects, separate F tests were conducted; therefore, the reported F statistics are not included in this table, following the reporting approach used by Pennycook et al. (2021).

Table 7. Headline Veracity, Intervention and Participant Type on Sharing Intentions.

Again, replicating the main findings for digital twins (model A), accuracy nudges improve truth discernment when sharing entertaining news (figure 2, panel B). As in Pennycook et al. (2021), there were no significant differences in sharing intentions between the control and active control conditions (Headline Veracity × Intervention b = 0.02, 95% CI= [-0.02, 0.06], F(1, 4046) = 1.17, p = .28) and both treatments significantly increased sharing discernment relative to the controls (Headline Veracity × Intervention: Treatment, b = 0.14, 95%CI= [0.08, 0.20], F(1, 7208) = 23.30, p = <.001; Importance Treatment, b = 0.12, 95% CI = [0.07, 0.17], F(1, 7208) = 21.80, p = <.001). See Table 8 for the full regression results.



	Twins: Participants who are willing to share content							
	Model 3: Controls only				Model 4: All conditions			
Predictors	*Estimate*	*SE*	*CI*	*t*	*Estimate*	*SE*	*CI*	*t*
(Intercept)	0.53 ***	0.03	0.46 – 0.60	15.97	0.53 ***	0.03	[0.46, 0.60]	15.97
Active Control	0.04 **	0.01	0.01 – 0.07	3.05	0.04 **	0.01	[0.01, 0.07]	3.05
Headline Veracity	-0.01	0.08	0.18 – 0.15	-0.20	-0.01	0.08	[0.18, 0.15]	-0.20
Active Control — Headline Veracity	0.02	0.02	0.02 – 0.07	1.08	0.02	0.02	[0.02, 0.07]	1.08
Treatment					-0.07 *	0.03	[0.13, 0.02]	-2.81
Importance Treatment					0.10 **	0.02	[0.15, 0.05]	-4.07
Condition — Headline Veracity					0.14 ***	0.03	[0.08, 0.20]	4.83
Importance Treatment — Headline Veracity					0.12 ***	0.03	[0.06, 0.17]	4.67
Observations	4064				7216			
R^2 / R^2 adjusted	0.018 / 0.017				0.058 / 0.057			

NOTE.—* *p<0.05,* ** *p<0.01,* *** *p<0.001.* When reporting the effects, separate F tests were conducted; therefore, the reported F statistics are not included in this table, following the reporting approach used by Pennycook et al. (2021).

Table 8. Headline Veracity, Intervention and Participant Type on Sharing Intentions.

Finally, Figure 3 displays the full distribution of individual sharing likelihood ratings for each participant type across all intervention conditions and headline veracity levels. These distributions reveal that humans were far more likely to report that they would not share headlines, with many responses clustering at the low end of the scale. In contrast, digital twins tended to give more moderate responses, with sharing likelihood ratings concentrated around the midpoint and a greater proportion of “likely to share” responses overall. This pattern suggests that, unlike humans—who default to inaction—digital twins are generally more inclined to share content.

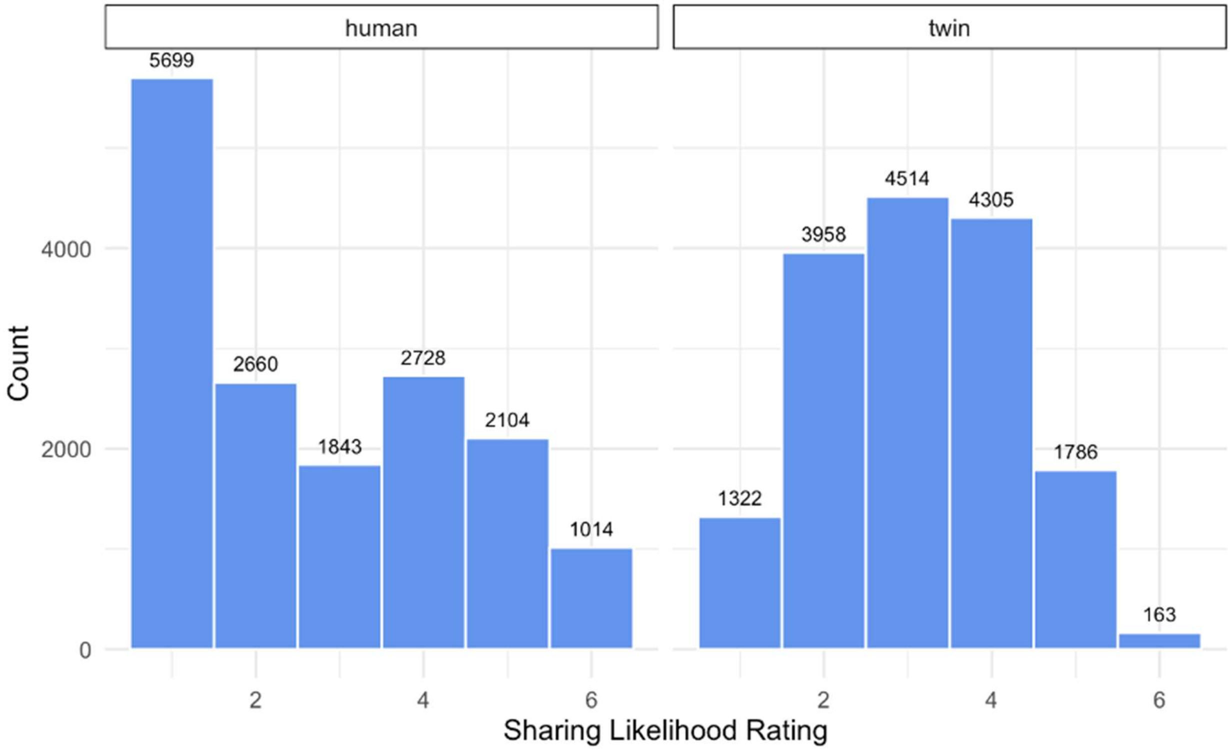


Figure 3. Full Distribution of Individual Sharing Likelihood Ratings.

These additional analyses strengthen confidence in the findings and provide a straightforward comparison for humans and their digital twins, separately, with the reported findings from Pennycook et al. (2021).

Discussion

These findings extend prior research studying the effectiveness of accuracy nudges in combating the spread of misinformation (Pennycook et al. 2021). Critically, accuracy nudges may have limited effectiveness for stopping the spread of false entertainment news. While prior research has shown that nudges can redirect attention toward accuracy and reduce the spread of false information in political contexts (Pennycook et al. 2021), this finding does not seem to generalize to more playful, emotionally engaging content. Specifically, even when primed to consider accuracy or reminded of its importance, humans showed no statistically reliable reduction in their willingness to share false versus true entertainment headlines. Even more, when first prompted to rate the entertainingness of unrelated headlines (i.e., Active Control condition), participants were marginally more likely to share false versus true entertainment headlines, indicating worse discernment. Interestingly, in contrast to humans, digital twins responded as the prior work would predict, demonstrating robust improvements in truth discernment in both the Treatment and Importance Treatment conditions.

One possible reason for this divergence is that digital twins, unlike humans, do not experience the hedonic appeal of entertaining content. For humans, entertainment may override epistemic goals—especially when sharing functions as a form of social play or bonding (working paper, Lane and Brucks 2025). In contrast, digital twins may more rigidly apply the prioritization embedded in their programming or training data, defaulting to accuracy in the absence of competing motivations. In this way, digital twins act as idealized rational agents—models of how people ought to behave if their sole goal was to share accurate information. But humans operate in a messier reality, where attention is easily hijacked by humor, surprise, or emotional resonance. This gap underscores a broader challenge in misinformation interventions: even the most well-intentioned nudges may fall short when the content itself draws attention away from accuracy.

From a practical standpoint, the results highlight the limits of one-size-fits-all interventions. If humans are less responsive to accuracy nudges in hedonic contexts, future interventions may need to do more than redirect attention—they may need to reshape the perceived function of sharing itself. For example, interventions could incentivize or socially reward accurate sharing (working paper, Lane and Brucks 2025), rebalancing the tradeoff between entertainment and truth. More ambitiously, platforms might redesign their interfaces to elevate epistemic cues without dampening the joy of social engagement. At a theoretical level, these findings invite deeper reflection on how models of decision-making should incorporate context-dependent goals, and how digital twins can help surface not just what people do, but why they do it.

References

Berger, Jonah (2011), "Arousal Increases Social Transmission of Information," Psychological Science, 22 (7), 891-93.

Berger, Jonah, and Katherine L. Milkman (2012), "What makes online content viral?," Journal of Marketing Research, 49 (2), 192-205.

Berger, Jonah, and Eric M. Schwartz (2011), "What Drives Immediate and Ongoing Word of Mouth?," Journal of Marketing Research, 48 (5), 869-80.

Harcup, Tony, and Deirdre O'Neill (2017), "What is News? News values revisited (again)," Journalism Studies, 18 (12), 1470-88.

Lane, Kristen, and Merrie Brucks (2025, working paper), "A Marketing Perspective on Misinformation Sharing: How the Target Audience Shapes Consumers' Focus on Accuracy vs. Entertainment".

Pennycook, Gordon, and David G. Rand (2019), "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," Cognition, 188, 39-50.

Pennycook, Gordon, and David G. Rand (2021), "The Psychology of Fake News," Trends in Cognitive Sciences, 25 (5), 388-402.

Reuters (2023), Digital News Report 2023, University of Oxford. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

Widjaya, Regina, Samuel Bestvater, and Aaron Smith (2024), “Who U.S. Adults Follow on TikTok,” Pew Research Center, October 8, Retrieved from https://www.pewresearch.org/data-labs/2024/10/08/who-u-s-adults-follow-on-tiktok/

Digital Certificates for Luxury Consumption

Main Questions/Hypotheses

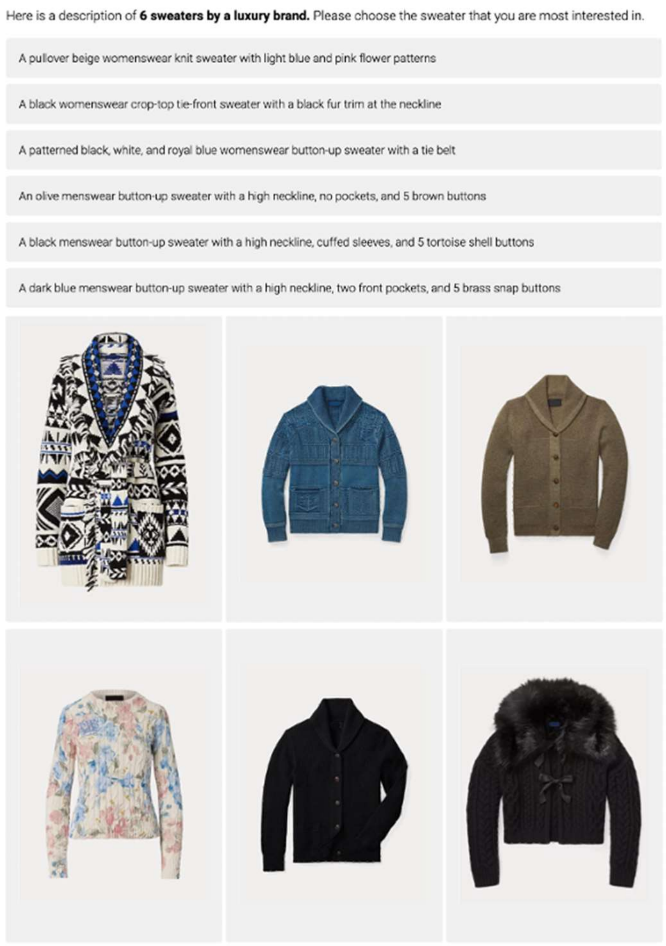
Luxury brands face the perpetual problem of balancing exclusivity with growth (Keller 2017). Diffusion of products into the marketplace inevitably erodes exclusivity and negatively impacts product and brand value (Bellezza and Keinan 2014). Recent work has demonstrated that a new technology—digital certification—reinforces value when luxury products become diffused (Park, Lane, and Bellezza 2022). In this study, we aim to replicate this pattern with human participants, and test whether their digital twins respond similarly to digital certification.

Specifically, we test whether attaching a digital passport to a luxury product that has become diffused increases its perceived value (i.e., status perceptions) and monetary value (i.e., price perceptions). Digital passports are a novel Blockchain-based certification system inspired by a recent European Union regulation to create more transparent and sustainable supply chains for physical products (EU Parliament Approves Supply Chain Law 2024). For example, the Aura Blockchain Collective (a non-profit collective of major luxury firms) provides digital passports that allow consumers to follow their purchases through the entire product journey, from the origin of the raw materials to recycling ("Solutions - Aura Blockchain Consortium" n.d.).

Methods

We opened the study to 600 human participants on Prolific (46.50% female, 53.50% male; Mage = 49.86, SD = 14.32). The study was subsequently run on their 600 digital twins, and the final sample size is 1200 for our analysis. Participants and their corresponding digital twins were randomly assigned to the same condition in a two-cell (Digital Certification: Digital Passport vs. Control) between-subjects design.

All participants were asked to select one of six luxury sweaters: "Here is a selection of 6 sweaters by a luxury brand. Please choose the sweater that you are most interested in". Participants were given a description of each sweater, instead of an image, to enable comparison between humans and their digital counterparts, who cannot process visual images (Figure 1, top). The description of each sweater was based on real sweaters designed by Ralph Lauren (M_{price} = $531.90; Figure 1, bottom), but to avoid potential confounds, the brand was not mentioned. These six target sweaters were chosen from a set of 20 after an extensive pretest to ensure the products were equally liked, perceived as high-end, suitable for the scenario (i.e., as an outfit for going out), and conspicuous. Additional exploratory analysis controlled for participants' selected sweater to account for differences across product stimuli (e.g., sweater color, style) that could influence value perceptions.



Here is a description of **6 sweaters by a luxury brand.** Please choose the sweater that you are most interested in.

A pullover beige womenswear knit sweater with light blue and pink flower patterns

A black womenswear crop-top tie-front sweater with a black fur trim at the neckline

A patterned black, white, and royal blue womenswear button-up sweater with a tie belt

An olive menswear button-up sweater with a high neckline, no pockets, and 5 brown buttons

A black menswear button-up sweater with a high neckline, cuffed sleeves, and 5 tortoise shell buttons

A dark blue menswear button-up sweater with a high neckline, two front pockets, and 5 brass snap buttons

Figure 1. Stimuli for Digital Certification Study.

After choosing a sweater, participants in the digital passport condition read as follows:

> We will ask you a few questions about the luxury sweater you selected. This luxury sweater is linked to a permanent digital record available on a public website, showcasing the current owner. This allows anyone interested to verify the ownership of your specific luxury sweater through this accessible record. Essentially, the digital record serves as unequivocal evidence of your luxury sweater's ownership.

Participants in the control condition alternatively read: "We will ask you a few questions about the luxury sweater you selected." All conditions next featured a scenario in which the luxury product has become diffused. Specifically, all participants read, "Imagine you decided to purchase this luxury sweater and style it for your night out, but during your night out you noticed many other people wearing the same luxury sweater." Participants in the digital passport condition also read, "Nevertheless, you know there is a digital record associated with the particular luxury sweater you are wearing, reassuring your ownership." Finally, all participants completed a writing task to reinforce the manipulation: "Please imagine your night out in detail. Briefly describe how you look and how you would feel wearing this sweater during your night out."

Then, participants responded to the dependent variables by reporting their perceptions of product value ("How expensive do you think the sweater is? (Please enter number of $)") and status perceptions ("Please answer the following questions. To what extent do you think the sweater is…"; luxurious, high status, prestigious; 1 = Not at all; 7 = To a great extent; Ward and Dahl 2014). These three items were averaged for analysis (α = .94). We log-normalized the price perception measure to deal with skewness (μ = 4.06> 2; i.e., right skewed; Curran, West, and Finch 1996). Finally, participants reported the extent to which they are familiar with digital passports ("How familiar are you with digital passports?" (Very Unfamiliar, 7-Very Familiar), the perceived difficulty of the task to account for different prompt lengths per condition ("How difficult was it to imagine the sweater scenario?" 1 = Extremely Easy, 7 = Extremely Difficult), and whether they owned real luxury products (0 = No, 1 = Yes).

Results - Pre-registered Analyses

We estimated a linear mixed-effects regression with participant as a random effect, log transformed price perceptions as the dependent variable, and participant type (human = 0, digital twin = 1), digital certification (0 = control, 1 = digital passport) and their interaction as predictors. We found a significant effect of digital certification (b = .25, SE = .05; t(1190.45) = 4.94, p < .001) along with a significant effect of participant type (b = .83, SE = .05; t(597.99) = 16.91, p < .001) and a significant interaction (b = -.21, SE = .07; t(596.99) = -3.08, p < .01; Figure 2), indicating that, at an aggregate level, the digital twins rated price perceptions differently than the humans. Pairwise comparisons confirm that for humans, price perceptions were significantly higher when their purchase was connected to a digital certification (M = 5.26, SE = .04) compared to no digital certification (M = 5.01, SE = .04; t(1190) = -4.94, p < .001). However, for the digital twins, price perceptions were not significantly different when their purchase was connected to a digital certification (M = 5.88, SE = .04) compared to no digital certification (M = 5.84, SE = .04; t(1190) = -.73, p = .463).

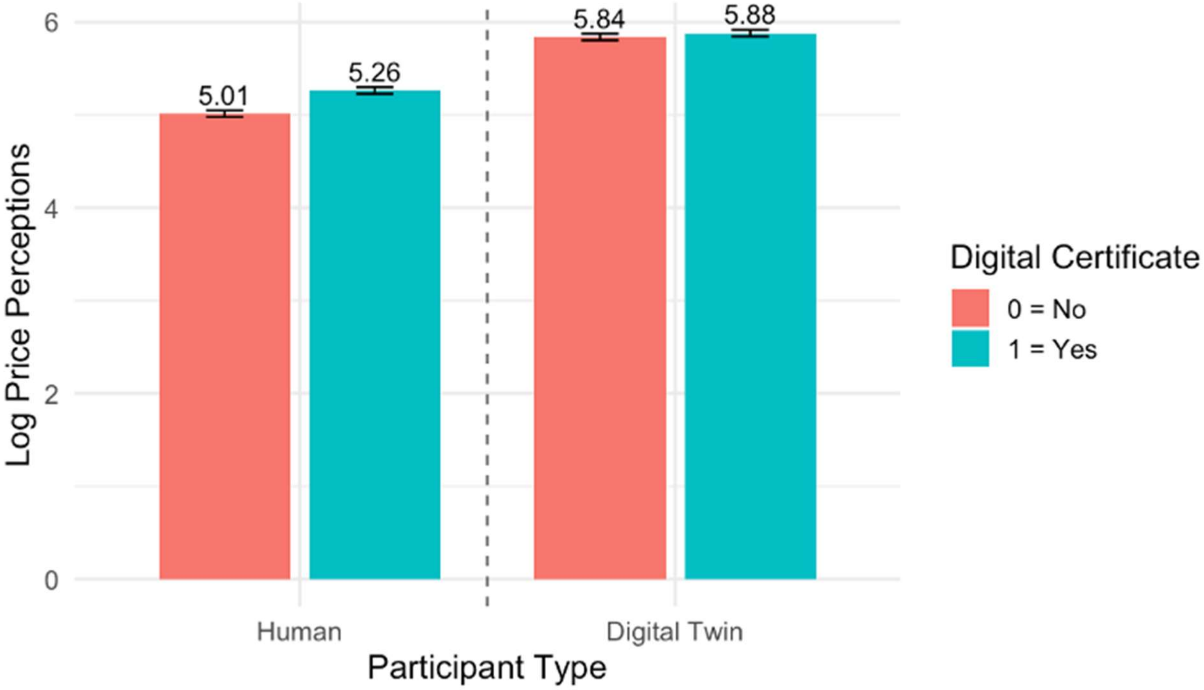


Figure 2. Price Perceptions by Digital Certification for Humans vs. Twins. NOTE: Error bars indicate ±1 standard error of the estimated marginal means.

We next estimated a linear mixed-effects regression with participant as a random effect, status perceptions as the dependent variable, and participant type (human = 0, digital twin = 1), digital certification (0 = control, 1 = digital passport) and their interaction as predictors. We found a marginally significant effect of digital certification (b = .14, SE = .099; t(1175.69) = 1.44, p = .151) along with a significant effect of participant type (b = .19, SE = .09; t(597.99) = 2.06, p = .04) and a significant interaction (b = -.55, SE = .13; t(597.99) = -4.24, p < .001; Figure 3), indicating that, at an aggregate level, the digital twins rated status perceptions differently than the humans. Pairwise comparisons confirm that for humans, status perceptions were marginally higher when their purchase was connected to a digital certification (M = 4.81, SE = .07) compared to no digital certification (M = 4.66, SE = .07, t(1176) = -1.44, p = .15). However, for the digital twins, status perceptions were significantly lower when their purchase was connected to a digital certification (M = 4.44, SE = .07) compared to no digital certification (M = 4.85, SE = .07, t(1176) = 4.16, p < .001).

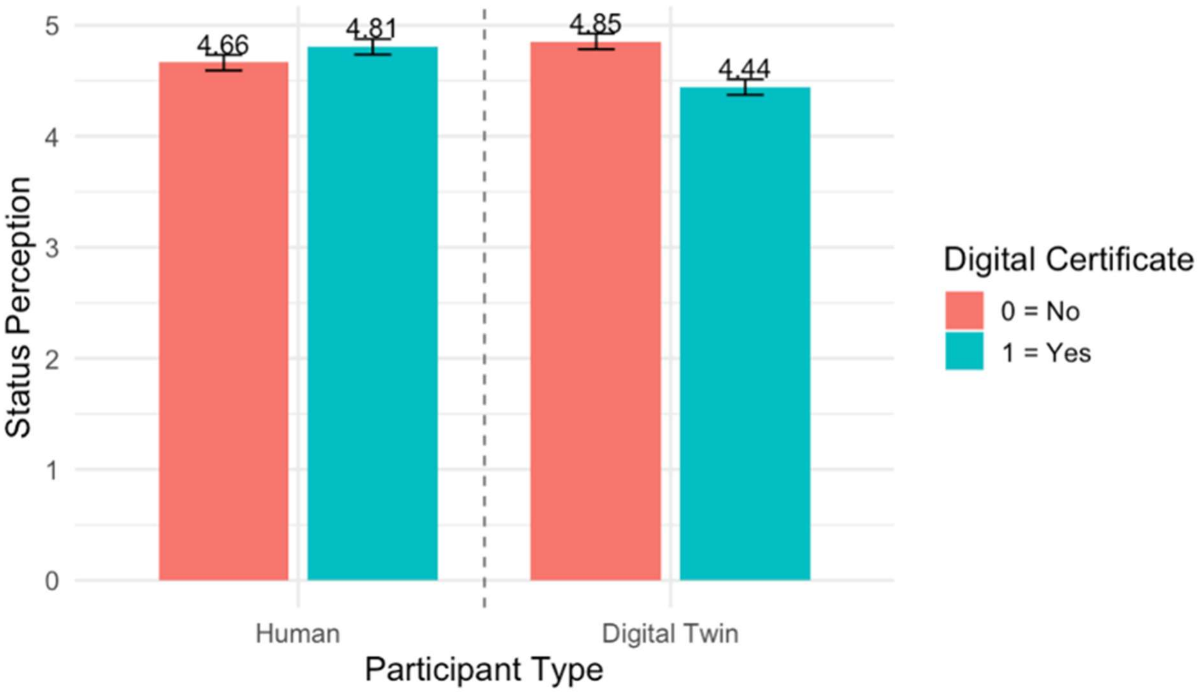


Figure 3. Status Perceptions by Digital Certification for Humans vs. Twins. NOTE: Error bars indicate ±1 standard error of the estimated marginal means.

Results - Additional Analyses (Non-Preregistered)

Although not part of our pre-registered plan, we ran two additional linear mixed-effects regressions on log-transformed price perceptions and status perceptions controlling for the specific product selected by the participant, number of words generated in the imagination prompt, participants' familiarity with digital passports, difficulty of imagining the scenario, and personal ownership of luxury products. Before running these regressions, we examined differences between humans and their digital twins on each of these variables.

We first tested whether humans and their twins differed in the sweaters they selected. The results indicate weak but statistically significant agreement between twins and humans (κ = .166, p < .001). A chi-squared test of independence also revealed a significant and positive association between human and twin sweater choices (χ^2(1, N = 1200) = 297.4, p < .001). Breaking down choice by the gender association of the sweaters, digital twins tend to mirror their human counterparts: when humans selected one of the three feminine sweaters, their twins also selected a feminine sweater 86.31% of the time (χ^2(1, N = 241) = 125.63, p < .001); when humans selected one of the three masculine sweaters, their twins chose a masculine sweater 85.24% of the time (χ^2(1, N = 359) = 176.89, p < .001).

We next checked whether all participants wrote roughly an equal number of words (log-corrected) between conditions. While the number of words generated did not differ between conditions for humans ($M_{\text{Digital Passport}}$ = 3.54 vs. M_{Control} = 3.47; F(1, 598) = 1.70, p = 0.20; d = .11), the digital twins generated more words in the digital passport condition (M = 4.42) compared to the control (M = 4.36; F(1, 598) = 66.92, p <.001, d = .67). Additionally, the digital twins generated more text (M = 4.42) compared to their human counterparts (M = 3.50; F(1, 599) = 1305.28, p < .001, d = 1.47).

We then tested for perceived familiarity with digital passports. Across humans and twins, there were no differences in perceived familiarity among participants in the digital passport condition (M = 2.40) compared to the control condition (M = 2.36; F(1, 1198) = 0.20, p = 0.65, d = 0.03). Human participants, however, reported being more familiar with digital passports (M = 2.68) compared to twins (M = 2.08; F(1, 599) = 74.54, p < .001, d = 0.35).

We then tested for differences in difficulty of imagining the scenario between conditions. Human participants in the digital passport condition found the imagination task more difficult (M = 2.68) than in the control condition (M = 2.03; F(1, 598) = 32.70, p < .001, d = 0.47). Digital twins also found the imagination task more difficult in the digital passport condition (M = 2.59) than in the control condition (M = 2.31; F(1, 598) = 22.32, p < .001, d = 0.39). There were no differences in difficulty of imagining the tasks between humans (M = 2.35) and twins (M = 2.45; F(1, 1198) = 2.26, p = .13, d = 0.09).

We finally tested for differences in reported ownership of luxury products. Across humans and twins, there were no differences in luxury product ownership among participants in the digital passport condition (40.4%) compared to the control condition (39.3%; $\chi^2(1)$ = 0.12, p = .73). Human participants, however, reported higher luxury ownership (53.8%) compared to their digital twins (25.8%; $\chi^2(1)$ = 96.97, p < .001).

To test whether our findings persist controlling for these differences, we ran a linear mixed effects regression with participant as a random effect, status perceptions as the dependent variable, and participant type (human = 0, digital twin = 1), digital certification (0 = control, 1 = digital passport) and their interaction as predictors, controlling for the product participants selected, as well as familiarity with digital passports, difficulty of imagination, and ownership of luxury products on log-normalized perceived price as the dependent variable. We find a significant effect of familiarity (b = .05, SE = .01; t(1190.90) = 3.48, p < .001) and luxury product ownership (b = .20, SE = .04; t(1170.09) = 5.10, p < .001), but a non-significant effect of product choice (b = .02, SE = .01; t(1190.85) = 1.48, p = .14) and difficulty of imagination (b = -.02, SE = .02; t(1190.82) = -1.51, p = .13). Nevertheless, the analysis revealed a significant effect of digital certification (b = .28, SE = .05; t(1189.85) = 5.53, p < .001) and a significant effect of participant type (b = .96, SE = .05; t(726.87) = 17.88, p < .001), qualified by a significant interaction between digital certification and participant type (b = -.25, SE = .07; t(603.74) = -3.66, p < .001).

Next, we ran a linear mixed effects regression with participant as a random effect, status perceptions as the dependent variable, and participant type (human = 0, digital twin = 1), digital certification (0 = control, 1 = digital passport) and their interaction as predictors, controlling for the product participants selected, as well as familiarity with digital passports, difficulty of imagination, and ownership of luxury products on status perceptions as the dependent variable. We again find a significant effect of familiarity (b = .07, SE = .03; t(1185.69) = 2.43, p = .02) and luxury product ownership (b = .46, SE = .08; t(1180.77) = 6.09, p < .001), but a non-significant effect of product choice (b = -.03, SE = .02; t(1,185.33) = -1.43, p = .15) and difficulty of imagination (b = -.02, SE = .03; t(1191.82) = -.70, p = .48). Nevertheless, the analysis revealed a marginal effect of digital certification (b = .17, SE = .10; t(1183.86) = 1.72, p = .08) and a significant effect of participant type (b = .34, SE = .10; t(722.27) = 3.30, p < .01), qualified a significant interaction between digital certification and participant type (b = -.59, SE = .13; t(598.73) = -4.51, p < .001).

These additional analyses strengthen our confidence that the effects are robust across individual product choice, familiarity with the digital certification context (i.e., digital passports), ease of completing the task, and luxury product ownership.

Discussion

These findings demonstrate that providing digital certification via a digital passport increases the value perception of a diffused luxury product for humans. Digital twins' value perceptions, however, are not impacted by the inclusion of a digital passport. The findings have important implications for our understanding of when digital twins will not act as their human counterparts.

In this study, the divergence may stem from the psychological reassurance of product authenticity that certification provides to human consumers (Park, Lane, and Bellezza 2022), which digital twins cannot fully internalize. This highlights a potential blind spot in current AI models: a limited ability to model symbolic consumption, ownership signaling, or status restoration processes. More broadly, these findings call for caution when relying on digital twins to simulate human judgments in domains involving identity and social perception—especially in luxury or prestige-based contexts. Future work should explore how these divergences arise and whether digital twins can be improved to better simulate symbolic inference processes, or whether some aspects of human value perception are fundamentally non-transferrable.

References

Bellezza, Silvia and Anat Keinan (2014), "Brand Tourists: How Non–Core Users Enhance the Brand Image by Eliciting Pride," Journal of Consumer Research, 41 (2), 397–417.

Curran, Patrick J., Stephen G. West, and John F. Finch (1996), "The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis," Psychological Methods, 1 (1), 16–29.

"EU Parliament Approves Supply Chain Law" (2024), Brussels.

Geylan, Zeynep (2022), "Total NFT market cap grows 11,664% in 2 years," Cryptoslate, (accessed January 2, 2023), [available at https://cryptoslate.com/total-nft-market-cap-grows-11664-in-2-years/].

Keller, Kevin Lane (2017), “Managing the Growth Tradeoff: Challenges and Opportunities in Luxury Branding,” Advances in Luxury Brand Management, 179–98.

Park, Eric, Kristen Lane, and Silvia Bellezza (2022, October). “NFT for Conspicuous Consumption,” in NA - Advances in Consumer Research Volume 50, eds. Haipeng (Allan) Chen, Giana Eckhardt, and Rebecca Hamilton, Duluth, MN : Association for Consumer Research, Denver, CO.

“Solutions - Aura Blockchain Consortium” (n.d.), (accessed February 18, 2025), [available at https://auraconsortium.com/solutions].

Ward, Morgan K. and Darren W. Dahl (2014), “Should the Devil Sell Prada? Retail Rejection Increases Aspiring Consumers’ Desire for the Brand,” Journal of Consumer Research, 41 (3), 590–609.

Affective Primes

Main Questions/Hypotheses

Research in the social sciences often manipulates emotional states using affective priming, where participants are asked to reflect on a recent event. In this study, we test whether affective priming works with digital twins. Specifically, we ask:

1. Do affective priming manipulations induce "states" in digital twins (e.g., does writing about gratitude or lack of control momentarily influence digital twins' responses?)
2. Is the influence of affective priming similar for digital twins and their human counterparts?
3. Is the influence of affective priming on digital twins dependent on the valence of affective prime (i.e., between positive primes like gratitude or negative primes like lack of control) or the proximal nature of the dependent measures (i.e., does it "spill over" to other, related dimensions or only influence proximal dimensions)?

Methods

One thousand human participants on Prolific (and their twins) participated in the study. They were randomly assigned the same condition in a 2 (affective domain: gratitude vs. lack of control) x 2 (prompt: prime vs. baseline) between-subjects design.

Participants first responded to a condition-specific prompt for two minutes. For the gratitude domain, participants received either a gratitude prime ("Please recall carefully and in detail a specific experience in the past when you felt sincerely grateful for someone's kindness or help") or baseline prompt ("Please recall carefully and in detail the sequence of your morning routine, such as brushing teeth or changing clothes"), using procedures from Emmons and McCullough (2003), DeSteno (2014), and Oguni and Ishii (2024). In the lack of control domain, participants received either the lack of control prime ("Please recall a particular incident in which something happened and you did not have control over the situation. Please describe the situation in which you felt a lack of control – what happened, how you felt, etc.") or a baseline prompt ("Please recall a particular incident in which something happened and you were in control of the situation. Please describe the situation in which you felt in control – what happened, how you felt, etc."), using procedures from Bukowski et al. (2024), Chen, Lee, and Yap (2017), Lembregts and Pandelaere (2019), and Whitson and Galinsky (2008).

After the writing task, participants completed a manipulation check, followed by proximal dependent measures (closely related to the target affective state) and distal dependent measures (more distantly related to the affective state). For the gratitude domain, the manipulation check assessed gratitude-related emotions (grateful, thankful, appreciative) on a 7-point scale from "not at all" to "extremely." The proximal measure used the Elevation scale (Schnall, Roper, and Fessler, 2010; Walsh et al., 2022), which asked participants to rate their current feelings on six items such as "Feeling optimistic about humanity" using the same 7-point scale. The distal measure employed an adapted version of the Empathic Concern scale (Oliveira et al., 2021), measuring agreement with two statements, "I am very concerned about those most vulnerable to the effects of tariffs" and "I feel compassion for those most affected by rising prices or job losses due to tariffs," on a 7-point scale from "completely disagree" to "completely agree."

For the lack of control domain, the manipulation check used items from Greenaway et al. (2015), asking participants to rate their agreement with three statements ("I feel in control of my life", "I am free to live my life how I wish", "My experiences in life are due to my own actions", on a 7-point scale from "not at all" to "extremely." These items were reverse coded for the analysis, but we present the raw means before reverse coding in the graph. The proximal measure was the Fatigue scale (Sedek & Kofta, 1990; Bulowski et al., 2024), consisting of nine items such as "I had a hard time thinking about the event" rated on the same 7-point scale. The distal measure employed the Desire for Predictability scale (Webster and Kruglanski, 1994; Lembregts and Pandelaere, 2019), containing eight items like "would not like to go into a situation without knowing what I can expect from it" rated on a 7-point scale from "completely disagree" to "completely agree."

Results - Pre-registered Analyses

First, the scales demonstrated strong internal consistency for both humans and digital twins (Digital Twins: gratitude manipulation check (α = .998), elevation scale (α = .964), empathic concern measure (α = .992), lack of control manipulation check (α = .951), fatigue scale (α = .897), desire for predictability measure (α = .931); Humans: gratitude manipulation check (α = .979), elevation scale (α = .917) empathic concern measure (α = .909), lack of control manipulation check (α = .830), fatigue scale (α = .766), and desire for predictability measure (α = .866)). As a result, we created a composite variable for each measure by averaging their respective items.

Following our pre-registered analysis plan, we then averaged the outcome measures in the gratitude domain (elevation, empathic concern, r = .38) to create an aggregate downstream dependent variable for gratitude and averaged the outcome measures in the lack of control domain (fatigue and desire for predictability, r = .09) to create an aggregate downstream dependent variable for lack of control. We also examine effects by individual measure.

Manipulation Check

To examine the effect of affective priming on humans versus digital twins for both gratitude and lack of control, we ran a linear mixed-effects regression using the lmerTest package in R (Kuznetsova, Brockhoff, and Christensen 2017) with the manipulation check measure as the dependent variable, affective domain (gratitude vs. feeling in control), prompt (prime vs. baseline), twin (human vs. digital) and their interactions as independent variables, and twin identifier as a random intercept. We do not observe a significant three-way interaction (b = -0.17, SE = 0.17, t(996) = -1.01, p = .314), but several interesting two-way interactions emerge, which we detail below.

First, we find that affective priming manipulations can successfully induce "states" in both digital twins and humans. The prompt (prime vs. baseline) had a significant effect on manipulation checks for digital twins in both the gratitude domain ($M_{baseline}$ = 4.34, SD =

1.17, M_{prime} = 5.67, SD = .71, b = 1.33, SE = 0.10, t(1760) = 12.93, p < .001) and lack of control domain ($M_{baseline}$ = 2.79, SD = .86, M_{prime} = 3.79, SD = .95, b = 1.00, SE = 0.10, t(1760) = 9.55, p < .001), as well as for humans in the gratitude domain ($M_{baseline}$ = 5.19, SD = 1.56, M_{prime} = 6.00 SD = 1.33, b = 0.81, SE = 0.10, t(1760) = 7.91, p < .001) and lack of control domain ($M_{baseline}$ = 2.81, SD = 1.16, M_{prime} = 3.12, SD = 1.29, b = 0.32, SE = 0.10, t(1760) = 3.02, p = .003).

Second, the effect of priming was significantly stronger for digital twins than humans in both affective domains. Digital twins showed greater responsiveness to the prompt (prime vs. baseline) in both the gratitude domain (interaction: b = -0.52, SE = 0.12, t(996) = -4.45, p < .001) and lack of control domain (interaction: b = -0.68, SE = 0.12, t(996) = -5.79, p < .001) compared to humans (see Figure 1 below).

Finally, the prompt (prime vs. baseline) produced stronger effects in the gratitude domain than the lack of control domain for both digital twins (b = 0.33, SE = 0.15, t(1760) = 2.26, p = .024) and humans (b = 0.50, SE = 0.15, t(1760) = 3.39, p < .001).

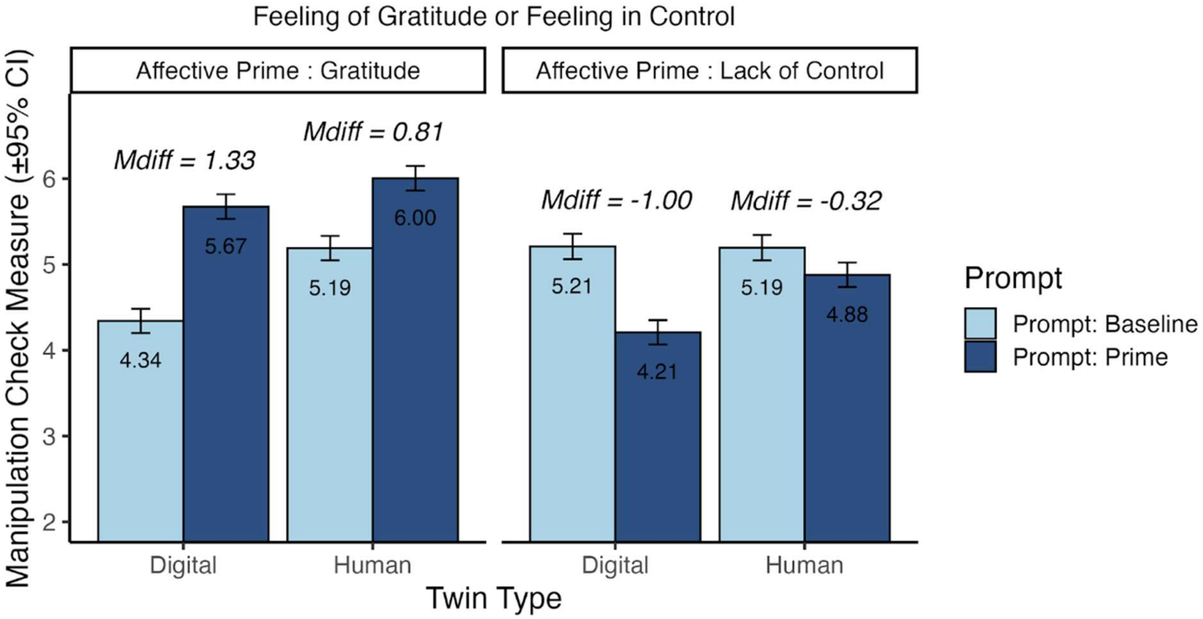


Figure 1. Effect of Affective Prime on Manipulation Check.

Dependent Measures

We next examined how affective priming in digital twins affects downstream dependent measures—that is, do these “affective states” spill over onto other measures? To test this, we ran a linear mixed-effects regression with the aggregate dependent variable measure as the dependent variable, affective domain (gratitude vs. lack of control), prompt (prime vs. baseline) and twin (human vs. digital) and their interaction as independent variables, and twin identifier as a random intercept. We do not observe a significant three-way interaction (b = -0.13, SE = 0.11, t(996) = -1.17, p = .242), but two significant two-way interactions emerge, which we detail below.

First, we find that affective priming manipulations spill over onto downstream dependent measures for both digital twins and humans. The prompt (prime vs. baseline) had a significant effect on downstream measures for digital twins in both the gratitude domain ($M_{baseline}$ = 4.39, SD = 1.11, M_{prime} = 4.90, SD = 1.05, b = 0.50, SE = 0.08, t(1538) = 6.01, p < .001) and lack of control domain ($M_{baseline}$= 3.92, SD = .60, M_{prime} = 4.38, SD = .68, b = 0.46, SE = 0.09, t(1538) = 5.45, p < .001), as well as for humans in the gratitude domain ($M_{baseline}$ = 4.82, SD = 1.15, M_{prime} = 5.21, SD = 1.23, b = 0.81, SE = 0.10, t(1760) = 7.91, p < .001) and lack of control domain ($M_{baseline}$ = 3.83, SD = .76, M_{prime} = 4.05, SD = .69, b = 0.32, SE = 0.10, t(1760) = 3.02, p = .003). See Figure 2.

Second, the effect of priming on downstream dependent measures was significantly stronger for digital twins than humans for lack of control priming (b = -0.25, SE = 0.08, t(996) = -3.05, p = .002), but not for gratitude (b = -0.11, SE = 0.08, t(996) = -1.43, p = .153).

Finally, the gratitude prime did not produce stronger effects than the lack of control prime for digital twins (b = 0.04, SE = 0.12, t(1538) = 0.33, p = .744), but gratitude produced stronger effects for humans (b = 0.50, SE = 0.15, t(1760) = 3.39, p < .001).

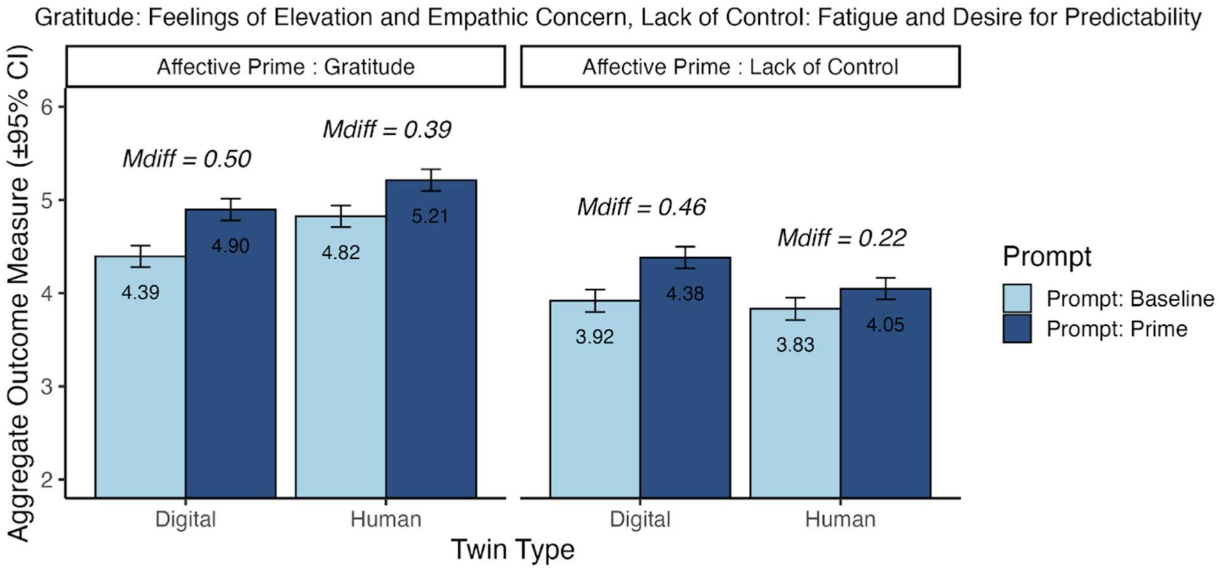


Figure 2. Effect of Affective Prime on Aggregate Downstream Outcomes.

Individual Dependent Measures

We also pre-registered looking at each dependent measure independently. Thus, we ran four linear mixed-effects regressions with (1) elevation, (2) empathic concern, (3) fatigue, and (4) desire for predictability as the dependent measure and condition (manipulation vs. control), twin (human vs. digital), and their interaction as independent variables, and twin identifier as a random intercept. Each measure's simple effect for prompt (prime vs. baseline) was significant for both humans and digital twins except empathic concern, which was marginal for twins (Mbaseline = 5.04, SD = 1.48, Mprime = 5.28, SD = 1.40, b = 0.238, SE = 0.136, t(731) = 1.744, p = .082) and not significant for humans (Mbaseline = 5.45, SD = 1.61, Mprime= 5.44, SD = 1.64, b = -0.01, SE = 0.14, t(731) = -0.07, p = .947), and desire for predictability, which was not significant for twins (Mbaseline = 5.54, SD = 1.04, Mprime= 5.59, SD = 1.04, b = 0.05, SE = 0.09, t(718) = 0.47, p = .636).

The effect of prompt (prime vs. baseline) on elevation did not differ for digital twins and humans (b = 0.02, SE = 0.2 t(506) = 0.16, p = .872). The effect of prompt (prime vs. baseline) was stronger for humans than digital twins for empathic concern (b = 0.25, SE = 0.12, t(506) = 2.08, p = .039) and fatigue (b = 0.70, SE = 0.09, t(490) = 8.26, p < .001). The effect of prime on desire for predictability was stronger for humans than for digital twins (b = 0.21, SE = 0.08, t(490) = 2.45, p = .015). See Figure 3.

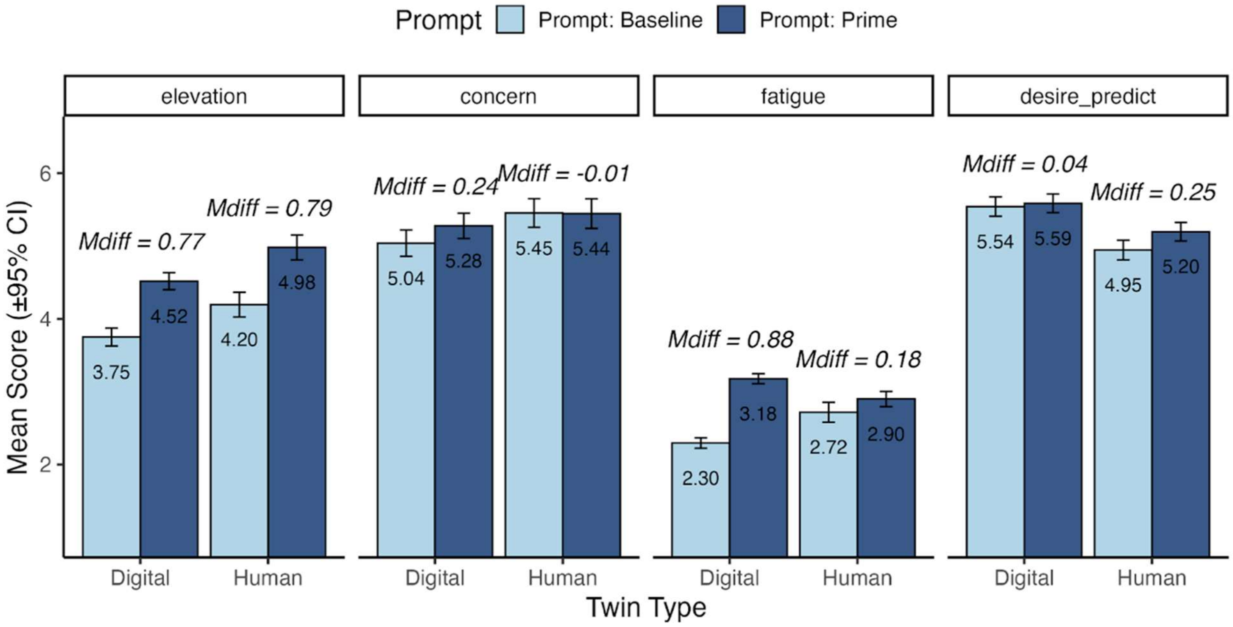


Figure 3. Measures by Prompt (Prime vs. Baseline) Grouped by Measure and Twin Type.

Results - Additional Analyses (Non-Preregistered)

Here, we examined the language used when responding to the prompt. Specifically, we examined the language similarity between a human participant's response and their twin versus a human participant's response and a randomly selected twin using the BERT metric, the "sum of cosine similarities between their tokens' embeddings" (Zhang, Kishore, Wu, Weinberger, & Artzi, 2019). In a random effects regression with language matching as the dependent variable, same twin versus different twin pairing as the independent variable, condition as a covariate, and human participant identifier as random effects, we find that the language between a human participant and their twin is slightly more similar (*M* = .11, SD= 0.09) than between a human participant and a random twin (*M* = 0.13, *SD* = 0.10, *b* = 0.014, *SE* = 0.002, *t*(999) = 5.54, *p* < .001). Figure 4 shows this by condition.

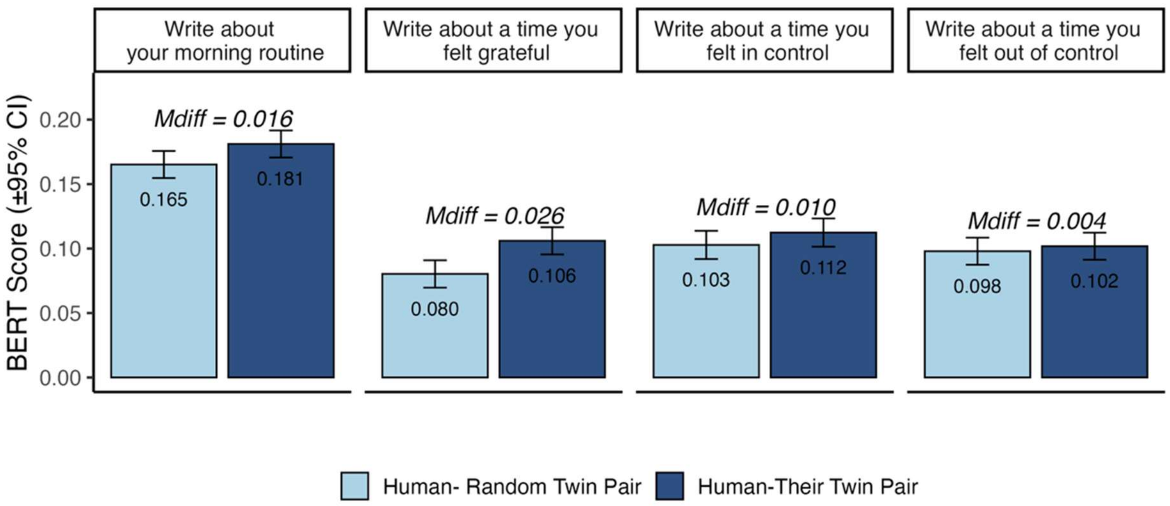


Figure 4. Similarity Between Human Prompt Text and Their Digital Twin's Prompt Text vs. a Random Digital Twin's Prompt Text.

Discussion

Returning to each research question:

1. Do affective priming manipulations induce "states" in digital twins (e.g., does writing about gratitude or lack of control momentarily influence digital twins' responses?)

Yes. Writing about feeling grateful or out of control significantly increased corresponding feelings of gratitude and lack of control in digital twins, with these induced "affective states" subsequently influencing downstream outcomes.

2. Is the influence of affective priming similar for digital twins and their human counterparts?

No. Affective priming had stronger effects on digital twins' affective responses compared to humans for both gratitude and lack of control manipulations.

3. Is the influence of affective priming on digital twins dependent on the valence of affective prime (i.e., between positive primes like gratitude or negative primes like lack of control) or the proximal nature of the dependent measures (i.e., does it "spill over" to other, related dimensions or only influence proximal dimensions)?

Valence. Yes, valence affects twins' responses, but in a similar way to humans: The gratitude prime had a stronger effect on the manipulation check than the lack of control prime for both humans and their twins.

Proximal Nature of DVs. Yes, the influence of affective priming on digital twins is dependent on how closely related the dependent measure is to the manipulation. Specifically, the impact of affective prime on proximal dependent measures was consistent and as expected for both twins and humans, but impact of affective prime on distal downstream measures was inconsistent between twins and humans.

For proximal dependent measures (measures closely linked to the manipulation), we observe a “spillover” effect of prime in both the gratitude prime and lack of control prime for humans and twins. For the distal downstream measure (more distantly related to the manipulation) in the gratitude condition, the impact of affective prime was marginal for twins and non-significant for humans. For the distal downstream measure in the lack of control condition, the impact of affective prime was non-significant for twins and significant for humans.

4. Is there a relationship between what digital twins and their human counterparts say in their response to the manipulation?

Yes. The language between a human participant and their twin is slightly more similar than between a human participant and a random twin.

References

Bukowski, M., Potoczek, A., Barzykowski, K., Lautenbacher, J., & Inzlicht, M. (2024). What do we manipulate when reminding people of (not) having control? In search of construct validity. Behavior Research Methods, 56(4), 3706-3724.

Chen, C. Y., Lee, L., & Yap, A. J. (2017). Control deprivation motivates acquisition of utilitarian products. Journal of Consumer Research, 43(6), 1031-1047.

DeSteno, D., Li, Y., Dickens, L., & Lerner, J. S. (2014). Gratitude: A tool for reducing economic impatience. Psychological Science, 25(6), 1262-1267.

Emmons, R. A., & McCullough, M. E. (2003). Counting blessings versus burdens: An experimental investigation of gratitude and subjective well-being in daily life. Journal of Personality and Social Psychology, 84(2), 377-389.

Greenaway, K. H., Haslam, S. A., Cruwys, T., Branscombe, N. R., Ysseldyk, R., & Heldreth, C. (2015). From “we” to “me”: Group identification enhances perceived personal control with consequences for health and well-being. Journal of personality and social psychology, 109(1), 53.

Lembregts, C., & Pandelaere, M. (2019). Falling back on numbers: When preference for numerical product information increases after a personal control threat. Journal of Marketing Research, 56(1), 104-122.

Greenaway, K. H., Haslam, S. A., Cruwys, T., Branscombe, N. R., Ysseldyk, R., & Heldreth, C. (2015). From “we” to “me”: Group identification enhances perceived personal control with consequences for health and well-being. Journal of personality and social psychology, 109(1), 53.

Oguni, R., & Ishii, C. (2024). Gratitude promotes prosocial behavior even in uncertain situation. Scientific Reports, 14, 14379.

Oliveira, R., Baldé, A., Madeira, M., Ribeiro, T., & Arriaga, P. (2021). The impact of writing about gratitude on the intention to engage in prosocial behaviors during the COVID-19 outbreak. Frontiers in Psychology, 12, 588691.

Schnall, S., Roper, J., & Fessler, D. M. T. (2010). Elevation leads to altruistic behavior. Psychological Science, 21(3), 315-320.

Sedek, G., & Kofta, M. (1990). When cognitive exertion does not yield cognitive gain: toward an informational explanation of learned helplessness. Journal of personality and Social Psychology, 58(4), 729.

Walsh, Lisa C., et al. "More than merely positive: The immediate affective and motivational consequences of gratitude." Sustainability 14.14 (2022): 8679.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. Journal of personality and social psychology, 67(6), 1049.

Whitson, J. A., & Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. Science, 322(5898), 115-117

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675

Obedient Twins

Main Questions/Hypotheses

Surveys often require participants to follow instructions—for example, to earnestly consider a viewpoint or imagine a scenario. Given that LLMs are trained to be obedient and deferential, might digital twins be more sensitive to survey instructions relative to their human counterparts? We tested this in three tasks:

1. Self-persuasion (Briñol, McCaslin, and Petty 2012). When prompted to consider the other side, do digital twins abandon their attitudes more readily than their human counterparts?
2. Scenarios. Do digital twins more "obediently" follow the instructions to imagine themselves in different scenarios, leading to more sensitivity to scenario manipulations?
3. Absurd Scenarios. Do digital twins "earnestly" respond to instructions that would be non-sensical to their human counterparts?

Methods

One thousand and one human participants on Prolific participated in the study. The study was subsequently run on their 1,001 digital twins, for a total of 2,002 participants. Participants and their corresponding digital twins took part in three tasks.

For the first task, we followed the self-persuasion condition in Catapano, Tormala, and Tucker (2019). Specifically, all participants first indicated their attitude toward universal basic income (UBI) using a 0-100 scale in response to the statement "America should have a universal basic income system"? where 0 = Strongly disagree (I am against universal basic income) and 100 = Strongly agree (I am in favor of universal basic income). Then, following the attitude measure, participants were instructed to take the alternate point of view. Due to technical constraints with dynamic surveys, the wording differed slightly between human and digital twin participants.

Human participants received the following prompt: "We'd like you now to consider the alternate point of view. That is, think about convincing reasons why universal basic income in America might be a good idea/a bad idea. After reflecting, please write one argument which supports/opposes universal basic income system in America that you find personally compelling. One compelling argument in support/against universal basic income is…"

Digital twins received a modified version: "We'd like you now to consider the alternate point of view. After reflecting, please write one argument which takes a different position from yours on universal basic income system in America that you find personally compelling. One compelling argument for the alternative point of view on universal basic income system in America is…".

Lastly, participants read "Now, after considering the alternate point of view we would like to ask you again…" and indicated their attitude on the same scale. Our dependent variable was attitude change, calculated as the difference between post- and pre-reflection attitude scores. We coded this variable such that positive values indicate movement toward the opposing viewpoint (i.e., participants initially supporting UBI becoming less supportive, or participants initially opposing UBI becoming more supportive), while negative values indicate further polarization in the participant's original direction (i.e., supporters becoming more supportive, or opponents becoming more opposed). For participants with neutral initial attitudes, any shift was coded as positive.

For the second task, participants were asked to imagine three scenarios. For each scenario, we randomly assigned participants to a baseline condition or treatment condition, reflecting a 2 (control vs. treatment) × 3 (scenario) mixed design. Digital twins were assigned the same conditions as their paired human participants.

In all scenarios, the baseline condition established the context (S1: “Imagine you're working on a team project”; S2: “Imagine you're given a puzzle to solve”; S3: “Imagine you're lying in bed, ready to fall asleep, and you have a headache”), while the treatment condition was designed to elicit a psychological response (S1: “Imagine you're working on a team project with a tight deadline. One of your teammates hasn't delivered their part on time”; S2: “Imagine you're given a difficult puzzle to solve under time pressure, and others are watching you attempt it”; S3: “Imagine you're lying in bed, ready to fall asleep, and you feel a sharp pain in your foot”). We were interested in how responsive to the imagined “treatment” scenarios digital twins would be compared to humans.

To test this, after each scenario, participants responded to two questions gauging their response to the scenario (S1: “How confident are you in your ability to manage conflict on this team?” and “How much do you trust your team?”; S2: “How anxious do you feel?” and “How confident are you in your ability to solve this puzzle?”; S3: “How certain are you that this is serious?” and “How likely are you to seek help?”; all rated from 1 = not at all to 7 = very much). For the third task, participants were given three absurd scenarios. Similar to the second task, for each scenario, we randomly assigned participants to one of two conditions, reflecting a 2 (condition A vs. condition B) × 3 (scenario) mixed design. Digital twins were assigned the same conditions as their paired human participants.

The conditions were absurd manipulations that would ostensibly be nonsensical to participants and therefore difficult to imagine. The first scenario asked participants to imagine either, “You are an echo that occurs before the sound it repeats. You emerge into the world slightly ahead of time, announcing something no one has said yet. People are confused when they hear you, unsure what you're responding to” or “You are an echo that occurs during the sound it repeats. You overlap almost perfectly with the original noise, tangled in its vibration, indistinguishable yet still somehow separate. People hear you and feel something is slightly off, but they can't say why.” Participants then indicated “How content are you with your existence?” (1 = not at all to 7 = very content). In the second scenario, participants were asked to imagine either, “You can hear certain smells” or “You can smell certain sounds,” then indicated “How powerful do you feel?” (1 = not at all to 7 = very powerful).

Finally, participants were asked to imagine either, “You are a mirror. You exist in a reality where photons have never been invented. No one has ever seen their reflection in you, yet you hold the capacity for it” or “You are a mirror. When someone looks at you, they don't see themselves—they see a version from 11 minutes ago, always 11 minutes. You don't reflect appearance, only temporal shadows.” They then indicated “How authentic do you feel?” (1 = not at all to 7 = very authentic).

For these absurd scenarios, we had no directional hypothesis and simply compared the effect of these absurd manipulations on humans and digital twins. We were curious whether, because these arbitrary and absurd situations are difficult to imagine, humans would be less sensitive to the manipulation than digital twins, who would earnestly follow the instruction.

Results - Pre-registered Analyses

Task 1. To examine whether humans and digital twins differ in the extent to which they are willing to adjust attitudes, we ran a linear mixed-effects regression using the lmerTest package in R (Kuznetsova, Brockhoff, and Christensen 2017), with attitude change as the dependent variable, twin (human vs. digital) as the independent variable, baseline attitude as a control, and twin identifier as a random intercept. We find that both humans and digital twins significantly change their attitude after considering the other side, but humans changed their attitudes more (M = 6.17, SD = 14.8) than digital twins (M =2.80, SD = 2.43, b = 3.45, SE = 0.47, t(1001) = 7.37, p < .001).

Task 2. We pre-registered to average the two measures for each scenario, and we determined that the correlations were high enough to justify this decision (S1: r_{human} = .57, $r_{digital}$ = .36; S2: r_{human} = .42, $r_{digital}$ = .60; S3: r_{human} = .65, $r_{digital}$ = .71).

To examine the extent to which humans and digital twins are sensitive to manipulations in imagined scenarios, we ran a linear mixed-effects regression with the scenario response scores as the dependent variable, manipulation (baseline vs. treatment), twin (human vs. digital) and their interaction as independent variables, and scenario and twin identifier as a random intercept. The results revealed a significant interaction (b = -0.16, SE = 0.06, t(5000) = -2.77, p = .006). Specifically, twins were significantly less affected by the manipulation (control: M = 2.62, SD = 1.16; treatment: M = 3.07, SD = 1.27, b = 0.43, SE = 0.04, t(5734) = 9.92, p < .001) than humans (control: M = 2.90, SD = 1.32; treatment: M = 3.51, SD = 1.51, b = 0.59, SE = 0.04, t(5734) = 13.58, p < .001).

Task 3. To examine the extent to which humans and digital twins are sensitive to manipulations in absurd scenarios, we ran three linear mixed-effects regression with the response as the dependent variable, absurd manipulation (A vs. B), twin (human vs. digital) and their interaction as independent variables, and twin identifier as a random intercept. Results revealed significant interactions for all three absurd scenarios (see figure 1 below). Unexpectedly, in the absurd scenario regarding echoes and absurd scenario regarding smells and sounds, we observed a crossover interaction. Because the direction of the effect was not meaningful to us (after all, these are intentionally absurd), we re-coded the condition so that the direction of the effect was the same to determine whether twins were more sensitive to the condition than humans (regardless of effect direction).

Once removing the cross-over, we find no significant interaction for the echoes scenario (b = 0.02, SE = 0.16, t(999) = 0.14, p = .892) or smells/sounds scenario (b = -0.07, SE = 0.14, t(999) = -0.49, p = .625). The effect of the condition on twins (echoes scenario: M_{diff} = .15, b = -0.15, SE = 0.11, t(1961) = -1.43, p = .152; smells/sounds scenario: M_{diff} = .19, b = -0.19, SE = 0.09, t(1982) = -2.04, p = .042) was similar to the effect of condition on humans (echoes scenario: M_{diff} = .13, b = -0.13, SE = 0.11, t(1961) = -1.23, p = .220; smells/sounds scenario: M_{diff} = .26, b = -0.26, SE = 0.09, t(1982) = -2.76, p = .006) for both the echoes and smelss/sounds absurd scenarios. For the mirrors absurd scenario, we observe a significant interaction (b = -0.33, SE = 0.14, t(999) = -2.35, p = .019). The effect was stronger for humans (M_{diff} = .88, b = -0.88, SE = 0.10, t(1981) = -8.54, p < .001) relative to twins (M_{diff} = .56, b = -0.55, SE = 0.10, t(1980.81) = -5.38, p < .001). See Figure 1.

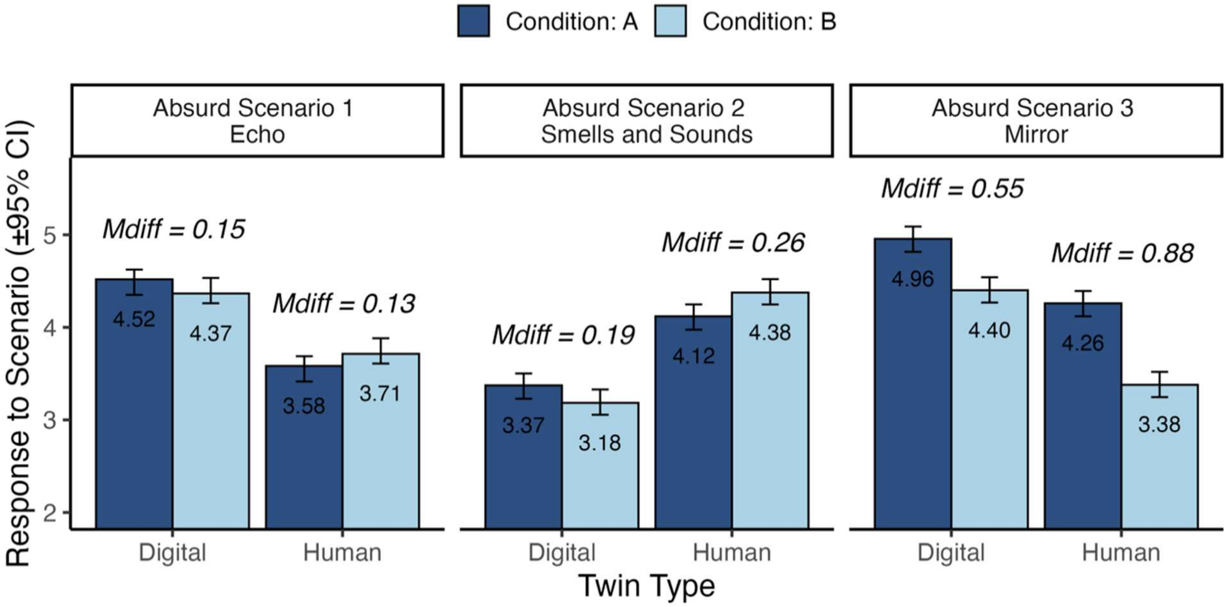


Figure 1. Effect of Absurd Scenarios on Digital Twins versus Humans.

Results - Additional Analyses (Non-Preregistered)

Task 1. We wondered if twins' lack of attitude change could be because they may not "experience" the metacognition such as attitude certainty. To examine this possibility, we examine the moderating role of attitude extremity. Research suggests that attitude extremity is one determinant of attitude strength—in other words, people are more certain when extreme and thus should be less likely to change their mind when their attitudes (Petty & Krosnick, 2014).

To test the effect of attitude certainty on attitude change in humans and digital twins, we first re-coded pre-attitude into attitude extremity—where 0 reflects neutral, and any number reflects the absolute difference from 0 (so 50 could be totally in favor or against).

We then ran a linear mixed-effects regression with attitude change as the dependent variable, attitude extremity (from 0 to 50), twin (human vs. digital) and their interaction as independent variables, and twin identifier as a random intercept. We find a significant interaction (b = -0.11, SE = 0.03, t(1795) = -3.18, p = .002). For humans, as attitude extremity increases, their attitude change decreases (b = -0.04, SE = 0.02, t(1998) = -2.13, p = .033). The opposite pattern is present in digital twins (b = 0.06, SE = 0.03, t(1998) = 2.36, p = .018), which is consistent with a regression to the mean. This provides initial evidence that twins might not reflect the metacognition of their human counterparts.

We also examined the language used when considering the other side. Specifically, we examined the language similarity between a human participant's text response and their twin's text response versus the text response of a randomly selected twin using the BERT metric, which is the "sum of cosine similarities between their tokens' embeddings." (Zhang, Kishore, Wu, Weinberger, & Artzi, 2019). In a random effects regression with language matching as the dependent variable, same twin versus different twin pairing as the independent variable, and a human participant identifier as a random effect, we find that the language between a human participant and their twin is slightly more similar (M = .19, SD= 0.09) than between a human participant and a random twin (M = 0.18, SD = 0.09; b = 0.009, SE = 0.001, t(1000) = 5.84, p < .001).

Task 2. To examine the extent to which humans and digital twins are sensitive to manipulations for each of the three imagined scenarios, we ran three linear mixed-effects regressions with the response to scenario for each of the three scenarios (group project, puzzle, lying in bed) as the dependent variables, manipulation (baseline vs treatment), twin (human vs. digital) and their interaction as independent variables, and twin identifier as a random intercept.

Each scenario yielded a significant interaction (group project: b = 0.34, SE = 0.10, t(4994) = 3.48, p < .001; puzzle: b = 0.59, SE = 0.10, t(4994) = 5.99, p < .001; lying in bed: b = -0.45, SE = 0.10, t(4994) = -4.60, p < .001). For the team project scenario, humans (control: M = 2.97, SD = 1.17; treatment: M = 3.45, SD = 1.23; b = 0.44, SE = 0.07, t(5710) = 5.92, p < .001) were more affected by imagining conflict than digital twins (control: M = 2.83, SD = 1.20; treatment: M = 2.97, SD = 1.18; b = 0.10, SE = 0.07, t(5710) = 1.30, p = .193). Similarly, for the puzzle scenario, humans were more affected by imagining time pressure (control: M = 3.15, SD = 1.24; treatment: M = 4.37, SD = 1.34; b = 1.21, SE = 0.07, t(5711) = 16.38, p < .001) than digital twins (control: M = 2.82, SD = 1.36; treatment: M = 3.45, SD = 1.52; b = 0.62, SE = 0.07, t(5711.16) = 8.44, p < .001). For lying in bed scenario, digital twins showed larger responses to imagining leg pain (control: M = 2.21, SD = 0.713; treatment: M = 2.78, SD = 0.94; b = 0.57, SE = 0.07, t(5710) = 7.74, p < .001) than humans (control: M = 2.59, SD = 1.47; treatment: M = 2.71, SD = 1.46; b = 0.12, SE = 0.07, t(5710) = 1.65, p = .100). See Figure 2.

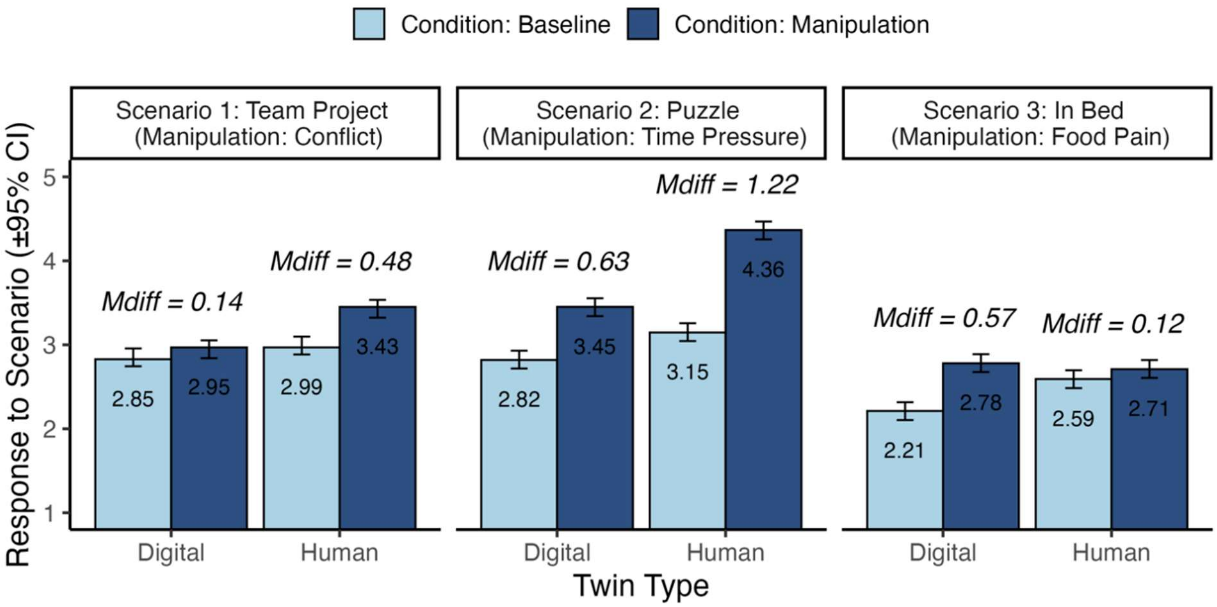


Figure 2. Effect of Imagined Scenarios on Digital Twins vs. Humans.

Discussion

Returning to each research question:

1. Self-persuasion (Briñol, McCaslin, and Petty 2012). When prompted to consider the other side, do digital twins abandon their attitudes more readily than their human counterparts?

No. In fact, twins were less sensitive to considering a different point of view than humans (although this could be explained by our slight wording modification between humans and twins). Additional analyses suggest this also could be due to twins not reflecting the metacognition of their human counterparts.

2. Scenarios. Do digital twins more “obediently” follow the instructions to imagine themselves in different scenarios, leading to more sensitivity to scenario manipulations?

No. In fact, twins were less sensitive overall to these scenario manipulations than humans. Looking at individual scenarios, twins were less sensitive than humans for imagined conflict and time pressure, but more sensitive when imagining pain.

3. Absurd Scenarios. Do digital twins “earnestly” respond to instructions that would be non-sensical to their human counterparts?

No. If anything, humans were more sensitive to the instructions to forecast feelings in absurd scenarios.

In sum, across all three task types, we do not find evidence supporting the possibility that because twins obediently follow orders, they are more sensitive to survey instructions, such as "consider the alternative point of view" or "imagine..." If anything, we find that humans, on average, respond more to instructions in this survey than digital twins.

References

Catapano, R., Tormala, Z. L., & Rucker, D. D. (2019). Perspective taking and self-persuasion: Why “putting yourself in their shoes” reduces openness to attitude change. Psychological science, 30(3), 424-435.

Briñol, P., McCaslin, M. J., & Petty, R. E. (2012). Self-generated persuasion: effects of the target and direction of arguments. Journal of Personality and Social Psychology, 102(5), 925.

Petty, R. E., & Krosnick, J. A. (2014). Attitude strength: Antecedents and consequences. Psychology Press.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Hiring Algorithms

Main Questions/Hypotheses

Digital twins, or AI agents trained to simulate individual behaviors, have the potential to reshape how job seekers and employers interact. From the candidate side, digital twins can help streamline the job search process by identifying roles aligned with the job seeker's preferences, tailoring job application materials, and even interacting with recruiters on the job seeker's behalf. For firms, digital twins enable the screening of candidates according to the firm's preferences, the simulation of candidate responses to different human resource policies, and even candidate outreach and outbound recruiting efforts. The advent of digital twins, and a growing ecosystem of start-ups offering such services, may thus fundamentally reshape the dynamics of hiring, and employee-employer matching more generally.

Despite the growing interest in digital twins, we lack evidence of their effectiveness to capture workplace preferences. While digital twins have shown promise in modeling preferences in more structured domains, there are unique challenges in the workplace context. For example, workplace preferences are multi-dimensional and candidates must weigh trade-offs between attributes like compensation, skill development, flexibility, and culture. Moreover, workplace preferences are shaped by social context and may vary over time. These features call into question the extent to which current digital twins can accurately represent how employees and firms navigate the matching process.

In this sub-study, we examine the ability of digital twins to predict job candidate preferences over workplace attributes using a two-stage study. In the first stage, we elicit the stated preferences of job candidates over a range of workplace characteristics, including work-life balance, career development, and company culture. We then ask each participant's digital twin to report their preferences over the same attributes. By comparing the stated and simulated preferences, we can understand how well digital twins predict the stated preferences of their human counterparts. In the second stage, we use a within-subjects experiment where participants evaluate four job postings. Our manipulation here changes whether the firm's hiring is done by algorithms or humans, and we randomly assign two of the four job postings to have AI recruiters. Both humans and their digital twins evaluate each of the four job postings on eight key attributes, including transparency, fairness, and likelihood of applying. This design thus allows us to understand the efficacy of digital twins in predicting workplace preferences and experiments in the HR domain.

Methods

We run a two-stage study to examine these questions. We outline these stages in Figure 1. Both humans and their digital twins completed the study in the same order.

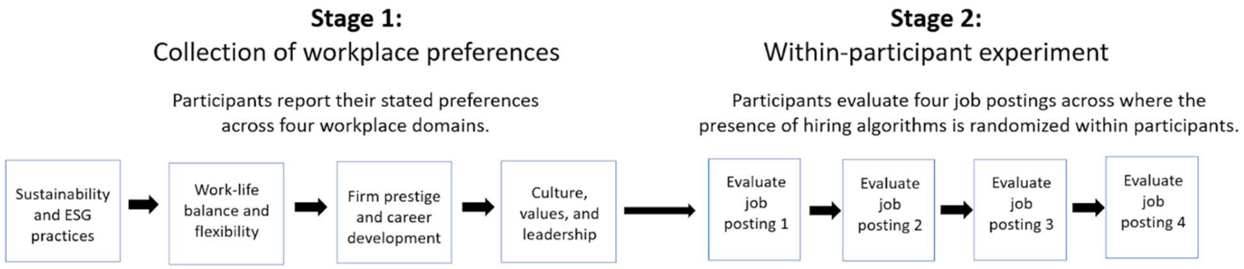

Stage 1:
Collection of workplace preferences
Participants report their stated preferences across four workplace domains.
Sustainability and ESG practices
Work-life balance and flexibility
Firm prestige and career development
Culture, values, and leadership
Stage 2:
Within-participant experiment
Participants evaluate four job postings across where the presence of hiring algorithms is randomized within participants.
Evaluate job posting 1
Evaluate job posting 2
Evaluate job posting 3
Evaluate job posting 4

Figure 2. Outline of the study.

The first stage of our study collects data on respondent workplace preferences. We asked respondents a total of eight questions across four distinct themes related to workplace preferences. These themes included:

- **Sustainability and ESG practices**
  - "How important is it to you that your employer actively invests in environmental sustainability (e.g., reducing carbon emissions, minimizing waste)?”
  - "Would you prefer to work at a company that is outspoken about social and political issues, or one that remains neutral?"
- **Work-life balance and flexibility**
  - "How likely are you to accept a job that requires you to be in the office five days a week?"
  - "Which of the following is more important to you when choosing a job? Where options included: (i) A flexible schedule that allows me to manage personal responsibilities; (ii) Clear boundaries between work and personal time, (iii) Strict 9-to-5 hours (iv) A predictable, fixed schedule[6]
- **Firm prestige & career development**
  - "How much do you value opportunities for formal training and upskilling in your current or future role?"
  - "Imagine two job offers: one from a prestigious company with a competitive culture, and one from a lesser-known company with a collaborative, supportive environment. Which would you choose?"[7]
- **Culture, values, and leadership**

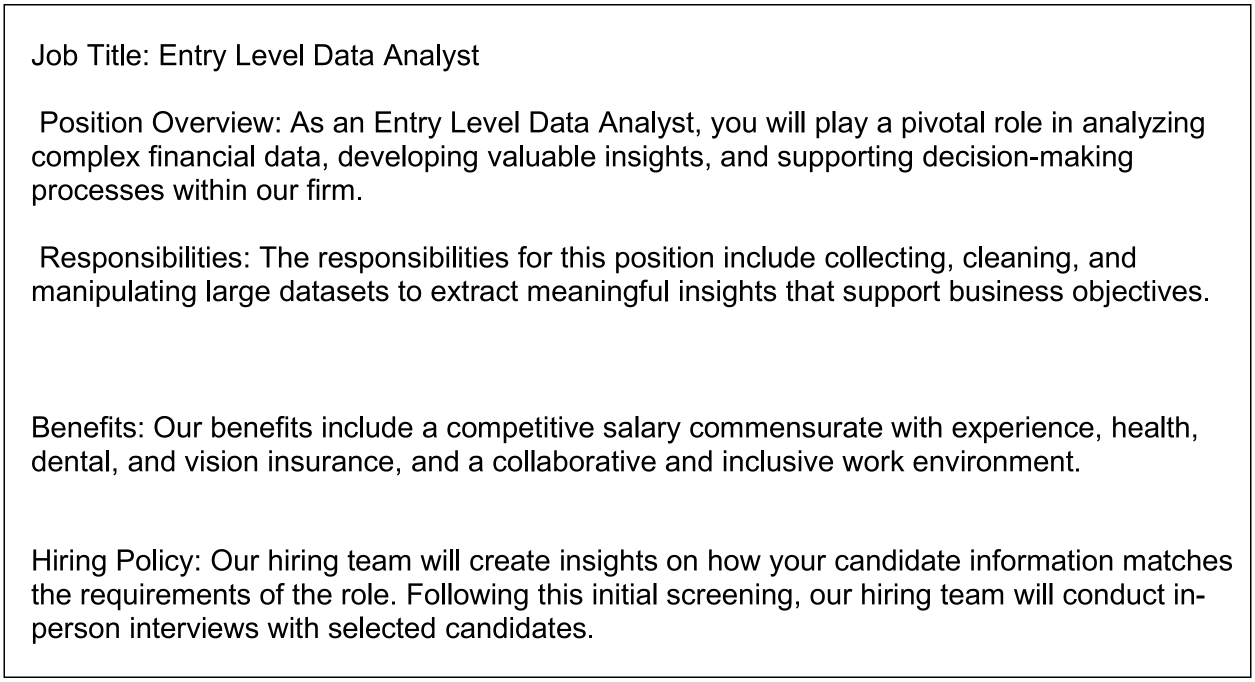
[6] To construct the index measuring respondents’ preferences for flexible work–life balance arrangements, we ordered the four response options from least to most flexible as follows: (1) *Strict 9–5 hours*, (2) *Predictable, fixed schedule*, (3) *Clear boundaries between work and personal time*, and (4) *Flexible schedule that allows me to manage personal responsibilities*. This index captures increasing preference for autonomy in managing one’s work schedule.

[7] To construct the index for collaborative culture preference, we ordered the responses to reflect increasing preference for collaborative environments: (1) *Prestigious*, (2) *Depends on compensation*, (3) *Not sure*, and (4) *Collaborative*. This index captures respondents’ orientation toward workplace culture, with higher scores indicating stronger preferences for collaboration over prestige or ambiguity.

- "How important is it that your company's leadership communicates transparently about the firm's goals and challenges?"
- "How much do you care about your company's stated mission and values aligning with your personal beliefs?"

We collect these measures for both human participants and their digital twins. Thus for each participant, we have their stated preference regarding (i) sustainability & ESG practices; (ii) work-life balance and flexibility; (iii) firm prestige and career development; and (iv) culture and values. We also have the corresponding responses from the digital twins, allowing us to compare how well LLMs can predict job candidate preferences in the hiring process.

The second stage of our study uses an experiment with within-subjects variation to understand how candidates respond to algorithms in the hiring process. In this part, participants evaluate hiring policies from the perspective of a job seeker. We present participants with four different job postings, which display a position overview, the responsibilities, benefits, and hiring policy (See Figure 2 for an example). Our randomization manipulates whether hiring uses either human resume screeners or algorithmic ones. Following each posting, participants evaluated the job on eight questions (displayed in Table 1 below), using a 7-point Likert scale from strongly disagree to strongly agree.



Job Title: Entry Level Data Analyst

Position Overview: As an Entry Level Data Analyst, you will play a pivotal role in analyzing complex financial data, developing valuable insights, and supporting decision-making processes within our firm.

Responsibilities: The responsibilities for this position include collecting, cleaning, and manipulating large datasets to extract meaningful insights that support business objectives.

Benefits: Our benefits include a competitive salary commensurate with experience, health, dental, and vision insurance, and a collaborative and inclusive work environment.

Hiring Policy: Our hiring team will create insights on how your candidate information matches the requirements of the role. Following this initial screening, our hiring team will conduct in-person interviews with selected candidates.

Figure 3. Sample job posting.

#	Question	Theme
1	I would apply for this position.	Likelihood of applying
2	I would be hired for this position.	Likelihood of receiving offer
3	I will receive clear and transparent communication throughout the hiring process.	Transparency
4	The firm's hiring policies are fair and unbiased.	Fairness
5	The hiring process will move at a reasonable pace and ensure a timely decision-making process.	Efficiency and timeliness
6	The hiring process will give me a clear understanding of the job requirements, expectations, and the company culture.	Information
7	The hiring process will effectively identify top job applicants for the position.	Identifies top talent
8	The firm's hiring process suggests the firm fosters a high degree of social interaction.	Sociality

Table 7. Questions regarding job postings.

Results - Pre-registered Analyses

We first examine whether digital twins can accurately predict human workplace preferences using two methods. First, we run paired t-tests to test whether on average the twin outcomes equal the actual human outcomes. Second, we use regression to estimate the size and direction of the bias.

Table 2 displays the results of the paired t-tests across all of our questions. The top panel contains the outcomes from the first phase of the study, while the bottom panel displays them from the second phase of the study. Column 1 displays the average human response while column 2 displays the average digital twin response. Column 3 displays the average difference in responses for the digital twin versus the human, with the corresponding t-test p-value in column 4. Lastly, column 5 displays the sample size for each group.

The results in Table 2 indicate that, on average, the outcomes generated by digital twins are statistically different from those generated by humans. Across the eight workplace preferences in phase 1 of the study (Panel A of Table 2), we only find evidence of no differences between the two for workplace flexibility preferences. For the other categories, we find statistically significant differences between the two. For most, the digital twins over-estimate human preferences for workplace attributes (for example, sustainability, ESG practices, collaboration, and transparency in leadership). Meanwhile, the digital twins under-estimate their human counterparts' preferences for career development opportunities.

We find qualitatively similar results for the second phase of the study, which we display in Panel B of Table 2. We find statistically significant differences across all outcomes here. Like before, the digital twins over-estimate some responses (for example, the likelihood of being hired for the position) and under-estimate others (the firm's degree of social interaction, and receiving clear information regarding job requirements, expectations, and company culture).

Panel A: Phase 1 Outcome

	Average		Difference	T-test	
	Human	Digital Twin	DT - H	p-value	N
Outcomes	(1)	(2)	(3)	(4)	(5)
Sustainability	2.88	3.45	0.57	0.00	999
ESG Practices	3.30	3.67	0.37	0.00	999
Work Life Balance	3.36	3.33	-0.03	0.40	999
Flexibility	2.97	2.95	-0.02	0.70	999
Collaboration over prestige	2.98	3.65	0.67	0.00	999
Career Development	3.67	3.56	-0.11	0.01	999
Transparent Leadership	3.83	4.43	0.60	0.00	999
Culture & Values	3.39	3.54	0.15	0.00	999

Panel B: Phase 2 Outcomes

	Average		Difference	T-test	
	Human	Digital Twin	DT - H	p-value	N
Outcomes	(1)	(2)	(3)	(4)	(5)
I would apply for this position	3.34	3.48	0.14	0.00	999
I would be hired for this position	3.46	4.22	0.76	0.00	999
I will receive clear and transparent communication throughout the hiring process	4.89	4.79	-0.10	0.00	999
The firm's hiring policies are fair and unbiased	4.88	4.59	-0.29	0.00	999
The hiring process will move at a reasonable pace and ensure a timely decision-making process	5.01	4.55	-0.46	0.00	999
The hiring process will give me a clear understanding of the job requirements, expectations, and the company culture	5.20	4.70	-0.50	0.00	999

The hiring process will effectively identify top job applicants for the position	5.05	4.60	-0.45	0.00	999
The firm's hiring process suggests the firm fosters a high degree of social interaction	4.61	4.05	-0.56	0.00	999

Table 2: Comparison of human versus digital twin responses using t-tests.

The regression tests of Hypothesis 1 lead to similar takeaways. In Table 3, we present the results of a regression of the human response on the twin's response with robust standard errors. Columns 1 and 2 display the coefficient and standard error on the twin's response, respectively. Column 3 displays the p-value from a test of whether the coefficient on the twin response equals one. The corresponding null hypothesis is that digital twins can perfectly predict human responses (i.e., β=1). As the table illustrates, however, we can reject the null across all of our outcomes. Thus, while the digital twins can predict workplace preferences for some participants, on average they struggle in this domain.

Panel A: Phase 1 Outcomes

	β	Std. Err.	p-value
Outcomes	(1)	(2)	(3)
Sustainability	0.68	0.02	0.00
ESG Practices	0.55	0.02	0.00
Work Life Balance	0.11	0.03	0.00
Flexibility	0.14	0.03	0.00
Collaboration over prestige	0.12	0.04	0.00
Career Development	0.32	0.04	0.00
Transparent Leadership	0.45	0.05	0.00
Culture & Values	0.46	0.04	0.00

Panel B: Phase 2 Outcomes

	β	Std. Err.	p-value
Outcomes	(1)	(2)	(3)
Apply	0.70	0.12	0.08

Hired	0.27	0.06	0.00
Communication	0.39	0.07	0.00
Unbiased hiring policies	0.41	0.06	0.00
Timely decision-making process	0.38	0.06	0.00
Clear information	0.39	0.07	0.00
Identify top applicants	0.39	0.07	0.00
High degree of sociality	0.04	0.04	0.00

Table 3: Comparison of human versus digital twin responses using regression

In our pre-analysis plan, we were also interested in understanding if digital twins can better predict outcomes for various demographic groups. We were specifically interested in differences by gender, age, and race. For this analysis, we use the t-test strategy above. We conduct a t-test for each specific subsample and report the p-values in Table 4 below. The last row of the table counts the number of times we can reject the null using a p-value cut-off of p<0.05. In other words, this row calculates the number of outcomes for the given subgroup where digital twin responses are statistically indistinguishable from their human responses. Our results provide suggestive evidence that digital twins do better at predicting outcomes for women versus men, as we can reject the null of no differences between the human and twin responses for three outcomes for women (versus two for men). Similarly, the twins seem to do better for older responses versus younger ones. Overall, however, the differences across subgroups are quite limited.

	P-value from t-test							
	Gender		Race		Age			
	Female	Male	White	Non-White	18-29	30-49	50-64	65+
Outcomes	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sustainability	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ESG Practices	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Work Life Balance	0.33	0.80	0.00	0.00	0.05	0.02	0.21	0.73
Flexibility	0.83	0.74	0.12	0.00	0.19	0.37	0.00	0.41
Collaboration	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Career Development	0.17	0.01	0.03	0.10	0.00	0.00	0.25	0.07
Transparent Leadership	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Culture & Values	0.00	0.01	0.00	0.01	0.00	0.03	0.00	0.00

Table 4: Comparison of human versus digital twin responses using t-tests, by subgroup

We next examine whether experimentation on the digital twins can uncover the true treatment effect in the human sample. Although the previous results show that digital twins responses are biased across workplace attributes, valid causal inference may still be possible. For example, if the digital twins over-estimate the responses from all participants, estimation using the digital twins may still uncover the true average treatment effect in the human population if this bias is constant across treatment arms.

Our experimental manipulation is the use of algorithms in a firm's hiring policy. Each participant saw four job postings, where we randomly assigned two to use hiring algorithms (vs human screeners). For our estimation here, we regress our outcomes on a binary indicator for whether the job included a hiring algorithm, participant fixed effects, and job-type fixed effects. We do so for both our human sample and our digital twins sample, and plot the results in Figure 3. Moreover, we use a stacked regression to compare the treatment effects from our human vs digital twin sample. For this, we append the human and digital twin samples and create a binary indicator for the twin sample. We then regress each outcome on an indicator for the hiring algorithm condition, an indicator for the twin sample, and an interaction between the two. The interaction captures the extent to which the treatment effect in the twin sample is different from the effect in the human sample. We display these results in Table 5.

The results reveal that experimentation on digital twins can uncover unbiased treatment effects even in scenarios where digital twins struggle to predict workplace preferences. For example, the digital twins accurately recover the treatment effect from the human sample when examining how hiring algorithms impact the likelihood of applying, being hired, and receiving timely information during the hiring process. These measures were where the digital twins failed to accurately predict human preferences at baseline.

In other cases, however, digital twins fail to recover the true treatment effect. The largest differences here are for beliefs regarding sociality at work and the ability of the firm to identify top talent. In these scenarios, digital twins substantially under-estimate the extent to which hiring algorithms depress human beliefs regarding the firm's hiring policy. We find similar results for clear communication, information regarding company culture, and whether hiring is fair and unbiased. This suggests that overall, digital twins are less adverse to algorithms compared to humans.

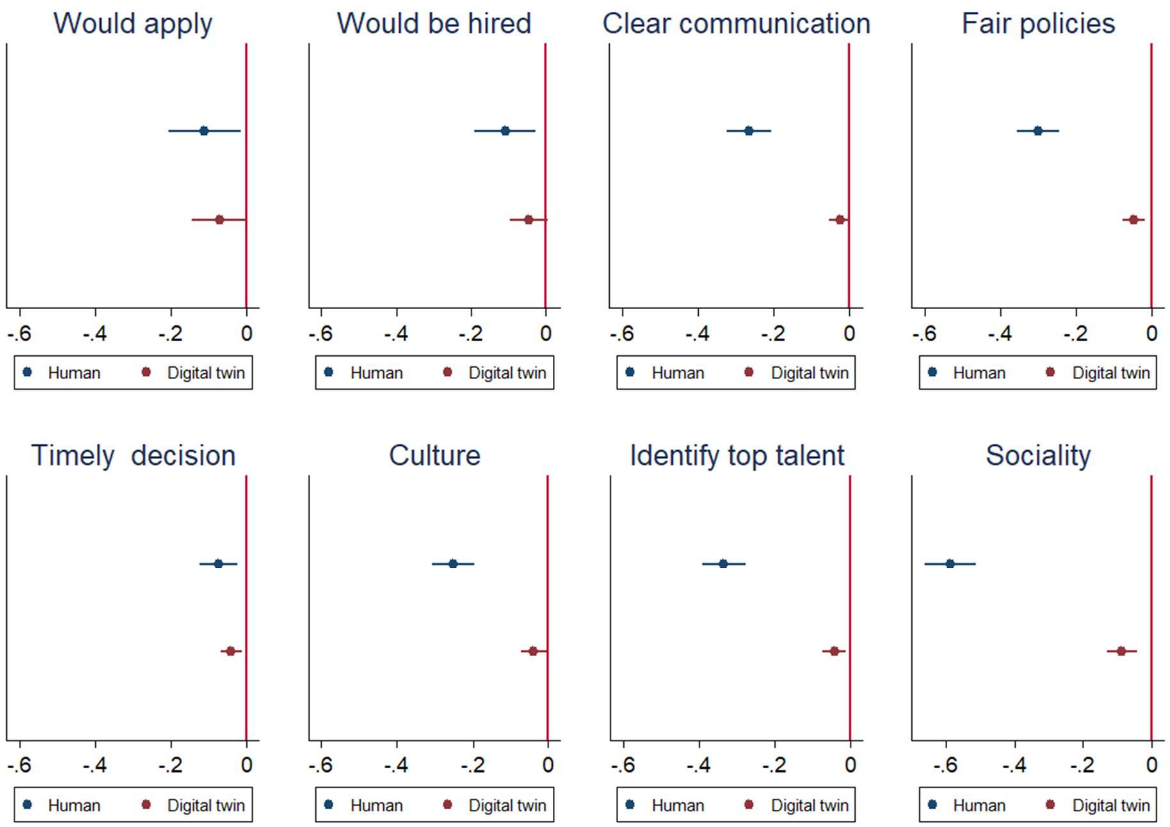


Figure 3: Impact of Hiring Algorithms on Job Candidate Beliefs, For Humans Vs Their Digital Twins

	Outcomes							
	Would Apply (1)	Would Be Hired (2)	Clear Communication (3)	Fair Policies (4)	Timely Decision (5)	Clear Expectations (6)	Identify Talent (7)	High Sociality (8)
AI	-0.11**	-0.11***	-0.27***	-0.30***	-0.07***	-0.25***	-0.33***	-0.59***
	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Digital Twin	0.13***	0.73***	-0.22***	-0.42***	-0.47***	-0.61***	-0.60***	-0.81***
	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
AI * Digital Twin	0.04	0.06	0.24***	0.26***	0.03	0.22***	0.29***	0.50***

	(0.07)	(0.06)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
N	7992	7992	7992	7992	7992	7992	7992	7992

Table 5: Comparison of within-participant effects using stacked regressions, by outcome