ESSAY

A simulated customer is never undecided.

By Mike Millett with Digilu

October 6, 2026 · 9 minute read

Pew built AI twins of its own panelists and they missed by 12 points on average. The quieter failure matters more to marketers: the twins said 'not sure' about a quarter as often as the people they copied.

A language model is given a description of a person, asked to answer as that person, and the answers are treated as a preview of what the market thinks. Pew Research Center tested that idea on its own panel and reported an accuracy problem. The more consequential result is a different one: the simulated respondents rarely said they were not sure. A simulated customer is never undecided, and the uncertainty it erases is the signal marketing needs to adapt to a decision still in motion.

Pew's AI digital twins missed real panelists by 12 points on average

On September 30, 2026, Pew Research Center published "Can AI Stand In for Human Survey-Takers? Not Really." The report tested whether AI stand-ins for its own survey panelists could reproduce what those panelists said. Pew used what it calls a digital twins approach: an AI model was asked to adopt the personas of real members of its American Trends Panel, using what Pew already knew about each person, and then to answer the same surveys those people had answered.

Across nearly 300 survey questions, the estimates from the AI respondents differed from the human results by an average of 12 percentage points. On around 28% of the questions the average difference exceeded 15 points. Some misses were far larger. The AI poll indicated that 97% of Hispanic adults were at least somewhat likely to follow the World Cup, where the measured share was 43%. Pew's description of the pattern is that the model took attitudes that are reasonably common in a group and portrayed them as nearly universal.

Two further findings in Pew's study matter more for anyone studying customer behavior. Nearly half the questions had at least one answer choice that no AI respondent selected at all. And where a 'not sure' option was offered, the human panelists were around four times as likely as the AI model to choose it.

The missing 'not sure' deletes the people a campaign exists to move

A 12 point error and a missing 'not sure' are different kinds of failure in synthetic research. The 12 point error puts a simulated answer in the wrong place on a scale that still exists. The missing 'not sure' removes a condition from the data: the simulation reports a settled person where the real person was unsettled, and no adjustment to the percentages restores a state the output never contained.

Pew's knowledge questions show what that removal costs. Pew wrote that its AI respondents "didn't like to admit to uncertainty," and the clearest case is factual. Asked what the First Amendment guarantees, 52% of U.S. adults chose the correct answer; the synthetic respondents put the figure at 98%. On the central focus of NATO, the numbers were 56% and 99%. Across questions that fewer than six in ten Americans answer correctly, the AI poll estimated that nearly every member of the public knew the answer.

Any effort whose job is to change those answers, a civics program or a public information campaign, has one addressable audience: the people who do not yet know. In Pew's measured public, that is nearly half the country on the First Amendment question. In the simulation, it is almost nobody. A plan built on the synthetic poll would conclude the work was finished before it began, and it would be wrong by exactly the population the work was for.

A purchase decision has the same structure. The customers a message can still move are the ones whose answer is not yet settled, and that is the group a simulation reports as empty. In the Marketing Helix, alignment between a customer and a brand is temporary, and Pew's unsure and unknowing respondents are the measurable form of that condition: people on whom trust, relevance and timing are still acting. A research method that reports every position as closed has measured the end of a decision and labeled it the whole decision.

Synthetic research error should be largest where a decision is still open

The Marketing Helix predicts that synthetic research will fail unevenly, and the prediction can be tested. A funnel model treats a customer as holding a position, so a table of positions is all it asks research to supply, and it gives no reason to expect synthetic output to fail in one place more than another. If alignment is temporary, the error should be smallest where alignment is stable, such as a habitual repurchase, and largest where a decision is long and still open, such as a first purchase in a considered category. Nobody has published that comparison. A company with both kinds of customer can run it.

Timing, in the Helix, is the customer's readiness state, not the brand's calendar. The questions that locate readiness are where confidence is forming, where it is eroding, and what changed since the last observation. A respondent that is never uncertain cannot register any of the three.

A second benchmark shows the same overconfidence from another angle. In "When Synthetic Users Fail: A Cross-Domain Benchmark of LLM-Simulated Human Survey Responses," a July 2026 arXiv preprint, Zihan Chen, Di Zhu and Lei Nico Zheng tested four models against the General Social Survey and the World Values Survey. The models treated identity as far more predictive of attitudes than it is among real people, and on a segment targeting task they inflated the gaps between segments two to fourfold. Sharper segments and fewer undecided people describe the same output: the simulation is more certain about customers than customers are about themselves.

A payroll campaign planned on synthetic segments lands before readiness

A team that plans a campaign against synthetic segments commits budget to a decision its customers have not reached. The payroll case below is an illustration, not a reported one.

A company selling payroll software to small firms asks a synthetic panel of simulated office managers what would make them switch providers. The panel returns a ranking: price first, integrations second, support third. Every simulated manager has a preference, so the table has no empty cells. The team commits the quarter's budget to a price message on a fixed twelve week schedule.

The real office managers are in a different state. Many have not decided whether to switch at all this year. Their open question is whether, not which, and what will settle it is a date the vendor does not control, such as a year end filing that goes badly or a renewal notice with a higher number on it. The price message reaches people who have not yet arrived at the question it answers. It is accurate, it is relevant to the eventual choice, and it lands before readiness.

The cost is larger than one quarter's budget. When the campaign underperforms, the team has no explanation, because its research said the segment was settled and price was the lever. The likely next move is a deeper discount, which answers the same question louder.

A real panel with a 'have not decided' option would have shown how many managers were still open. The follow-up question, what would settle this for you, is the instrument that finds the turning point: a manager who answers 'when our renewal notice arrives' has named the week a price message becomes relevant, and a manager who answers 'if year end filing goes badly again' has named the month after it. Neither date exists in a synthetic table, because no simulated manager is waiting for anything.

The strongest objection: a survey 'not sure' is not an undecided buyer

A survey 'not sure' is not the same thing as an undecided buyer, and the strongest case against this essay starts there. Pew measured public opinion on politics, civic knowledge and current events. It did not study purchase decisions, and it did not measure how long a buyer stays undecided. Survey researchers have long known that some 'not sure' answers reflect low effort, not real ambivalence. A

model can also be instructed to express uncertainty more often, and Pew tested a specific setup on a specific model. Vendors of synthetic research will say, fairly, that their tools are meant for early screening of concepts, not for replacing fieldwork.

Each point narrows the claim without removing it. The evidence is one study from a polling organization and two academic studies in neighboring domains. No one has published a comparable test on buyers in the middle of a real decision, and until someone does, the transfer from polling to purchasing is an argument, not a finding.

The part of the argument that survives the objection concerns what a tuned model can deliver. Suppose a vendor adjusts its model until simulated respondents say 'not sure' at exactly the human rate. The total now matches. The assignment does not: the uncertainty has been distributed by quota, with no connection to which actual people are unsure or what would settle the question for them. A matched total is weak evidence on its own. In "Questioning the Survey Responses of Large Language Models," presented at NeurIPS 2024, Ricardo Dominguez-Olmedo, Moritz Hardt and Celestine Mendler-Dünner found that once answer order was randomized, the 43 models they tested trended toward uniformly random responses, and appeared to represent best the subgroups whose statistics were closest to uniform. A marketer does not act on the total. A marketer acts on who is open and why. Calibrated uncertainty and observed uncertainty are different measurements, and only the second one says anything about timing.

Adaptive marketing reads the 'not sure' share as a result

Adaptive marketing treats the 'not sure' share as a measurement, not as missing data. If synthetic error is largest where decisions are open, that share is the one number a simulation cannot supply, and the uneven-error prediction implies three readings a team can check against its own sales record.

A falling 'not sure' share with flat sales should mean decisions are closing elsewhere: toward a competitor or toward doing nothing. The test is to ask the newly decided what settled it before adding spend to the current message.

A rising share among prospects should mean something reopened the category, such as a new entrant or a price change, and the window for a well timed message has widened. The same rise among existing customers should be an early reading of decaying alignment, visible before it reaches renewals.

A falling share with rising sales should mean alignment is forming, and the useful record is which message reached those customers shortly before they decided.

On Monday, a team can make four changes. Add an explicit 'have not decided' option to every preference question in customer surveys and sales intake forms, and stop forcing a choice. Follow it with one question: what would settle this for you? Ask the same people again in thirty days. Report the 'not sure' share beside the preference share in every plan, so a schedule is never set against a segment whose decision is still open.

Simulated customers retain a narrower use: stress testing the wording of a questionnaire, surfacing objections a team had not considered, and generating hypotheses to take to real people. Pew's bottom line was that AI polling is not a replacement for rigorously surveying real humans.

> A research method that reports every position as closed has measured the end of a decision and labeled it the whole decision.

A model can be tuned to say 'not sure' at the human rate. It cannot say which customer is unsure or what that customer is waiting for, and a schedule built without those two answers is set against a decision that has not yet happened.

Further reading

1. Pew Research Center, the primary study: AI digital twins of American Trends Panel members answered nearly 300 survey questions, with the average error, the share of large misses and the 'not sure' gap cited here pewresearch.org/data-labs/2026/09/30/can-ai-stand-in-for-human-survey-takers-not-really
2. Chen, Zhu and Zheng, a cross-domain benchmark preprint showing models over-determine demographics and inflate gaps between segments, cited as independent support for simulated respondents being more certain than real people arxiv.org/abs/2607.26348
3. Dominguez-Olmedo, Hardt and Mendler-Dünner, an earlier study of language model survey responses, cited in the objection section on the difference between matching a distribution and representing a respondent arxiv.org/abs/2306.07951

marketing helix

marketinghelix.com

A customer behavior model describing how a customer in motion gives a message attention, and how trust, relevance, and timing then shape whether it draws closer, drifts away, or returns.