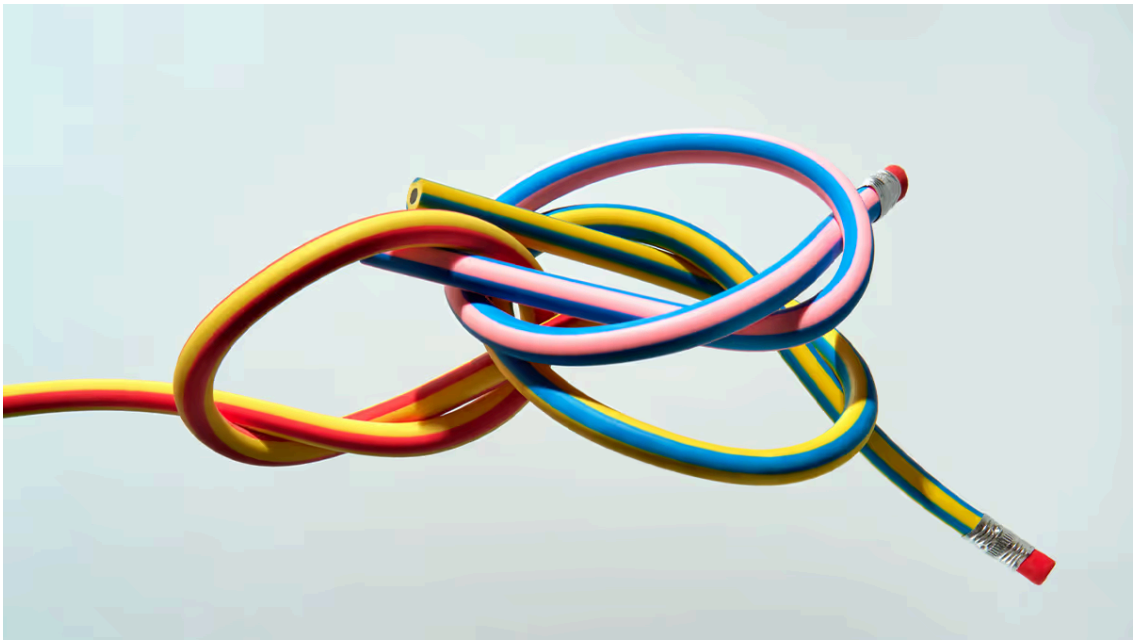
HARVARD BUSINESS REVIEW

AI And Machine Learning

Research: The Innovation Problems AI Can’t Solve

by Julian De Freitas, Ayelet Israeli, Gideon Nave, Artem Timoshenko and Olivier Toubia

August 14, 2026



Tatyana Lavrova/Stocksy

Summary. Generative AI can accelerate innovation, but the article argues that its default use often reinforces the human bottlenecks it is meant to solve. In ideation, models steer teams toward familiar ideas and make people more fixated on them. In screening,... **more**

Every innovation team now has the same tools: the same foundation models, similar prompt libraries. Yet the results are wildly uneven. Some teams report a creative renaissance while others report a flood of homogenized, forgettable ideas that all sound like they came from the same person. The reason isn’t

which model you're using. It's that generative AI acts on the *human bottlenecks* buried inside that process. These bottlenecks respond to AI in different—sometimes opposite—ways.

That's the core argument of a new working paper, invited by the International Journal of Research in Marketing, that we—researchers at Harvard, Wharton, Northwestern, and Columbia—have written. Instead of asking what generative AI can do, we ask a more durable question: What are the cognitive, social, and organizational constraints that have always slowed innovation down, and how does AI interact with each one?

To answer this, we synthesized recent empirical studies on generative AI and innovation—on idea diversity, AI-simulated consumers, algorithmic screening of ideas, and post-launch analysis—against decades of established research on creativity, judgment under uncertainty, and consumer behavior. For each of the four stages below, we asked the same three questions: What is the underlying human constraint, how does generative AI affect it when used "naively," and what would a more deliberate redesign look like? That consistent structure allowed us to apply the same diagnostic question to bottlenecks that look nothing alike on the surface, such as a screening committee's bias against novelty, or a consumer's inability to say what they want before they've actually experienced it. Is AI dissolving this bottleneck or quietly making it worse?

The answer matters more than innovation managers currently appreciate. Some of these constraints, it turns out, generative AI can genuinely dissolve. Others, it makes worse while looking like

it's helping. And a few, no plausible future model is going to touch.

The Problem with Assuming AI Will Fix Innovation

Most conversations about AI and innovation treat AI as uniformly good at removing friction in the innovation process. Usually, the assumption is it will achieve this by completely displacing the roles of human innovators: more ideas, faster testing, cheaper research, quicker learning. Our framework pushes back on that assumption. We trace bottlenecks through four stages of the innovation pipeline—ideation, screening, consumer insight, and market learning—and find that at nearly every stage, AI's default effect (used "naively," without deliberate redesign) is to *deepen* the underlying bottleneck, not shrink it.

Here's why: Most innovation bottlenecks are human problems, not technology ones. They're issues of how people get anchored to familiar ideas, how panels develop tribal instincts about what counts as "promising," how consumers can't tell you what they want until they've seen it. Generative AI, trained on the aggregate of existing human output, tends to reproduce exactly the tendencies that caused the bottleneck in the first place; just faster and at scale.

Here are the ways that generative AI *deepens* human bottlenecks at each stage of the innovation process—and what to do about it.

Ideation: Over-fixating on the typical. Ideation is the process of generating candidate concepts for a new product, service, or feature. Because a single exceptional idea is worth more than dozens of merely good ones, the task isn't to produce a large pile of solid options—it's to surface genuine outliers, which makes ideation fundamentally a search for exceptions rather than an exercise in efficient production.

When people brainstorm with an LLM without a deliberate structure, two things happen simultaneously. The model itself gravitates toward the statistically typical response—after all, it was trained to autocomplete phrases by using the most likely next words. And once a person reads that typical response, they become fixated on it, preventing them from thinking of the more unusual ideas they might have generated unprompted. The result is a subtle, two-step narrowing: the AI presents a narrow idea, then this also narrows the human's thinking.

Worse, this happens at a systemic level. Because most generative models draw from similar training data, independent teams at different companies using AI "creatively" tend to converge on the same handful of design directions. So, even if individual productivity goes up, collective diversity across a market goes down.

In this case, the fix is on the model, not the human. Using chain-of-thought prompting—asking the model to draft, then explicitly revise toward bolder, more distinct territory—works remarkably well on the AI. However, the same approach does not help a *person* improve their own thinking. Once a person is mentally fixated on the AI's initial idea, telling them to "think more broadly" typically does not help—if anything, they may become even more fixated on the AI's idea (this is a version of the classic "don't think of a white bear" problem).

Screening: Biasing toward more "polished" LLM ideas. Screening is where a firm decides which of those ideas get funded and which get shelved, usually via a stage-gate committee or portfolio review board. It's judgment made under real uncertainty—before the market has spoken, and often before the idea itself is fully specified—so whatever the panel is consciously or subconsciously optimizing for ends up shaping everything the company eventually builds.

Evaluators have always been biased against novel ideas—original concepts feel riskier and harder to place, so panels unconsciously reward safer, more familiar submissions. Generative AI supercharges this bias in a new way: AI-generated pitches are fluent and well-structured by design, and fluency gets mistaken for quality. A polished-sounding idea gets a better score, independent of whether it's actually the better idea. Here the fix is structural, not technological: strip submissions to a common format before anyone scores them, and hide whether an idea came from a human or a model.

That's not the only bottleneck generative AI can quietly deepen at the screening stage. Here's a subtler trap: if you train a screening LLM on your own panel's historical accept/reject decisions, you're not removing bias from the process—you're encoding it permanently, at scale, and making it *look* objective. A recent field experiment found something even more troubling: when evaluators were given an LLM recommendation plus a written rationale, they became more likely to rubber-stamp *rejections*—without getting any better at telling correct AI judgments from incorrect ones. The LLM's rationale substituted for their own judgment rather than supporting it. And the effect was strongest exactly where human judgment matters most: borderline cases.

The intervention here is almost the opposite of what most companies are building toward: strip the LLM explanations out.

In that same experiment, a "black box" AI recommendation without a rationale improved decisions, as compared to keeping the LLM rationales in the mix.

Consumer Insight: Faster research, less authentic answers. Once an idea survives screening, it gets developed into a concrete concept with real features, pricing, and positioning—and testing that concept means understanding what customers actually want. This is the preference-measurement stage: the surveys, interviews, and conjoint studies firms have long relied on to turn a concept into a defensible read on demand.

This is where the paper is most sobering for anyone using AI to simulate customers. As an example, if you ask an LLM to role-play a thousand different consumer segments reacting to a price change, you'll get a smooth, downward-sloping curve that's mostly wrong in a specific, problematic way. Because the model wasn't drawn from a real population, it doesn't treat a price variation as an experiment; it treats it as a *clue* about context (a higher price reads as "premium," not "the same product, just more expensive"). So instead of the downward-sloping demand curve you'd get from actual shoppers as the price goes up, you can get a flat or even upward-sloping one.

"Digital twins"—AI personas built from rich behavioral profiles of real people—fare better, but they are still currently missing something essential: real customers behave *irrationally* in ways that matter enormously for whether they'll adopt something new. They cling to sunk costs. They overweight the pain of switching. Simulated twins, by contrast, tend to make the rational choice far more often than actual humans do—which means they systematically underestimate the very frictions that kill new product launches.

There's no clean fix for this one yet. Digital twins aren't useless—they're a reasonable stand-in for some research questions—but they shouldn't be trusted on precisely the behaviors that separate a genuinely new product's fate from an incremental one. For any decision that hinges on sunk costs, switching costs, or how a choice is framed, the safer move is to keep that judgment with human research: lead-user panels and ethnographic fieldwork that watch how people actually behave, not just what a model predicts they'll choose.

Diffusion & Market Learning: Same confirmation bias, with a more persuasive voice. Diffusion and market learning covers everything that happens after launch: getting an innovation adopted at scale, then reading the resulting signals to inform the next round of decisions. Adoption is a social process as much as an individual one—driven by early adopters and word of mouth—and the feedback it produces has to be collected, prioritized, and interpreted correctly before it's useful.

Once a product ships, it throws off more feedback than any team can absorb: reviews across platforms, social chatter, support tickets, returns, forum threads. Generative AI is genuinely useful here: it can ingest, cluster, and summarize all of that noise at a scale no analyst team could match, extracting structured customer needs with a quality that starts to approach professional analysts. For the first time in this framework, a bottleneck actually shrinks.

But the relief creates a new problem in its place. Out of the hundreds of themes an AI system surfaces each week, which ones deserve action, and which are just the loudest? The signals that matter most for innovation, such as an emerging use case or a segment quietly failing to adopt, are exactly the ones a frequency-weighted summary is built to bury.

The deeper problem shows up once that signal reaches a room full of people with a stake in the outcome. Organizations have always read market feedback charitably toward decisions they’ve already made. An AI system has no career riding on last year’s product bet, so in principle it should flag a wave of complaints about a champion’s feature exactly as readily as a wave of praise. But the same fluent, confident summary that makes AI synthesis useful also makes it easy to selectively cite as a voice in support of whatever the team already believed going in—especially if the AI is queried in a way that makes it clear what answer the team would prefer, given AI’s tendency to sycophantically say what people want. Whether AI synthesis nets out to more honest reporting, or just a more convincing way to dress up the same rationalizations, is one of the most important open questions in the entire framework.

The Major Takeaways

Look closely at these four examples and a single principle emerges: the mechanism behind a bottleneck determines whether AI helps or hurts it—not the sophistication of the AI. Bottlenecks that are essentially informational, high-volume, or procedural (too many submissions to review, too much review text to read, a rigid format that hides quality) are exactly the kind of problem generative AI is built to solve, provided you design the system deliberately.

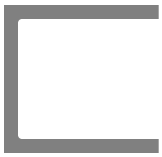
In contrast, bottlenecks rooted in lived experience, irrational human behavior, or organizational self-interest are a different animal entirely. No amount of model scaling fixes the fact that consumers can’t articulate a preference for a product category they’ve never seen, or that a manager with two years of political capital invested in a failing project will read the data charitably no matter how good the data is.

This gives leaders something more useful than a list of AI use cases: a diagnostic question to ask before adopting any new tool or workflow. What, specifically, is slowing this down: a problem of information, a problem of human judgment, or a problem of organizational incentives? This alternative perspective yields clear lessons at each ideation stage:

- **Ideation**: Left unchecked, generative AI narrows creative thinking instead of expanding it. The fix has to target the model's prompting, not the person's mindset.
- **Screening**: AI doesn't remove a screening panel's bias against novelty. It either flatters that bias with polish or quietly encodes and scales it, unless the process is deliberately redesigned.
- **Consumer insight**: Simulated customers are a fast, useful filter, but they underestimate the irrational frictions that actually decide whether something new gets adopted.
- **Diffusion and marketing learning:** AI solves the volume problem in post-launch feedback, but it can just as easily hand teams a more convincing way to defend a decision they'd already made.

...

The paper's most unsettling argument isn't about any single bottleneck—it's about what happens to organizations over time. As AI absorbs more of ideation, screening, and consumer simulation, the junior employees who used to build judgment by doing that work firsthand no longer get the reps. You end up with sophisticated *users* of AI output who never developed the instinct to know when that output is wrong.



Compound that across an entire innovation function, and you get what is called the automation trap: every individual AI tool tests fine in isolation, but the whole feedback loop slowly drifts away from real customers, because ideation, screening, testing, and interviewing are all now mediated by models trained on each other’s outputs rather than on the market. Everything gets faster. Nothing stays grounded.

Our answer isn’t to slow down AI adoption, but to be deliberate about which parts of the process still require unmediated contact with reality. Lead-user immersion. Ethnographic fieldwork. A human who owns the consequential call and has to stand behind it. These aren’t nostalgic holdovers from a pre-AI era; they’re the load-bearing walls that keep the rest of the automated structure from drifting off the ground truth it’s supposed to be tracking.

The technology will keep changing. The bottlenecks—cognitive fixation, status-driven evaluation, the gap between what consumers say and what they’ll actually do—have been remarkably stable for decades. Betting your innovation strategy on the mechanism rather than the model is likely to age a lot better.

Julian De Freitas is an assistant professor in the marketing unit at Harvard Business School.

Ayelet Israeli is the Marvin Bower Associate Professor of Business Administration at Harvard Business School. She is also the

cofounder of the Customer Intelligence Lab at the Digital Data Design Institute at Harvard Business School.

Gideon Nave is Carlos and Rosa de la Cruz Associate Professor, Marketing Department, Wharton School of Business.

Artem Timoshenko is an assistant professor of marketing at the Kellogg School of Management, at Northwestern University. His research focuses on applications of AI to marketing analytics and customer insights.

Olivier Toubia is a professor of marketing at Columbia Business School and a recognized leader in the field of quantitative marketing.





Read more on **AI and machine learning** or related topics **Generative AI**, **Innovation** and **Disruptive innovation**