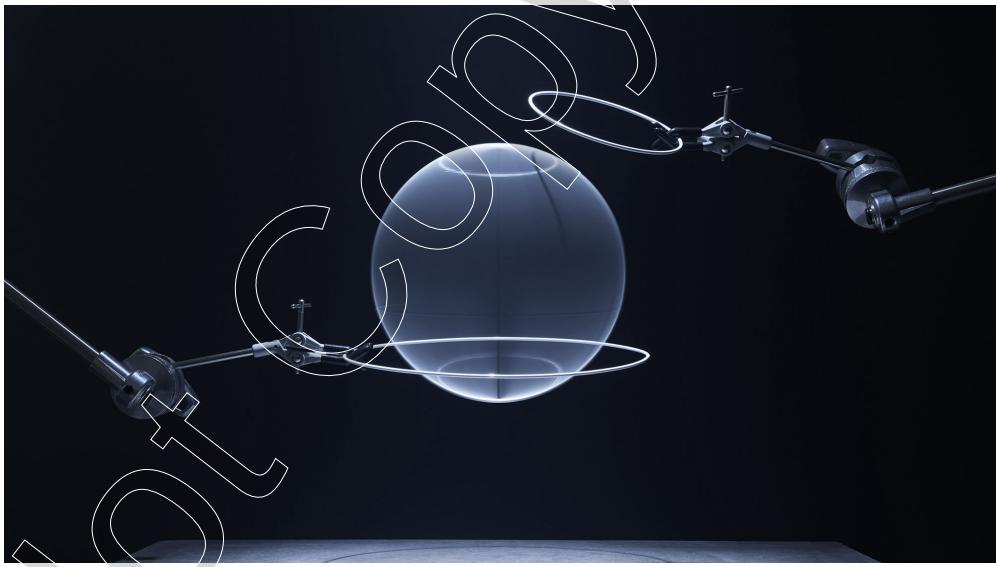
HARVARD BUSINESS REVIEW

Magazine Article / Generative AI

A Systematic Approach to Experimenting with Gen AI

To reduce risk, refine their strategies, and optimize adoption at scale, companies need more testing at the organizational level. *by Johannes Berndt, Florian Englmaier, Raffaella Sadun, Jorge Tamayo, and Nikolaus von Hesler*

From the Magazine (upcoming) / Reprint R2601W



Greg White

After taking the software industry by storm, generative AI is now moving into a broad set of industries, including manufacturing, where it is helping manage unpredictability and support real-time decision-making. Gen AI's ability to codify, automate, and distribute organizational expertise may eventually reshape work structures from the shop floor to the C-suite. Already some companies are using it to

analyze the flood of information generated in factories and to predict problems, simulate complex scenarios, and optimize processes in real time. By working with a wide range of manufacturing industry data—from maintenance manuals and machine automation code to complex diagrams, 3D drawings, and process data—gen AI has the potential to establish new ways for people and machines to collaborate.

But who will benefit from these changes, and how quickly? That's not a simple question. Like electricity and the printing press, gen AI is a general-purpose technology—the adoption of which, history teaches us, is rarely straightforward. Managers often fail to recognize the true economic potential of new technologies and struggle to reorganize tasks, skills, and workflows to suit them. As a result, performance gains typically lag behind technological diffusion, giving rise to what's known as the "productivity J curve": an initial dip in productivity as organizations adapt to a new technology, followed by sustained gains once complementary investments pay off. Recent data on gen AI is consistent with that pattern: A 2025 McKinsey survey, for instance, found that although many firms had rapidly adopted gen AI, more than 80% reported that it had had no significant impact on earnings yet.

Because it's not clear how firms will adopt gen AI, managers face a strategic dilemma: Wait for more clarity and risk falling behind? Or act too soon and invest in applications that don't deliver?

To address this tension, leaders need to think about gen AI adoption not as a single decision but as a portfolio of organizational experiments. Like A/B testing in digital-product development, these experiments should aim to isolate causal effects—focusing not just on whether gen AI works but also on how it works, for whom, and under what conditions. By testing gen AI applications before scaling them up, managers can reduce risk, refine their strategies, and build internal

momentum for change. Experts have been advocating for this approach, but many firms are struggling to implement it. Experimentation therefore remains a relatively novel practice in many organizations.

That needs to change. Experimentation allows companies to transform gen AI uncertainty into a strategic advantage. It helps firms move through their own adoption phase more successfully than their competitors do. And the knowledge generated through experimentation can be leveraged to reinforce existing relationships—or create new ones—within their ecosystems. In this article we'll describe how firms are getting better at adopting gen AI through experimentation—within their own organizations and across entire ecosystems. Software organizations have been pioneers in this work, but some companies, such as Siemens, are starting to carry it out successfully in the physical world of production.

The Adoption Challenge

Although the promise of gen AI is great, many organizations have yet to fully embrace it. The fact that gen AI tools generate hallucinations and unreliable results is one reason firms have balked at using them in high-stakes settings. But a deeper reason, various experts say, is that gen AI's true economic potential lies in creating entirely new systems of value, which are hard for organizations to recognize, let alone pursue. For a historical parallel, think about electricity: Manufacturing plants took almost 40 years to adapt to the technology and optimize themselves around it, according to the late economic historian Paul David.

Embedding gen AI technology at the organizational level will require companies to carefully work out how to integrate it with existing processes, routines, and teams. In manufacturing, the challenge will be even greater, because in that domain the need for performance,

reliability, security, and smooth integration with human workers is so strong.

Seen in this light, the slow rate of successful adoption is not surprising. It reflects the larger challenge of making gen AI organizationally useful, not just technically impressive. That's where organizational experiments can help.

Engines of Learning and Adaptation

At its core an organizational experiment is an application of the scientific method. In a real-work setting, it establishes a treatment group (for example, employees or teams using a new AI system) and a control group (those operating as usual, without the new system). The experiment is based on a specified research design (starting with a clear and testable hypothesis) and may run for an extended period (weeks or months) to capture both initial and sustained effects. Data is collected on key performance metrics and sometimes supplemented with qualitative feedback from participants. Random assignment or other controls are used to ensure comparability across the groups in the experiment.

To isolate an AI tool's effect on performance, a company might enable it for only half its workers at random. GitHub and Google, for example, conducted controlled trials in which developers were randomly assigned to do their coding manually or with AI assistants. Those who used the AI assistants completed their coding tasks 21% to 55% faster and had slightly higher completion rates than those who coded manually. They also reported feeling greater job satisfaction and reduced mental strain. The results— which show that AI assistants can speed up development and improve worker well-being—suggest that AI not only increases employees' performance but also can improve their job experience.

When randomization is difficult, some firms implement staggered rollouts, phasing an intervention in to different teams over time to create natural control groups. In one experiment a *Fortune* 500 company that specializes in business-process software for small and midsize companies in the United States staggered the rollout of a gen AI assistant to more than 5,000 customer-support agents and compared the performance of those who had access to the tool with that of those who didn't. It found that productivity rose by about 14% overall for those using the tool, with a 34% increase for less-experienced agents. Customer-sentiment scores also rose, as did the customer-retention rate. Those results led the company to scale the tool up across the organization.

Photographer Greg White uses in-camera techniques to playfully document the principles of physics in his series *Base Quantities*.

Another approach is to create a "lab in the field"—that is, a controlled environment where interactions with the new technology can be observed. For example, in a recent trial at Procter & Gamble, 776 product

developers were randomly assigned to work either with or without AI, and either solo or in pairs, during an innovation hackathon. (One of us— Raffaella—was involved in running the trial.) On average, solo AI users performed as well as teams that weren't using AI did, and both the individuals and the teams that used AI were better at blending technical and commercial ideas. P&G concluded that using gen AI could reduce siloed thinking—and might enable the creation of smaller cross-functional teams.

Although organizational experiments share some traits with traditional tech pilots and A/B tests, there are some fundamental differences. Pilots typically are informal tests that involve handpicked teams and anecdotal feedback; scaling decisions for them are often based on enthusiasm rather than evidence. Pilots lack clear hypotheses and control groups and therefore have only a limited ability to produce generalizable insights. A/B tests, for their part, work well for fine-tuning —choosing the digital features of a new product, for instance—but rarely capture the broad effects of a change on coordination, workflow, or people's experience. Organizational experiments go further than both pilots and A/B tests: They evaluate real-world impact and reveal whether gen AI works as well as how, for whom, and under what conditions. They are engines of strategic learning and adaptation.

When conducted properly, organizational gen-AI experiments can produce a host of benefits, including the following:

Causal insights. Experiments help distinguish correlation from causation, which is crucial. Without an experimental design that makes this distinction clear, organizations can't determine whether productivity gains come from gen AI or, say, from early adopters of the technology, who are often more skilled or motivated than average.

Granularity. Experiments can reveal how gen AI affects different types of workers or units differently. That's vital because gen AI's effectiveness depends heavily on context—on the specific task in question, a user's skill level, workflow integration, and an organization's culture, to name a few factors. What works brilliantly for one team may fail spectacularly for another. Recent evidence, for example, suggests that in customer service, at least, gen AI copilots can provide large benefits for less-experienced workers but almost undetectable ones for more-experienced workers. This kind of evidence gives managers valuable insights into not only the impact of gen AI in their organization but also the investments needed to realize that impact—for example, by deploying new tools where they're likely to produce the greatest improvements.

Risk reduction. Experiments help managers spot possible implementation hurdles before they undertake a full-scale rollout. In *The Voltage Effect*, the economist John A. List expands on this idea, noting that experimentation can achieve multiple benefits. For instance, it can help you *avoid false positives* (ensure that initial positive results aren't just a fluke), *understand your audience* (avoid the risk that an idea that works with one specific, highly motivated group may not work with a broader, more diverse audience), *assess the scalability of ingredients* (make sure that an idea's success is not dependent on one unique person, such as a celebrity, but on a replicable process or product), *consider unintended consequences* (avoid unforeseen effects of scaling that negatively impact the original idea), and *manage costs* (evaluate whether the costs of an idea will remain sustainable as it grows). The last is especially relevant for gen AI, because its adoption requires nontrivial investments in technology, people, and organizational processes.

Strategic learning. Managers can overcome the decision-making paralysis often associated with uncertainty by launching a process of discovery that consists of small experiments based on testable hypotheses. Designing experiments forces managers to focus on strategic questions and develop a framework for thinking through problems in a structured way. When Siemens, for example, began its organizational experimentation, it first defined the hypotheses to be tested to capture the impact of gen AI on workers' productivity and well-being. Focusing on specific hypotheses led to a much clearer definition of the measures to be collected during the experiment to capture changes in behavior (for example, the time needed to solve a problem), changes in attitudes (for example, level of job satisfaction) and, ultimately, productivity outcomes on the shop floor. Adhering to the scientific method also provided clarity on primary and secondary effects of gen AI in manufacturing, such as whether gen AI would reduce workers' reliance on expert engineers. Overall, this process helped Siemens understand the path from technology adoption to value creation in a much richer and more precise way than it would have with a generic product rollout.

Ecosystem Experimentation

Gen AI experimentation does not simply benefit potential adopters. Innovators can see even greater returns: They can apply the insights gleaned to help prospective buyers understand which gen AI use cases really matter for them or which challenges might prevent them from integrating the technology into existing processes. Some innovators with large user bases are leading the experimentation surrounding new gen AI applications outside their own organizations—for example, in partnership with current or prospective buyers. In these cases, innovators orchestrate ecosystem experimentation.

Microsoft, for example, collaborated with a team of academics to study the adoption of Microsoft Copilot by more than 7,000 employees across 66 firms. The team ran a structured experiment that gave Copilot licenses to a targeted group of employees and then tracked how their email and meeting behaviors changed, comparing them with employees in similar roles who didn't have access to Copilot.

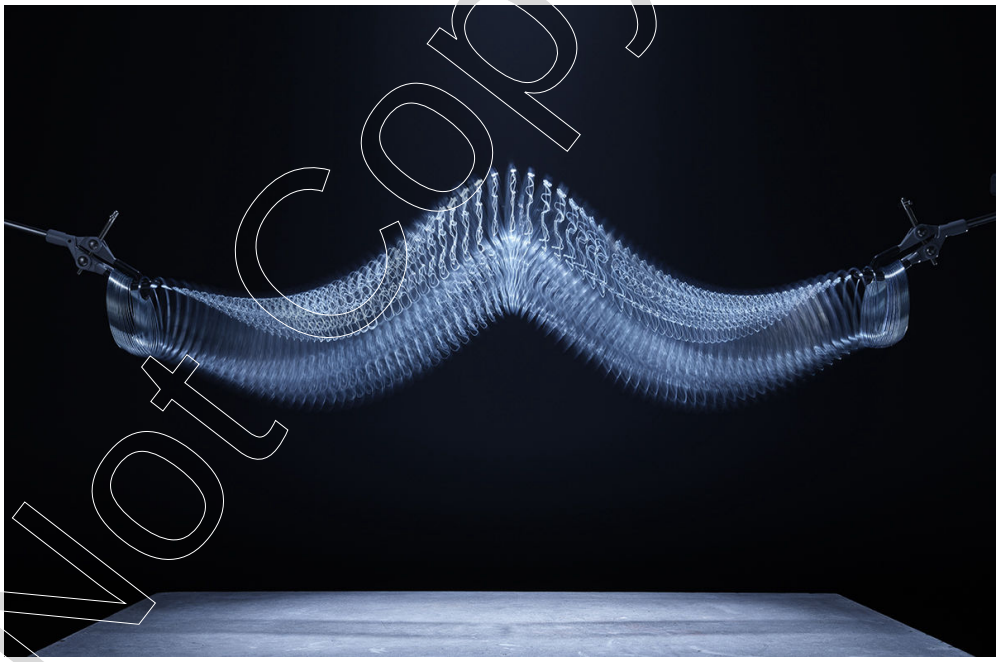
The research found that Copilot users spent 1.3 to 3.6 fewer hours per week on email and drafted documents faster, but they didn't change their meeting behavior. The study also found that offering training and establishing a change management program are key to adoption. Because the quality and scale of the experiment were robust, these findings are likely to shape how Microsoft moves forward with Copilot.

In a similar vein the software platform Grab is currently collaborating with a team of academics from Harvard Business School and INSEAD to examine the impact of an AI assistant on more than one million entrepreneurs in six countries. The scale of the experiment allows Grab to see with precision which tasks gen AI helps with most and how different kinds of businesses on the company's platform actually use it. This data enables Grab's product developers to keep improving how they design, deploy, and experiment with AI.

AI on the Factory Floor

Software companies aren't the only ones that can benefit from ecosystem experimentation with AI. By collaborating with several customers, Siemens has created what it calls a "generative-AI-powered assistant for the shop floor," which helps shop floor workers with the maintenance and repair of industrial machines. The gen AI assistant provides instant access to machine information embedded in static documents and live machine data through an intuitive chat interface. Siemens released an early version of the tool to targeted users to

help them understand how to integrate it into their daily work. The customers, who include everybody from machines' builders to their end users, have tested the minimum viable product across varied operational contexts in a set of exploratory experiments designed to surface technical, organizational, and commercial insights. The experiments help Siemens improve and refine its product (for example, it benchmarked the quality of responses to different questions to find the most effective prompting styles and to determine which areas in the handbooks needed improvement). Customers have embraced the approach, seeing it as an opportunity to start trying the tool out now and to prepare for what a more-powerful version of it might do in the future.



Greg White

Siemens ran the first test of its gen AI shop-floor assistant in 2024 at its experimental factory in Erlangen, Germany. (All the authors of this article were involved with the test.) Maintenance technicians

were instructed to use the tool during complex repairs of expensive machines. These repairs usually involve multiple steps to identify and replace worn or broken parts. The experiments tested whether the shop-floor assistant could streamline the process by providing step-by-step analysis and repair instructions to workers in the flow of work. Early results—based on structured before-and-after surveys matched with granular performance data—show that the shop-floor assistant cut the time needed to find information and helped workers handle tasks more independently.

Experimentation taught Siemens the following important lessons about its shop-floor assistant and, more generally, the adoption and use of AI by production workers:

Users are wary until they try it. Maintenance technicians at Siemens were initially skeptical about their personal future in a gen-AI-enhanced factory. Within a few weeks of using the shop-floor assistant, however, they reported feeling more secure in their jobs. Why? By dramatically reducing the time it took to find information, the tool allowed them to spend more time doing the important work that only they could do.

It's a valuable learning tool. Even without any onboarding, users quickly started using the shop-floor assistant to expand their understanding of the machines and potential causes of recurring incidents. This opens a whole new way of providing knowledge to production workers, who are not typically exposed to structured knowledge-sharing sessions as white-collar workers are. The tool also gives workers more autonomy over the time and place for learning, reducing their dependence on their more experienced colleagues' availability and willingness to act as teachers.

It empowers workers to take on more-complex work. Because maintenance technicians, for example, can handle many complicated incidents with the support of the shop- floor assistant, they have become less reliant on process engineers. In turn, with fewer requests from maintenance technicians, process engineers have more time for higher-value tasks, such as production-process optimization and technology updates.

It allows workers to complete their work more easily. During the 2024 test period, Siemens significantly downsized its teams overall, for reasons unrelated to the introduction of the gen AI tool. The smaller team was nonetheless able to maintain stable production output, even though members often had to tackle incidents alone, without the option of quickly calling a colleague for support. These workers even reported feeling less stress when they worked with the shop-floor assistant but lacked other support.

Siemens is leveraging this experience to develop a larger randomized control trial to test the causal impact of the shop-floor assistant in its factories, as well as across selected clients within its ecosystem. Additionally, the capabilities developed thanks to this exposure to experimentation led the company to apply similar methods to test whether and how the design of new "AI-intensive" jobs affects the quantity and quality of job applicants.

Becoming an Organizational Experimenter

We've outlined the many benefits of organizational experiments, but we're not claiming that they're easy to pull off. To implement them successfully, you'll need to focus on several critical areas.

Customer needs. At the heart of any successful gen-AI experiment lies a deep understanding of customer needs. Organizations must

focus on solving specific, high-impact problems. The experimental solutions must offer a clear potential return on investment—and a clearly articulated and testable potential impact. This requires conducting extensive customer interviews to ensure that solutions address urgent needs rather than merely offer "nice to have" enhancements. By distinguishing between strategic differentiators and minor inconveniences, companies can channel their resources into high-impact experiments and avoid scattering them across small pilots with questionable value. That is what P&G did in its experiment, for example: By building on a deep understanding of the traditional product-innovation process, the company recognized that a "cybernetic teammate" enabled by gen AI could help reduce the frictions that often arise between members of its marketing and R&D teams (the end users in this case), especially in the early stages of product development.

Usable prototypes. During the product development process, teams need to build early prototypes that people can actually use and test, and then they need to involve users in real-world experimentation that allows those prototypes to be quickly improved until they're ready for wider rollout. Doing that builds trust and makes it more likely that experiments will produce real insights and better results. This approach treats gen AI as a way not to replace workers but instead to help them do their jobs better.

A learning mindset. Traditional product development, which often moves slowly and has an inward focus, isn't well-suited for experimenting with gen AI. To drive innovation, companies need to embrace an experimental approach that brings customers into the process from the start, and in which cross-functional teams work in short sprints to test ideas and gather feedback quickly. While experimenting with its gen AI tool, Siemens relied on a preexisting tool called the Innovation Validation Engine, which ensures that everything

the company does is focused on solving real customer problems early and quickly. This approach hands more control to end users and makes product teams directly responsible for delivering value. It was exactly what Siemens needed to identify, validate, and develop gen AI applications in industrial settings with speed, precision, and market alignment.

Experimental expertise. Applying the scientific method inside companies requires a mix of skills. Teams need to know how to design and conduct good experiments (setting out clear and testable hypotheses, determining appropriate sample sizes) and how to run them well and keep them on track. Teams also need to be able to analyze results, explain what they mean, and use findings to make decisions. Because these skills are found in academia, some companies have turned to academic experts for help. In 2020, for example, Amazon hired Justine Hastings, a leading labor economist, to operationalize large-scale, people-focused experimentation. In 2022 Walmart brought on John A. List to help test and scale up myriad experiments, involving everything from gen AI merchandising tools to HR practices. Other companies are forging relationships with academic researchers that allow them to borrow experimental capabilities rather than bring them in-house. That's what Google, GitHub, and Procter & Gamble did to carry out the experiments described in this article.

Partnership capabilities. To drive effective prototyping, experimentation, discovery, and knowledge sharing, companies experimenting with gen AI need to develop active partnerships with a diverse set of players—suppliers, customers, and industry experts, as well as academics. The key is to create teams that have enough domain expertise and authority to design and conduct experiments that meet business needs and then to ensure that those teams can communicate reliably with the product development team. Above all, if you're a leader

who hopes to make experimentation an integral part of your company's strategic product development, you need to constantly make clear your commitment to continuous learning and data-driven decision-making.

...

Fast and rigorous experimentation is emerging as a strategic imperative in the gen AI era. Firms that develop the capacity to test, learn, and adapt in real time—both internally and across their ecosystems—will be better positioned to translate technological potential into organizational advantage. Here's the uncomfortable truth: While you're debating gen AI strategy, your competitors may be systematically learning what works. By embracing experimentation as a discipline, businesses can transform uncertainty into a source of strategic differentiation, and in doing so they can shape the future of work.

A version of this article appeared in the upcoming issue of Harvard Business Review.

JB

Johannes Berndt is a senior project manager in the People & Organizational Strategy division at Siemens.

FE

Florian Englmaier is a professor of organizational economics at LMU Munich.

Raffaella Sadun is the Charles E. Wilson Professor of Business Administration at Harvard Business School, a cochair of its Managing the Future of Work project, and the board chair of Harvard Business Publishing.

JT

Jorge Tamayo is an assistant professor in the Strategy Unit at Harvard Business School.

Nikolaus von Hesler is the global head of People & Organizational Strategy at Siemens.