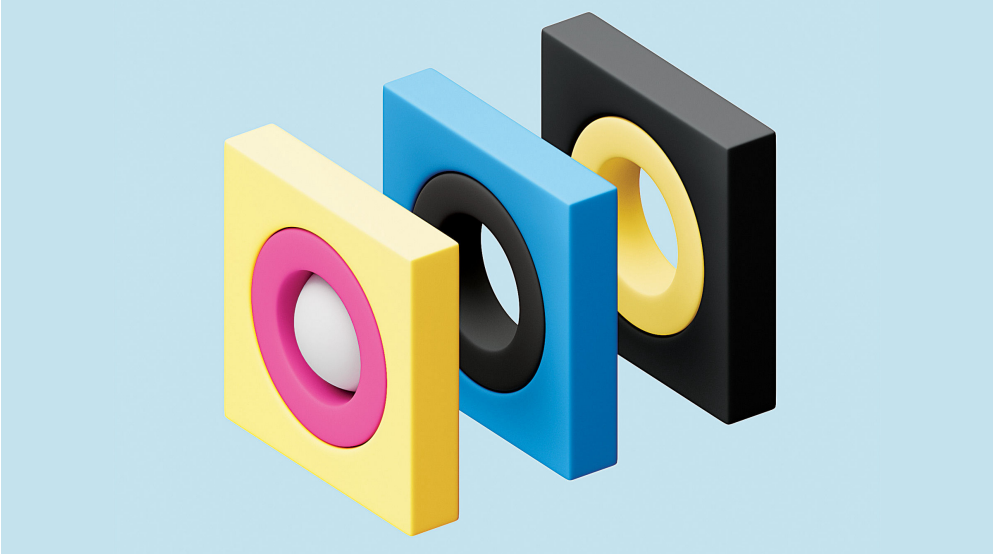
Harvard Business Review

Spotlight Article

AI and Machine Learning



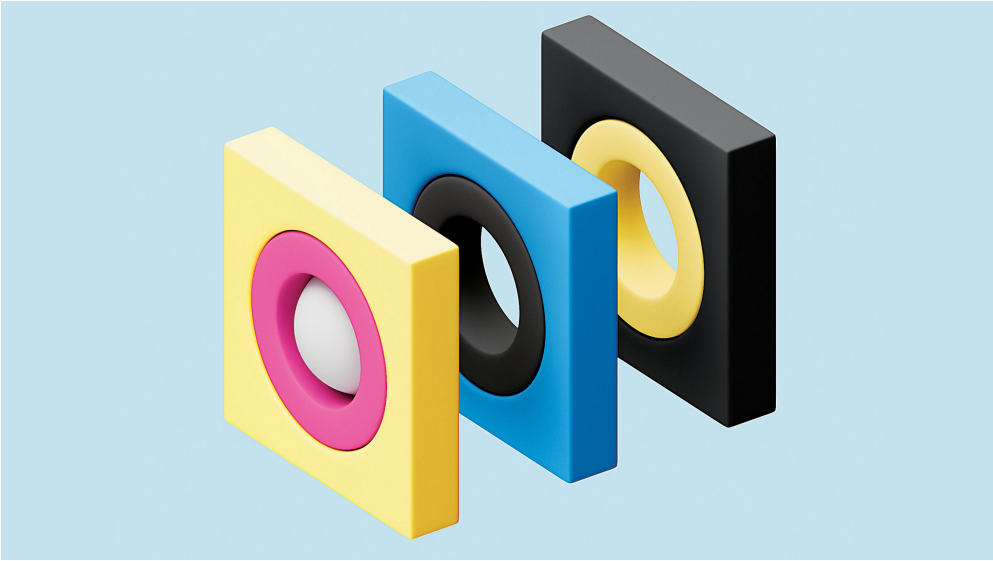
Keep Your AI Projects on Track

Most go off course. To make sure yours succeed, consider these five steps. **by Iavor Bojinov**

Keep Your AI Projects on Track

Most go off course. To make sure yours succeed, consider these five steps. **by Iavor Bojinov**

From the Magazine (November–December 2023) / Reprint S23063



Michael Brandon Myers

When I worked as a data scientist at LinkedIn in 2018 and 2019, AI was of interest only to a small team of people in the data science organization with advanced degrees in statistics or computer science. AI —and especially its newest star, generative AI—is now a central theme in corporate boardrooms, leadership discussions, and casual exchanges among employees eager to supercharge their productivity. The topic is so fundamental that "Data Science for Managers," a course I helped create to teach MBA students how to develop, leverage, and manage AI, is now a first-year requirement at Harvard Business School.

Sadly, beneath the aspirational headlines and the tantalizing potential lies a sobering reality: Most AI projects fail. Some estimates place the failure rate as high as 80%—almost double the rate of corporate IT project failures a decade ago. There are ways, however, to increase the odds of success. Through my experience in industry and academia and my consulting work, I have found that companies can greatly reduce their risk of failure by carefully navigating five critical steps that every AI project traverses on its way to becoming a product: *selection, development, evaluation, adoption,* and *management*. This article isn't just about preventing failure, however. It's also about developing what I call "data science and AI operations" processes that can help companies compete and survive in an increasingly AI-driven business landscape.

[1]

Selection

Although effectively prioritizing and sequencing projects is a task most leaders are familiar with, AI projects have a few idiosyncrasies that require careful consideration when assessing their impact and feasibility.

Begin by considering internal- and external-facing projects separately. Internal-facing projects are designed to help a company's employees perform their jobs. For example, organizations now regularly rely on AI to provide their sales teams with insights on which customers to target and what products to offer; to enhance their supply chain management through predictive analytics; and to streamline HR functions using smart chatbots for employee queries, among many other things. External-facing projects develop and implement AI for end users who are the company's customers. They are the most visible applications of AI and were pioneered by technology companies; examples include

Netflix's recommendation engine, Google's search results, and Uber's matching algorithm.

The following checklist is useful in the selection process.

Strategic alignment. *Is the project in line with the organization's overarching strategy and goals?* That may sound like a no-brainer, but in reality data scientists all too often lack a comprehensive grasp of their company's strategy and thus focus on projects unlikely to deliver transformative change and significant value. Because data scientists and other technical experts are often siloed, rarely interacting with the rest of the business, overcoming this hurdle can be challenging. Some companies, such as Procter & Gamble, temporarily embed data scientists in business units. Others, such as LinkedIn, keep a centralized reporting structure but have teams that mirror the specific business units and are physically colocated with them to boost collaboration. Embedding strategies like these make it easier for technical experts to learn the business and select strategically aligned projects.

Measurable impact. *Can we objectively assess the project's potential financial and operational benefits?* Explicitly consider the project's purpose and direction. By specifying how success will be measured, leaders can align teams around concrete goals and hold them accountable for achieving them. A useful framework for choosing such measures is the "if, then, by, because" paradigm: *If* this project is implemented, *then* this business outcome or key performance indicator will improve *by* having this anticipated impact *because* of this rationale supporting the anticipated impact. Companies with robust data-driven cultures have championed that hypothesis-driven scientific approach. Quantifying the impact also allows for direct comparison among multiple projects—an exercise that counters the tendency of technical

teams to focus on adopting the latest and most-advanced technologies and helps prioritize projects that are likely to have a substantial impact.

Augment or replace. *Will it enhance current human operations or is it replacing an existing manual process?* Answering this question requires understanding the cost of making errors (or prediction mistakes). When the cost is low, automation is generally viable; when the cost is high, it's better to augment the AI system with a human decision-maker. Take product recommendation: The cost of proposing an item that a customer is not interested in buying is relatively small, so it's safe to automate that task. However, the risk associated with an inaccurate diagnosis of a serious illness is extremely high, which is why doctors consult AI recommendations but retain the final word.

The impact assessment will guide leaders as they seek to understand each project's potential and help them select the efforts that promise to generate substantial business value. In addition, the information it generates can be disseminated across various teams and divisions within the organization, fostering a shared understanding and cross-pollination of ideas.

You must also assess the feasibility of an AI project: Obtain a basic cost estimate and determine whether your organization has the necessary resources to implement the project. This process consists of exploring four things.

Nature of the problem. *Is this a problem AI can solve?* AI is great at finding trends, identifying patterns, and providing predictions for well-formulated problems, but it fails to understand context, practice emotional intelligence, and exercise moral or ethical judgment. So use it judiciously. For example, it is great at identifying a sales prospect but would do poorly at closing a deal.

Data availability. *Does the organization have access to the necessary data?* An AI application's success depends on the availability, the quantity, the freshness, and the overall quality of underlying data. Take a company aiming to construct a lead-prioritization tool for its sales force. It would need comprehensive data on potential customers—job title, industry, and company size—as well as basic information such as name and phone number.

Technological capability and skills. *Does the organization have the infrastructure and skill set necessary to build, deploy, and scale up the project?* Technological infrastructure is project-specific, but at a minimum companies must have data storage and management capabilities, sufficient computational resources to train and run an AI, and systems to protect the data going into and coming out of it. The needed skills vary, but employees versed in data science and data engineering are usually required.

Ethical considerations. *Have all the ethical implications been fully considered?* It's crucial to identify them and to formulate a plan for addressing them during the development phase. If that doesn't happen, the costs can be high—in terms of reputational harm, government fines, and the engineering time required to retrofit the system to deal with them.

AI ethics is a broad topic that often includes three central themes: bias, privacy, and transparency.

Bias occurs when the available training data does not accurately represent the population the AI is intended to serve. It can diminish a model's accuracy and lead it to produce unfair outcomes—that is, cause it to systematically underperform for, or essentially discriminate against, specific groups or individuals. Consider a biotech company

assessing the feasibility of developing an AI to diagnose a type of cancer that affects both men and women of various races. If the company has data on white men only, the resulting models will provide biased or inaccurate results for women and people of color.

> **AI is great at identifying patterns and providing predictions for well-formulated problems, but it fails to practice emotional intelligence and exercise moral or ethical judgment.**

Privacy requires that AI models safeguard personal data and provide guarantees that it won't be leaked. Embracing "privacy by design" principles—developed by Ann Cavoukian, a former information and privacy commissioner of Ontario—in combination with emerging technologies such as differential privacy (which provides a mathematical measure of data privacy) can help companies adequately protect customer information.

As for transparency, users need to understand how an AI model works, evaluate its functionality, and know its strengths and limitations. Identifying the desired degree of transparency before starting a project is vital, because AI models make an inherent trade-off: Those that are more transparent and easier to explain tend to be less accurate, whereas those that are hard to explain often display superior performance.

After completing the analysis of potential projects' impact and feasibility, companies can classify each project as high or low in those two qualities. That will allow managers to estimate the return on investment and determine which projects should be selected and how they should be sequenced. Projects with high impact

but low feasibility should be investigated further to ascertain the root cause of the low feasibility; that can help managers identify opportunities for strengthening their infrastructure and data. Low-impact, high-feasibility projects should generally be ignored unless implementation would be cheap enough to justify their small impact or they would provide a suitable sandbox for testing infrastructure and new technologies.

[2]

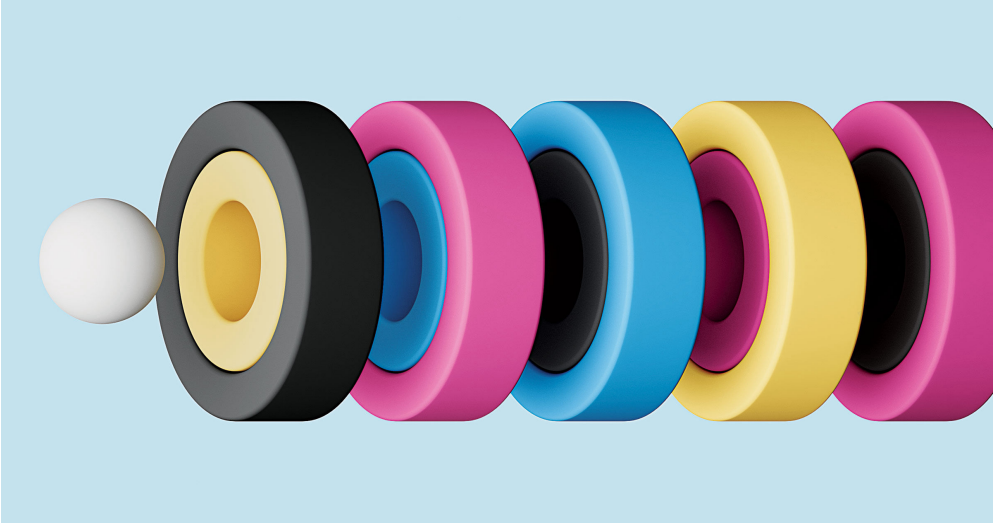
Development

Once a potential AI project has been given the green light, the complex, time-consuming development work begins. The intended users should be engaged throughout this process to ensure that the product meets their needs, which will pave the way for adoption.

Data scientists often go through multiple iterations of finding and cleaning data, performing exploratory data analysis, and training and evaluating AI models, and stop only when a model with the desired accuracy level is developed or the project is dropped. The next step is to build the means for integrating the model's outputs into the appropriate business processes. That integration, which turns the AI model into an AI product, typically entails developing software to transport data to the model for processing and then send the output to where it will be used. It might also require developing customized user interfaces or integrating the AI with other IT systems, such as customer-relationship-management tools.

Most companies conduct this process in an ad hoc manner with little standardization or specialization, resulting in an inefficient process that is prone to failure. Some technology firms have sought to create a better approach. LinkedIn, Netflix, and Uber are among those that

have developed internal tools for managing the entire AI-development process, from exploratory data analysis to deploying the product. Software companies such as Databricks and Snowflake, along with the large cloud-computing providers, offer this automation layer as a service. To take advantage of it, however, a company must build a centralized data repository.



Michael Brandon Myers

My colleagues Marco Iansiti and Karim Lakhani have dubbed this automation and standardization "the AI factory" (see "Competing in the Age of AI," HBR, January–February 2020). An AI factory increases the speed with which AI products are developed and standardizes key parts of the process, allowing for more monitoring and oversight. What's more, an AI factory can improve the overall quality of the models being developed. Indeed, an experiment I conducted in 2022 showed that providing data science with tools to automate parts of the development process led to a 30% improvement in the accuracy of the final model. The same experiment showed that by embedding AI-development knowledge into tools, an AI factory also reduces the skills needed to develop a product.

It's still too soon to know to what degree generative AI tools that can produce text, images, and code might improve the development process. But one controlled experiment involving GitHub Copilot suggests that software engineers using it to generate code can expect a significant boost in productivity, with the least experienced reaping the greatest benefit.

The takeaway for leaders is that they should create a center of excellence to build an easy-to-use AI factory and provide their employees with tool-specific training and education. In addition, they should ensure that any ethical issues identified during the selection phase have been addressed.

[3]

Evaluation

After an AI product has been developed, its impact should be evaluated before encouraging widescale adoption. Scientific experimentation, the simplest form of which is A/B testing, is the gold standard for quantifying the effect of a new AI model (see "The Surprising Power of Online Experiments," HBR, September–October 2017). Specifically, A/B tests pit the existing offering against an alternative version by randomly assigning users to either version and measuring engagement, satisfaction, and other relevant metrics. Companies often use experimentation platforms—either internally developed or purchased from providers such as LaunchDarkly, Optimizely, and Split—to execute such tests and analyze their results. Experimentation is also set to play a central role in the growth and adoption of generative AI by looking at how users respond to its various outputs.

AI products—even those that display amazing predictive accuracy during development—may not deliver sufficient value, for four

common reasons. First, AI doesn't exist in isolation: It interacts with other products, systems, and processes within the organization, leading to conflicts or issues not apparent during development. For example, a new AI-driven content-recommendation system might increase user engagement by reducing displayed ads, thereby lowering the company's profitability. Such changes can offset the potential benefits of deploying a system.

Second, the data used to train AI may not represent the actual users. When an AI model is presented with scenarios not covered in the training data, it attempts to extrapolate as best as it can, but its overall performance is usually adversely affected. That's part of the reason that self-driving cars are proving challenging to develop: So many unique driving situations exist that it's impossible to gather data covering all of them.

Third, deploying an AI model may inadvertently create negative feedback loops. For instance, an AI-driven content-recommendation system that shows only content very similar to what users have already engaged with might be acceptable to them once or twice, but by the 20th time they may find the recommendations boring and leave the platform.

> **Self-driving cars are proving challenging to develop in part because so many unique driving situations exist that it's impossible to gather data covering all of them.**

Finally, some models fail to adapt to changes in the real world. For instance, a pricing model trained on historical sales data might not anticipate trends, sudden market shifts (such as a global pandemic or

high inflation), or the reasons for consumers' purchase of one particular item as opposed to another, such as product availability or the store's return and exchange policies (see "A Step-by-Step Guide to Real-Time Pricing," by Marshall Fisher, Santiago Gallino, and Jun Li, in this issue).

In addition to quantifying the impact of an AI model, experimentation provides rapid user feedback that can help identify why a project fails to deliver adequate value. That early-stage feedback can substantially enhance the final product. In a study I conducted with LinkedIn in 2021, we estimated that incorporating such data results in a 20% improvement in business outcomes. In one instance a team at LinkedIn found that an AI model performed less well for some users than others because the training data was not representative of all the actual users. The team expanded the model to allow for more personalization—a simple change that drastically improved the model's performance.

Experimentation also lowers the risks of innovation by reducing the number of intended users exposed to potentially poor or negative experiences. New statistical methods even allow for automated continuous monitoring that stops experiments as soon as they are perceived to have a significant negative effect or unintended consequences. In one instance we reanalyzed an experiment that Netflix had run for two weeks and showed that this method would have stopped the test after a single day.

Beyond the typical pitfalls, which Guillaume Saint-Jacques, Martin Tingley, and I discussed in this magazine (see "Avoid the Pitfalls of A/B Testing," HBR, March–April 2020), A/B testing some AI projects can be hard for two reasons. First, the intended user base may be too small. This challenge is especially common with projects designed for just a handful of companies or employees. A strategy for overcoming it is to stagger users' exposure to the product being tested—comparing

their reactions before and after adoption—which reduces bias from idiosyncratic events.

Second, deploying the AI project to a subset of users might be impossible or impractical, particularly with AI that performs global optimization—a process for finding the best solution across all possible choices. Supply chain optimization, workforce scheduling, and product mix optimization all require global optimization. An A/B test with a subset of users could lead to an inferior solution and might not reflect all the potential advantages. In such a case, switchback experiments that intermittently alternate all users between operating with the AI system and without it can be helpful.

[4]

Adoption

After an AI project has been evaluated and shown to add enough value, it's time to focus on encouraging widescale adoption. When I was at LinkedIn, I led the team that launched an internal-facing tool for AI-driven data analysis. The tool automated large portions of the development process and made writing code unnecessary by allowing data scientists to describe the study they wanted to run through a simple user interface; our AI-driven back end handled the rest.

Early experimentation had shown that this approach drastically reduced the analysis time from days to a few hours. And yet after our launch few people used the product. Instead they continued to do what they had always done: perform custom analysis on their personal computers.

That was because the intended users didn't understand our product, weren't sure whether the AI was designed to work for them, and

didn't know who would be at fault if the analysis provided incorrect conclusions. Simply put, they didn't trust our product, so they didn't use it.

In my experience, trust in AI products has three pillars: the *algorithm,* the *developer,* and the *process.* People tend to trust AI when they believe it's working effectively, they have faith in the developer, and they think the process is designed to empower them without undue risk. Failure to adopt an AI product can almost always be traced back to a lack of trust in at least one of the three. The solution is to speak with people who tried the product but chose not to use it to identify the reason for their distrust.

Here are some questions skeptical users might have and an explanation of how to interpret them.

How does the AI product work? Is it free from bias? What are the assumptions that went into it? Why does the product make these predictions? Questions like this indicate a lack of trust in the algorithm. Begin by focusing on its efficacy. Quite possibly, certain forms of bias were missed during the development and evaluation stages. Next turn to harder-to-answer questions about understanding, transparency, and explanation. Because AI is often complicated, its intended users are unlikely to understand the algorithm's inner workings. So explain the assumptions encoded in the algorithm to show that the output is aligned with their intuition and business logic. Explain the data the algorithm was trained on to demonstrate the breadth and depth of knowledge ingrained in it. Provide use cases to show how the product performs in real-world scenarios.

The root cause of users' lack of trust in the LinkedIn product I helped develop was that in our push to make it as simple as possible, we had

created a black box. So we developed educational materials to describe the product's inner workings and explain our assumptions.

Did the developer have any hidden intentions? Did the developer listen to me and understand my needs? Is the algorithm replacing a task I enjoyed doing and consider to be valuable in performing my job? The developer may be an organization, a specific team within a company, or an individual. A lack of trust in the developer typically occurs when intended users were not part of the development process, in which case two things can go wrong: Users think the developer has a hidden agenda to replace their jobs with an algorithm, or they assume that the developer doesn't understand or care about their needs and won't deliver a product that addresses them. Overcoming both concerns requires clear communication and transparent explanations of the product's purpose.

At LinkedIn my team received input from several intended users throughout the development process. Because of our close partnership, those people trusted us. Unfortunately, other potential users jumped to the conclusion that our product was customized to solve just a handful of specific problems and was not flexible enough to deal with theirs. To overcome the lack of trust, we identified potential users with distinct problems and helped them use our product. Although that took a lot of time and energy, it demonstrated the product's flexibility.

If the AI gives an incorrect recommendation that I follow, am I at fault or is the AI? Do I have the authority to overrule the algorithm if I think it's wrong? AI products are inherently random and make mistakes. Understanding how to react is vital to ensuring trust in the process. When the AI is augmenting a human decision-maker, create a feedback loop and clear guidelines for resolving disagreements between the AI

and its users. When the AI is automating a process, ensure that strong protections have been put in place to identify and rectify mistakes.

To build trust in the process at LinkedIn, we created a certification board to review the results and subsequent recommendations before any derived insights could be shared. Thus responsibility for mistakes found after the certification process rested with the review board, not the users.

After we had taken all those steps to instill trust among users, the product was rapidly adopted throughout the organization. Five years later it's still in use.

[5]

Management

The journey is far from over when an AI product has been adopted. Ensuring its ongoing success requires a diligent, proactive management strategy to sustain and enhance results. A basic requirement is to provide engineering support—to fix bugs, for instance—and to monitor the product for changes in performance. The most common cause of a drop in performance is that the training data has become outdated. Consider, for example, a company that has developed an AI to predict customer purchasing behavior. If customers' preferences or market conditions shift over time, the model's predictions will be less accurate. It's important to regularly retrain models on fresh data, but because that can be expensive, many companies have opted to build mechanisms that monitor and alert managers to significant changes in the model's performance that might indicate a need for retraining.

In addition to monitoring, companies should perform AI audits to look for unintended consequences, ethical issues, and security flaws. For

instance, in a study published in 2022 in *Science,* my coauthors and I discovered that LinkedIn's People You May Know algorithm, designed to broaden a user's network by suggesting potential new connections, inadvertently altered users' career prospects. This effect was due in part to the relationship between individuals' social networks and their access to information about new jobs and opportunities. Although in this case the change was beneficial, AI projects can have the opposite effect, making audits essential.

• • •

Successful management isn't just about maintaining the status quo. It also involves continual improvement to ensure that the product evolves in step with technologies and users' changing needs. That entails a cyclical process of gathering more data, refining algorithms, and promoting increased usage. The five steps I've described can significantly increase the odds that leaders will choose the highest-value projects to pursue and can deliver on their promise.

A version of this article appeared in the November–December 2023 issue of Harvard Business Review.

Iavor Bojinov is a former data scientist at LinkedIn. He is currently an assistant professor of business administration and the Richard Hodgson Fellow at Harvard Business School.

