TECHNOLOGY

A BETTER WAY TO ONBOARD AI

Understand it as a tool to assist rather than replace people

AUTHORS

Boris Babic
Assistant professor, INSEAD

Daniel L. Chen
Professor, Toulouse School of Economics

Theodoros Evgeniou
Professor, INSEAD

Anne-Laure Fayard
Associate professor, NYU

PHOTOGRAPHER JEAN-BAPTISTE PERROT

ABOUT THE ART
Photographer Jean-Baptiste Perrot roots his work in the dysfunctions of new technologies, tktktk.

IN A 2018 WORKFORCE INSTITUTE SURVEY OF 3,000 MANAGERS ACROSS EIGHT INDUSTRIALIZED NATIONS, THE MAJORITY OF RESPONDENTS DESCRIBED ARTIFICIAL INTELLIGENCE AS A VALUABLE PRODUCTIVITY TOOL.

It's easy to see why: AI brings tangible benefits in processing speed, accuracy, and consistency (machines don't make mistakes because they're tired), which is why many professionals now rely on it. Some medical specialists, for example, use AI tools to help make diagnoses and decisions about treatment.

But respondents to that survey also expressed fears that AI would take their jobs. They are not alone. The *Guardian* recently reported that more than 6 million workers in the UK fear being replaced by machines. These fears are echoed by academics and executives we meet at conferences and seminars. AI's advantages can be cast in a much darker light: Why would humans be needed when machines can do a better job?

The prevalence of such fears suggests that organizations looking to reap the benefits of AI need to be careful when introducing it to the people expected to work with it. Andrew Wilson, until January 2020 Accenture's CIO, says, "The greater the degree of organizational focus on people helping AI, and AI helping people, the greater the value achieved." Accenture has found that when companies make it clear that they are using AI to help people rather than to replace them, they significantly outperform companies that don't set that objective (or are unclear about their AI goals) along most dimensions of managerial productivity—notably speed, scalability, and effectiveness of decision-making.

In other words, just as when new talent joins a team, AI must be set up to succeed rather than to fail. A smart

IDEA IN BRIEF

THE PROBLEM
Many companies struggle to apply AI and fail to achieve the productivity improvements they seek.

WHY IT HAPPENS
Executives often don't make clear that they are using AI to help people increase productivity rather than to replace them.

HOW TO FIX IT
Treat AI adoption as an onboarding process that consists of four phases: AI as an assistant, as a monitor, as a coach, and as a teammate.

employer trains new hires by giving them simple tasks that build hands-on experience in a noncritical context and assigns them mentors to offer help and advice. This allows the newcomers to learn while others focus on higher-value tasks. As they gain experience and demonstrate that they can do the job, their mentors increasingly rely on them as sounding boards and entrust them with more-substantive decisions. Over time an apprentice becomes a partner, contributing skills and insight.

We believe this approach can work for artificial intelligence as well. In the following pages we draw on our own and others' research and consulting on AI and information systems implementation, along with organizational studies of innovation and work practices, to present a four-phase approach to implementing AI. It allows enterprises to cultivate people's trust—a key condition for adoption—and to work toward a distributed human-AI cognitive system in which people and AI *both* continually improve. Many organizations have experimented with phase 1, and some have progressed to phases 2 and 3. For now, phase 4 may be mostly a "future-casting" exercise of which we see some early signs, but it is feasible from a technological perspective and would provide more value to companies as they engage with artificial intelligence.

PHASE 1

THE ASSISTANT

This first phase of onboarding artificial intelligence is rather like the process of training an assistant. You teach the new employee a few fundamental rules and hand over some basic but time-consuming tasks you normally do (such as filing online forms or summarizing documents), which frees you to focus on more-important aspects of the job. The

When portfolio managers are choosing stocks in which to invest, the information available is far more than a human can feasibly process.

trainee learns by watching you, performing the tasks, and asking questions.

One common task for AI assistants is sorting data. An example is the recommendation systems companies have used since the mid-1990s to help customers filter thousands of products and find the ones most relevant to them—Amazon and Netflix being among the leaders in this technology.

More and more business decisions now require this type of data sorting. When, for example, portfolio managers are choosing stocks in which to invest, the information available is far more than a human can feasibly process, and new information comes out all the time, adding to the historical record. Software can make the task more manageable by immediately filtering stocks to meet predefined investment criteria. Natural-language processing, meanwhile, can identify the news most relevant to a company and even assess the general sentiment about an upcoming corporate event as reflected in analysts' reports. Marble Bar Asset Management (MBAM), a London-based investment firm founded in 2002, is an early convert to using such technologies in the workplace. It has developed a proprietary state-of-the-art platform, called RAID (Research Analysis & Information Database), to help portfolio managers filter through high volumes of information about corporate events, news developments, and stock movements.

Another way AI can lend assistance is to model what a human might do. As anyone who uses Google will have noticed, prompts appear as a search phrase is typed in. Predictive text on a smartphone offers a similar way to speed up the process of typing. This kind of user modeling, related to what is sometimes called *judgmental bootstrapping,* was developed more than 30 years ago; it can easily be applied to decision-making. AI would use it to identify the choice an employee is most likely to make, given that employee's past choices, and would suggest that choice as a starting point when the employee is faced with multiple decisions—speeding up, rather than actually doing, the job.

Let's look at this in a specific context. When airline employees are deciding how much food and drink to put on a given flight, they fill out catering orders, which involve a certain amount of calculation together with assumptions based on their experience of previous flights. Making the wrong choices incurs costs: Underordering risks upsetting customers who may avoid future travel on the airline. Overordering means the excess food will go to waste and the plane will have increased its fuel consumption unnecessarily.

An algorithm can be very helpful in this context. AI can predict what the airline's catering manager would order by analyzing his or her past choices or using rules set by the manager. This "autocomplete" of "recommended orders" can be customized for every flight using all relevant historical data, including food and drink consumption on the route in question and even past purchasing behavior by passengers on the manifest for that flight. But as with predictive typing, human users can freely overwrite as needed; they are always in the driver's seat. AI simply assists them by imitating or anticipating their decision style.

It should not be a stretch for managers to work with AI in this way. We already do so in our personal lives, when we allow the autocomplete function to prefill forms for us online. In the workplace a manager can, for example, define specific rules for an AI assistant to follow when completing forms. In fact, many software tools currently used in the workplace (such as credit-rating programs) are already just that: collections of human-defined decision rules. The AI assistant can refine the rules by codifying the circumstances under which the manager actually follows them. This learning needn't involve any change in the manager's behavior, let alone any effort to "teach" the assistant.

PHASE 2
THE MONITOR

The next step is to set up the AI system to provide real-time feedback. Thanks to machine-learning programs, AI can be trained to accurately forecast what a user's decision would be in a given situation (absent lapses in rationality owing to, for example, overconfidence or fatigue). If a user is about to make a choice that is inconsistent with his or her choice history, the system can flag the discrepancy. This is especially helpful during high-volume decision-making, when human employees may be tired or distracted.

Research in psychology, behavioral economics, and cognitive science shows that humans have limited and imperfect reasoning capabilities, especially when it comes to statistical and probabilistic problems, which are ubiquitous in business. Several studies (of which one of us, Chen, is a coauthor) concerning legal decisions found that judges grant political asylum more frequently before lunch than after, that they give lighter prison sentences if their NFL team won the previous day than if it lost, and that they will go easier on a defendant on the latter's birthday. Clearly justice might be better served if human decision makers were assisted by software that told them when a decision they were planning to make was inconsistent with their prior decisions or with the decision that an analysis of purely legal variables would predict.

AI can deliver that kind of input. Another study (also with Chen as a coauthor) showed that AI programs processing a model made up of basic legal variables (constructed by the study's authors) can predict asylum decisions with roughly 80% accuracy on the date a case opens. The authors have added learning functionality to the program, which enables it to simulate the decision-making of an individual judge by drawing on that judge's past decisions.

The approach translates well to other contexts. For example, when portfolio managers (PMs) at Marble Bar Asset Management consider buy or sell decisions that may raise the overall portfolio risk—for example, by increasing exposure to a particular sector or geography—the system alerts them through a pop-up during a computerized transaction process so that they can adjust appropriately. A PM may ignore such feedback as long as company risk limits are observed. But in any case the feedback helps the PM reflect on his or her decisions.

Of course AI is not always "right." Often its suggestions don't take into account some reliable private information to which the human decision maker has access, so the AI might steer an employee off course rather than simply correct for possible behavioral biases. That's why using it should be like a dialogue, in which the algorithm provides nudges according to the data it has while the human teaches the AI by explaining why he or she overrode a particular nudge. This improves the AI's usefulness and preserves the autonomy of the human decision maker.

Unfortunately, many AI systems are set up to usurp that autonomy. Once an algorithm has flagged a bank transaction as possibly fraudulent, for example, employees are often unable to approve the transaction without clearing it with a supervisor or even an outside auditor. Sometimes undoing a machine's choice is next to impossible—a persistent source of frustration for both customers and customer service professionals. In many cases the rationale for an AI choice is opaque, and employees are in no position to question that choice even when mistakes have been made.

Privacy is another big issue when machines collect data on the decisions people make. In addition to giving humans control in their exchanges with AI, we need to guarantee that any data it collects on them is kept confidential. A wall ought to separate the engineering team from management; otherwise employees may worry that if they freely interact with the system and make mistakes, they might later suffer for them.

Also, companies should set rules about designing and interacting with AI to ensure organizational consistency in norms and practices. These rules might specify the level of predictive accuracy required to show a nudge or to offer a reason for one; criteria for the necessity of a nudge; and the conditions under which an employee should either follow the AI's instruction or refer it to a superior rather than accept or reject it.

To help employees retain their sense of control in phase 2, we advise managers and systems designers to involve them in design: Engage them as experts to define the data that will be used and to determine ground truth; familiarize them with models during development; and provide training and interaction as those models are deployed. In the process, employees will see how the models are built, how the data is managed, and why the machines make the recommendations they do.

PHASE 3

THE COACH

In a recent PwC survey nearly 60% of respondents said that they would like to get performance feedback on a daily or a weekly basis. It's not hard to see why. As Peter Drucker asserted in his famous 2005 *Harvard Business Review* article

Of course AI is not always “right.” That’s why using it should be like a dialogue, to improve its usefulness and preserve the autonomy of the human decision maker.

“Managing Yourself,” people generally don’t know what they are good at. And when they think they do know, they are usually wrong.

The trouble is that the only way to discover strengths and opportunities for improvement is through a careful analysis of key decisions and actions. That requires documenting expectations about outcomes and then, nine months to a year later, comparing those expectations with what actually happened. Thus the feedback employees get usually comes from hierarchical superiors during a review—not at a time or in a format of the recipient’s choosing. That is unfortunate, because, as Tessa West of New York University found in a recent neuroscience study, the more people feel that their autonomy is protected and that they are in control of the conversation—able to choose, for example, when feedback is given—the better they respond to it.

AI could address this problem. The capabilities we’ve already mentioned could easily generate feedback for employees, enabling them to look at their own performance and reflect on variations and errors. A monthly summary analyzing data drawn from their past behavior might help them better understand their decision patterns and practices. A few companies, notably in the financial sector, are taking this approach. Portfolio managers at MBAM, for example, receive feedback from a data analytics system that captures investment decisions at the individual level.

The data can reveal interesting and varying biases among PMs. Some may be more loss-averse than others, holding on to underperforming investments longer than they should. Others may be overconfident, possibly taking on too large a position in a given investment. The analysis identifies these behaviors and—like a coach—provides personalized feedback that highlights behavioral changes over time, suggesting how to improve decisions. But it is up to the PMs to decide how to incorporate the feedback. MBAM’s leadership believes this “trading enhancement” is becoming a core differentiator that both helps develop portfolio managers and makes the organization more attractive.

What’s more, just as a good mentor learns from the insights of the people who are being mentored, a machine-learning “coachbot” learns from the decisions of an empowered human employee. In the relationship we’ve described, a human can disagree with the coachbot—and that creates new data that will change the AI's implicit model. For example, if a portfolio manager decides not to trade a highlighted stock because of recent company events, he or she can provide an explanation to the system. With feedback, the system continually captures data that can be analyzed to provide insights.

If employees can relate to and control exchanges with artificial intelligence, they are more likely to see it as a safe channel for feedback that aims to help rather than to assess performance. Choosing the right interface is useful to this end. At MBAM, for example, trading enhancement tools—visuals, for instance—are personalized to reflect a PM's preferences.

As in phase 2, involving employees in designing the system is essential. When AI is a coach, people will be even more fearful of disempowerment. It can easily seem like a competitor as well as a partner—and who wants to feel less intelligent than a machine? Concerns about autonomy and privacy may be even stronger. Working with a coach requires honesty, and people may hesitate to be open with one that might share unflattering data with the folks in HR.

Deploying AI in the ways described in the first three phases does of course have some downsides. Over the long term new technologies create more jobs than they destroy, but meanwhile labor markets may be painfully disrupted. What's more, as Matt Beane argues in "Learning to Work with Intelligent Machines" (HBR, September–October 2019), companies that deploy AI can leave employees with fewer opportunities for hands-on learning and mentorship.

There is some risk, therefore, not only of losing entry-level jobs (because digital assistants can effectively replace human ones) but also of compromising the ability of future decision makers to think for themselves. That's not inevitable, however. As Beane suggests, companies could use their artificial intelligence to create different and better learning opportunities for their employees while improving the system by making it more transparent and giving employees more control. Because future entrants to the workforce will have grown up in a human-plus-machine workplace, they will almost certainly be faster than their pre-AI colleagues at spotting opportunities to innovate and introduce activities that add value and create jobs—which brings us to the final phase.

PHASE 4
THE TEAMMATE

Edwin Hutchins, a cognitive anthropologist, developed what is known as the theory of distributed cognition. It is based on his study of ship navigation, which, he showed, involved a combination of sailors, charts, rulers, compasses, and a plotting tool. The theory broadly relates to the concept of extended mind, which posits that cognitive processing, and associated mental acts such as belief and intention, are not necessarily limited to the brain, or even the body. External tools and instruments can, under the right conditions, play a role in cognitive processing and create what is known as a *coupled system.*

In line with this thinking, in the final phase of the AI implementation journey (which to our knowledge no organization has yet adopted) companies would develop a coupled network of humans and machines in which both contribute expertise. We believe that as AI improves through its interactions with individual users, analyzing and even modeling expert users by drawing on data about their past decisions and behaviors, a community of experts (humans and machines) will naturally emerge in organizations that have fully integrated AI coachbots. For example, a purchasing manager who—with one click at the moment of decision—could see what price someone else would give could benefit from a customized collective of experts.

Although the technology to create this kind of collective intelligence now exists, this phase is fraught with challenges. For example, any such integration of AI must avoid building in old or new biases and must respect human privacy concerns so that people can trust the AI as much as they would a human partner. That in itself is a pretty big challenge, given the volume of research demonstrating how hard it is to build trust among humans.

The best approaches to building trust in the workplace rely on the relationship beween trust and understanding—a subject of study by David Danks and colleagues at Carnegie Mellon. According to this model, I trust someone because I understand that person's values, desires, and intentions, and they demonstrate that he or she has my best interests at

When AI is a coach, people will be even more fearful of disempowerment. It can easily seem like a competitor as well as a partner.

heart. Although understanding has historically been a basis for building trust in human relationships, it is potentially well suited to cultivating human-AI partnerships as well, because employee's fear of artificial intelligence is usually grounded in a lack of understanding of how AI works. (See the sidebar "When AI Loses Its Way.")

In building understanding, a particular challenge is defining what "explanation" means—let alone "good explanation." This challenge is the focus of a lot of research. For example, one of us (Evgeniou) is working to open up machine-learning "black boxes" by means of so-called counterfactual explanations. A counterfactual explanation illuminates a particular decision of an AI system (for example, to approve credit for a given transaction) by identifying a short list of transaction characteristics that drove the decision one way or another. Had any of the characteristics been different (or counter to the fact), the system would have made a different decision (credit would have been denied).

WHEN AI LOSES ITS WAY

In 2016 the investigative newsroom *ProPublica* published an exposé of a risk-prediction AI program known as COMPAS, which judges in southern Florida use to determine a defendant's likelihood of re-offending within a specified time period.

The algorithm underlying COMPAS is held as a trade secret by its manufacturer, Northpointe (now Equivant), which means that we don't know how COMPAS generates its predictions, nor do we have access to the data the algorithm is trained on—so we cannot even inquire into its rationale. When it was reported that the algorithm produces disparate outcomes across race, COMPAS immediately became a leading example of why people cannot trust AI.

If businesses want employees to adopt, use, and ultimately trust AI systems, it will be important to open up the black box—to the extent legally possible—to those who are expected to engage with the technology. As Richard Socher, the chief scientist at Salesforce, puts it, "If businesses use AI to make predictions, they owe humans an explanation as to how the decisions are made."

Evgeniou is also exploring what people perceive as good explanations for AI decisions. For example, do they see an explanation as better when it's presented in terms of a logical combination of features ("The transaction was approved because it had X,Y,Z characteristics") or when it's presented relative to other decisions ("The transaction was approved because it looks like other approved transactions, and here they are for you to see")? As research into what makes AI explainable continues, AI systems should become more transparent, thus facilitating trust.

ADOPTING NEW TECHNOLOGIES has always been a major challenge—and the more impact a technology has, the bigger the challenge is. Because of its potential impact, artificial intelligence may be perceived as particularly difficult to implement. Yet if done mindfully, adoption can be fairly smooth. That is precisely why companies must ensure that its design and development are responsible—especially with regard to transparency, decision autonomy, and privacy—and that it engages the people who will be working with it. Otherwise they will quite reasonably fear being constrained—or even replaced—by machines that are making all sorts of decisions in ways they don't understand.

Getting past these fears to create a trusting relationship with AI is key. In all four phases described in these pages, humans determine the ground rules. With a responsible design, AI may become a true partner in the workplace—rapidly processing large volumes of varied data in a consistent manner to enhance the intuition and creativity of humans, who in turn teach the machine.

HBR Reprint R2004#

BORIS BABIC *is an assistant professor of decision sciences at INSEAD.* **DANIEL L. CHEN** *is a professor at the Institute for Advanced Study at the Toulouse School of Economics and lead investigator at the World Bank's Data and Evidence for Justice Reform program.* **THEODOROS EVGENIOU** *is a professor of decision sciences and technology management at INSEAD and a senior adviser to Marble Bar Asset Management (an investment firm named in this article).* **ANNE-LAURE FAYARD** *is an associate professor of innovation, design, and organization studies at NYU's Tandon School of Engineering.*