Onboarding AI

Boris Babic, Daniel L. Chen, Theodoros Evgeniou, Anne-Laure Fayard

Onboarding AI

By Boris Babic, Daniel L. Chen, Theodoros Evgeniou, and Anne-Laure Fayard

Working squib: Whether AI improves decisions and creates value or not hinges on smart adoption

Boris Babic is a professor at INSEAD a business school with campuses in Abu Dhabi, France, and Singapore. Daniel L. Chen is a professor at the Institute for Advanced Studies of the Toulouse School of Economics in France. Theodoros Evgeniou is a professor at INSEAD. Anne-Laure Fayard is a professor at NYU's Tandon School of Engineering in New York.

In a 2018 Workforce Institute survey of 3,000 managers across eight industrialized nations, the majority of respondents described AI 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. Medical specialists in many fields, for example, use AI tools to help diagnose illness and make decisions about treatment.

But the same respondents to that survey also expressed fears that AI would take their jobs. They are not alone. The UK's Guardian newspaper recently reported that "more than 6 million workers [in the UK alone] fear being replaced by machines." We hear these fears echoed by the academics and executives we talk to at conferences and seminars. The advantages of AI can be cast in a much darker light: why would we need humans 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 about how they introduce it to the people expected to work with it. As Accenture's CIO, Andrew Wilson, argues, "the greater the degree of organizational focus on people helping AI, and AI helping people, the greater the value achieved." Accenture's research confirms this: they find that companies using AI to improve human productivity (rather than replace humans directly) significantly outperform -- in decision making speed, scalability, and effectiveness, and other performance dimensions.

We need, in other words, to set AI up to succeed rather than to fail – just as we have to do when adopting new talent into a team. In that context, a smart employer will train the newcomer by setting her simple tasks to gain hands-on experience in a non-critical context and assign a mentor to offer help and advice. This offers them an opportunity to learn while giving others time to focus on more value-added tasks. As the newcomer gains experience and demonstrates that she can do the job, the mentor starts to rely more and more on her, both as a sounding board and as someone to

delegate more substantive decisions to. Over time the apprentice becomes an equal partner, contributing her skills and insight as an equal.

It's an approach that we believe can work for AI as well. In the following pages, we draw on our own and others' research and consulting on AI and IS (Information Systems) implementation, as well as on decision making and organizational studies of innovation and work practices, to present a four-phase approach to implementing AI. Our approach to this organizational journey allows organizations to efficiently implement AI while cultivating people's trust in it (a key condition to adoption) and working toward a distributed human-AI cognitive system, in which *both people and AI continuously improve*. Many organizations have experimented with Phase 1, and some have also progressed to Phases 2 and 3. For now, Phase 4 may be mostly a "future-casting" exercise for which we see some early signs, but one that is feasible from a technological perspective and one that would provide more value to organizations as they start truly engaging with AI.

1: Offloading: AI as Assistant

This first phase of onboarding AI is rather like taking on a trainee assistant. You teach the trainee some basic rules and get them to take on the more formulaic but time-consuming tasks you normally do (like photocopying or data entry), which frees you to focus on other more important aspects of the task. The trainee learns through doing the tasks, watching you, and asking questions.

AI assistants that take on routine jobs of sorting data input for decisions aren't an entirely new phenomenon. Recommender systems have already been used since the mid-90s to help customers filter thousands of products to decide which are the most relevant ones for them – Amazon and Netflix being among the leaders in this space.

There is plenty of room, though, to do more, as more and more business decisions require sorting through a lot of data. When portfolio managers, for example, pick stocks to invest, there is more information available than the human can feasibly process – think about just how much relevant new information comes out in a given day, pertaining to the stocks they have to choose from, on top of all the relevant information in the historical record.

Information filtering software, however, can help make the task more manageable by immediately reducing the list of stocks to those that meet predefined investment criteria. Natural Language Processing (NLP) technologies, for example, can identify the most relevant news for a company or even assess the general sentiment about an upcoming corporate event as that is reflected in news and 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. To this end, it has developed a proprietary state-of-the-art platform, called RAID, to help portfolio

managers (PMs) filter through high volumes of corporate events, news, and stock movements to make stock selections more effectively and manage their investment processes (see the sidebar).

AI can do more than simply filter data. As anyone who uses Google will have noticed that Google will prompt terms as you type in a search phrase. Predictive text on a smartphone works in a similar way to help speed up the process of typing text. This functionality, sometimes called "judgmental bootstrapping," was developed more than 30 years ago and it can easily be applied to decision-making. An AI would use this functionality to identify what an employee would most likely choose, given the employee's past choices and simply suggest that choice as a starting point when the employee is faced with multiple decisions, hence speeding up – instead of replacing – the job. If the prompt is based on the employee's past choices – rather than on some notion of an objective best choice, then the AI would be helping the employee decide faster rather than taking the decision away from the human.

Let's look at what this would mean in a specific context. When an airline has to decide how much food and drink to put on a given flight, employees will fill out catering orders, which will involve a certain amount of calculation and the application of rules of thumb born out of the employee's experience of previous flights. There are costs to making the wrong choices. If they order too little food and drink, they risk upsetting customers who may avoid traveling on the airline in the future. If they over-order, the excess food that the airline has paid for will go to waste and the plane's fuel consumption will be unnecessarily higher.

An algorithm can be very helpful in this context. Just like predictive typing predicts words from the letters we have already typed, AI can predict what our airline catering manager would order, based on analyzing past choices that the employee has made or even using rules defined by the employee. This "auto-complete" of "recommended orders" can be customized for every single flight, using AI trained on all relevant past data - historical data of food and drink consumption on the route in question and even information it might have access to from credit card records of past purchasing behavior of passengers on the manifest for the flight in question. But like predictive typing, the human users can freely overwrite as needed – keeping them always in the driver's seat. AI is *not* trained to replace the human, or to ignore their preferences, but to assist them by being *customized* to their decision style using past input and data from their choices.

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 search for products and find the autocomplete function prefilling forms for us online. In the workplace, a manager can easily initialize her own personal AI assistant by, for example, simply defining specific rules for completing forms. In fact, many existing software tools (credit-rating programs, for example) used in the workplace already are just that collections of human-defined

decision rules. Where the AI comes in is after the basic rules have been initialized, because the AI assistant can refine the rules by processing under what circumstances the manager actually follows the rules. This learning, moreover, doesn't have to involve any change in our behavior, let alone any effort in "teaching" the assistant.

2: Monitoring: AI as Monitor

The next step is to set up the AI system to provide real-time feedback. Thanks to machine learning programs, an AI system can be trained to accurately forecast based on past behavior what a user's decision would be in a situation *absent lapses in rationality*. It can detect if a user is about to make a choice that is inconsistent with their choice history and inform them of this discrepancy between past and present -- this is especially helpful in high-volume decision making where human employees may become tired or distracted.

Research in psychology, behavioral economics, and cognitive science shows that humans have limited and imperfect reasoning capacities, especially when it comes to statistical and probabilistic problems, which are ubiquitous in business. Several studies of legal decisions (of which Daniel L. Chen is a co-author) found that judges grant political asylum petitions more frequently before lunch than after, give lighter sentences the day after their NFL football team wins than if it loses, and will go easier on a defendant on the latter's birthday. Clearly, in this context, justice would be better served if human decision-makers were to be assisted by a software program that could highlight whether or not a decision they are planning to make would be inconsistent with their prior decisions or with the decision that would be predicted from an analysis of purely legal variables.

AI can deliver this kind of input. Another study (by the same co-author) 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 the date a case opens. The authors have, moreover, 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. This can tell a judge whether a decision he or she proposes in a given case is or is not consistent with what the judge using the system would likely conclude absent any non-legal influencing variables.

The approach translates well into other contexts. For example when portfolio managers at MBAM consider buy or sell decisions that may increase the overall portfolio risk – for example by increasing exposure to a particular sector or geography — the system alerts them through an on-screen pop-up during a computerized transaction process, so that they can adjust appropriately. The fund managers can ignore such feedback, when of course company policy risk limits are not broken, but in either case such feedback helps the fund manager to reflect on his or her decisions.

Of course, the AI system is not always "right". In many cases, its suggestions will not take into account some reliable private information that the human decisionmaker has access to. Where this happens, an AI might well steer an employee off course rather than simply correct for possible behavioral biases. This is why using AI should be like a dialogue, in which the algorithm provides nudges based on the data it has while the human teaches the AI by explaining why her or she had over-ridden a particular nudge. This helps the AI improve its usefulness and preserves the autonomy of the human decision-maker.

Unfortunately, many AI systems are actually set up to usurp the decision autonomy of human employees. Once an algorithm flags a bank transaction as possibly fraudulent, for example, human employees are often unable to approve it without going through several management layers -- clearing with their supervisor or even an outside auditor. Sometimes undoing a machine's choice is next to impossible, which is a persistent source of frustration for both customers and customer service professionals. In many cases the rationale for AI choices (like flagging a bank transaction as fraudulent) are opaque - an employee does not know why the AI did what it did and is not in a position to question that choice. This leads to skepticism and, ultimately, failures of adoption.

Privacy also becomes a big issue when machines collect data on the decisions people make. In addition to conferring control on the human in her exchanges with an AI, we need to guarantee that the data on the human collected and processed by the AI is kept private. If we are collecting data for machine learning purposes, for instance, we should probably build a wall of separation between the engineering team and various managerial/HR divisions. Otherwise, employees may worry that if they freely interact with the system, and "make mistakes", these might later be held against them.

Companies should also harmonize rules around designing and interacting with AI to ensure organizational consistency in the norms and practices associated with AI. These might include specifying what level of predictive accuracy is needed before showing a nudge or offering a reason for the nudge and setting criteria for determining when a nudge is necessary. And under what conditions is it necessary for the employee to either follow the instruction or refer it to a superior rather than accept or reject the nudge herself?

To help retain employees' sense of control in the deeper relationship they develop with AI on entering the second phase of the journey, we advise managers and designers of the systems to involve employees in design: engaging them as experts to define the data to use and to determine ground truth, familiarizing them with the models as they are developed, providing training and interaction as they are deployed. This has the further benefit of showing employees how the models are built, how the data is managed, and why the machines make the recommendations they do.

3. Supporting: AI as Coach

According to a recent PwC survey, nearly 60% of their respondents reported that they would like performance feedback on a daily or weekly basis. It's not hard to see why: as Peter Drucker noted in his famous *Harvard Business Review* article, "Manage Yourself", people generally don't know what they are good at. And when they think they do they are usually wrong.

The trouble is that the only way to discover strengths and improvement opportunities is through careful analysis of key decisions and actions, which require documenting expectations about outcomes, and then, 9 to 12 months later, comparing what effectively happened with the expectations. That requires investment of time and energy at critical points, which many employees will struggle to do systematically by themselves. As a result, the feedback they get is generally given by others, usually hierarchical superiors in the context of a review. The content of this feedback, therefore, is not given at a time or in a format of the recipient's choosing. This is unfortunate because, as NYU's Tessa West found in a recent neuroscience study, people respond better to feedback the more they feel that their autonomy is protected and that they are in control of the conversation – able to choose, for example, when the feedback conversation takes place.

AI can address this problem. The capabilities we've already documented can easily be used to generate feedback to employees, allowing users to look at their own performance and reflect on variations and errors. It can also use a case-based training system to help users better understand their decision patterns and practices.

A few companies, notably in the financial sector, are already some way down this path. MBAM portfolio managers (PMs), for example, receive feedback from a data analytics system that captures investment decisions at the individual level. The system can, for instance, model alternative portfolios that may be modifications of the real ones the PMs manage, indicating how certain changes in the way PMs trade would have made a difference in the past. It can also help identify what characteristics of stocks differentiate those that the PM traded from those that the PM did not trade.

The data can also reveal interesting and varying biases among PMs. Some may be more loss averse than others, keeping losing investments longer than they should. Others may be overconfident, possibly taking on too large a position in a given investment. The analytics identify these behaviors and provide rich, personalized feedback to the PMs about how their performance across many such dimensions. And like a coach, it provides personalized feedback through proprietary analytics and visuals that highlight behavioral changes over time and can be used to make suggestions for improving decisions. But it is up to the PMs to decide how to incorporate such feedback. MBAM believes that this type of "Trading Enhancement" is becoming a core differentiation that not only helps develop portfolio managers but also

makes the organizations using such practices more attractive. Financial professionals like what they see as "objective" feedback that can help them improve.

What's more, just as a good mentor learns from the insights of people that she mentors, a machine-learning "Coachbot" learns from the decisions that the empowered human employee makes. 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. The learning process goes in both directions, with humans input leading to a change in the AI's implicit models. At MBAM, for example, if a portfolio manager decides not to trade a highlighted stock because of specific recent company events, he or she will provide an explanation to the system. With this feedback, the system continuously captures data that can be analyzed to provide insights into investment decisions.

Engagement with AI can be greatly enhanced through the choice of interface. If the AI exchanges with the human employee in ways that the latter can relate to and control, it is easier to see it as a safe channel for feedback, predicated on helping rather than assessing the employee's performance. At MBAM, for example, the presentation feedback of the Trading Enhancement tool– its use of visuals, for instance – is personalized to reflect a PM's preferences, which makes the latter more comfortable using the tool.

Finally, as in phase 2 of the organizational journey, involvement of the human employees in the design of the AI is essential. When an AI is a Coach, people will be even more fearful of disempowerment – the AI can easily seem like a competitor as well as a partner, and who wants to feel less intelligent than a machine. For similar reasons, concerns about privacy and autonomy will rear their heads even higher. Working with a Coach requires honesty, and many people may hesitate to be too open with a Coach who may share unflattering data with the folks in HR.

4. Connecting: AI as Part of the Team

Edwin Hutchins, a cognitive anthropologist, developed what is known as the theory of distributed cognition based on his study of ship navigation, which he showed involved the combined effort of sailors, charts, rulers, compasses and the hoey (plotting tool). The theory broadly relates to the concept of extended mind, which posits that cognitive processing, and associated mental acts like 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 no organization has yet adopted, companies would develop a coupled network of humans and machines, in which each part contributes expertise. As AI improves through its interactions with different individual users and their feedback at each level, and as expert users are analyzed and even modeled using AI based on data of their past

decisions and behaviors, we believe that a community of experts (humans *and* machines) will naturally emerge in organizations that have fully integrated AI Coachbots. For example, if we can make it possible for a judge to see how *other* judges – instead of herself – might make a decision (based on data from those other judges) we would, essentially, create a customized collective of experts, both real and virtual, that each judge can invoke – with one click – at the moment of decision.

Although the technology now exists for organizations to create this kind of collective intelligence, this phase of the journey is fraught with challenges. To begin with, any such integration of AI should avoid building in old or new biases. Just as important, such a coupled system must be designed in a way that respects human privacy concerns, because people must be able to trust it as much as they would a human partner. And this is already a pretty big challenge, given the volume of research in organizational behavior and sociology demonstrating how hard it is to build trust among humans.

The best approaches to building trust in people in the workplace context are based on the principles of *Trust through Understanding*, a concept that David Danks and his colleagues at Carnegie Mellon have articulated.. In this model I will trust someone because I understand her values, desires, and intentions -- and given these features, I can know that she has my best interests at heart. Although understanding has historically been a basis for trust building in human-to-human relations, it is potentially well-suited to cultivating the human/AI partnership, as it can generalize to new situations and directly addresses the fear issue, which is usually grounded in a lack of understanding of how an AI works (for a particularly striking example, see the sidebar: *Why COMPAS Lost Its Way.)*.

In building understanding, a particular challenge is defining what an "explanation" means – let alone a "good explanation". Creating an AI people can trust will depend on solving this problem, and it is the focus of a lot of research. For example, one of us, Theos Evgeniou, is working on a project to develop methods that can find so called "counter-factual explanations" of Machine Learning "black boxes". The idea is to explain a particular decision of an AI system (for example, approval of credit for a particular transaction), by identifying a minimal list of the transaction characteristics that drove the decision one way or another - in other words, had any or some of the characteristics been different, the decision of the AI system would have changed (credit for the transaction would not have been approved).

The same author is also exploring questions around the nature of what people perceive as "good explanations" of AI decisions – for example, is an explanation perceived to be better if it is presented in terms of a logical combination of features (for example, "the transaction was approved because it had X,Y,Z characteristics") or when it is presented relative to other decisions (for instance, "the transaction was approved because it looks like those other ones approved before, and here they are all for you to

check")? As research into what makes AI "explainable" develops, AI systems should become more transparent, thus facilitating trust.

***

Adoption of new technologies has always been a major challenge - and the more impactful a technology, the more challenging its adoption. AI is both easy to use and very useful. This is precisely why great care must be taken to ensure that its design and development is morally responsible – especially as regards, transparency, decision autonomy and privacy - and that it engages the humans who will end up working with it. Otherwise people will quite reasonably fear that that they will end up being constrained – even replaced - by machines that will be taking all sorts of decisions humans have traditionally made in ways that we do not understand. Getting past these fears to create a trusting relationship with AI is the key adoption challenge. In the incremental, people-sensitive organizational journey we have described in these pages, it is humans who get to determine the ground rules for the design and implementation of AI as it evolves. And with a morally responsible design that engages human employees, AI might become a true partner in the workplace collective – bringing its ability to process varied data in large volumes and in a consistent manner very rapidly to enhance human intuition and creativity, which in their turn teach the machine.

Sidebar: When AI Loses Its Way

In 2016, the investigative magazine *Pro Publica* wrote an exposé of a risk prediction AI program known as "COMPAS", which judges in South Florida use to determine a defendants' risk of re-offence within a pre specified time period..

The algorithm underlying COMPAS is held as a trade secret by its manufacturer, Northpointe, which means that we do not know the method by which COMPAS generates its predictions, nor do we have access to the data the algoirthm is trained on. **N**ot only can we not explain why COMPAS produces its predictions, but we cannot even inquire into its rationale. So when *Pro Publica* reported that the algorithm produces disparite outcomes across race, COMPAS immediately became the hallmark example for why people cannot trust AI.

If businesses want their employees to adopt, use and ultimately trust, AI systems, it will be important to open up the black box, as much as possible, from a legal perspective, to employees and stakeholders who are expected to engage with it. As Richard Socher, the Chief Scientist of Salesforce puts it, "If businesses use AI to make predictions, they owe humans an explanation as to how the decisions are made".

Exhibit 1: An AI Organizational Journey