

Director's Cut: How Boards Can Help Ensure the Responsible Use of AI

THE QUESTION FOR BOARDS AND MANAGEMENT IS NOT WHETHER TO USE ARTIFICIAL INTELLIGENCE, BUT HOW TO ENSURE IT IS USED RESPONSIBLY

By Bob Doris, Diana Wagner, and Herbert Winokur

Artificial intelligence (AI) is being deployed rapidly by companies in almost every sector of the economy. Though powerful, AI presents new and different challenges to corporate management teams and to boards. Consider how you might have responded to the following “nightmare” incidents that illustrate the potential social and physical harms of advanced AI:

You're a member of the board of a health-care company that's testing a bot based on the latest Generative Pre-trained Transformer 3, or GPT-3, natural language processing algorithm. The system promises

to lower the cost and increase the accuracy of initial patient intake interviews. A reporter calls about why depressed patients using an experimental version of the bot are being advised to kill themselves. Nabla, a health-care startup in Paris, encountered this and other problems in attempting to develop such a bot. The company did note that OpenAI, a developer of GPT-3, issued a general warning against use of the algorithm in “high stakes” health-care settings.

As the CEO of a regional bank, you're surprised one morning by a message from the head of the local chapter of the NAACP, someone

GETTY IMAGES



you've enjoyed working with on several community projects and with whom you've developed a personal relationship. He's upset about a media report that minority applicants are being denied home loans at almost double the rate of white people despite having similar credit scores. Before returning the call, you check with your vice president of mortgage banking. Her response: "The problem is we're subject to a 'secret' algorithm that Fannie Mae and Freddie Mac developed. I don't know exactly how it works, but it does seem to me that it's hard on minorities." This incident is based on an extensive study published

in 2021 by The Markup; while our regional bank is fictional, the debate about mortgage lending algorithms is a serious one.

Or perhaps you encounter a situation like this: At Facebook in mid-2021, the social media platform's "Keep seeing videos about . . ." feature asked people who viewed a video of a group of Black men in an altercation with a white man, "Keep seeing videos about primates?" The video had been posted by the DailyMail.com. After a social media firestorm, a Facebook spokesperson said in a statement, "While we have made improvements to our AI, we know it's not perfect, and we have more progress to make. We apologize to anyone who may have seen these offensive recommendations."

These incidents would've certainly led you, too, to call your public relations head to say, "It's time for a very apologetic press release, but I'm not sure how we can prevent this in the future." They illustrate why companies and their leaders must pay increased attention to the responsible use of AI before any "nightmare" situation occurs.

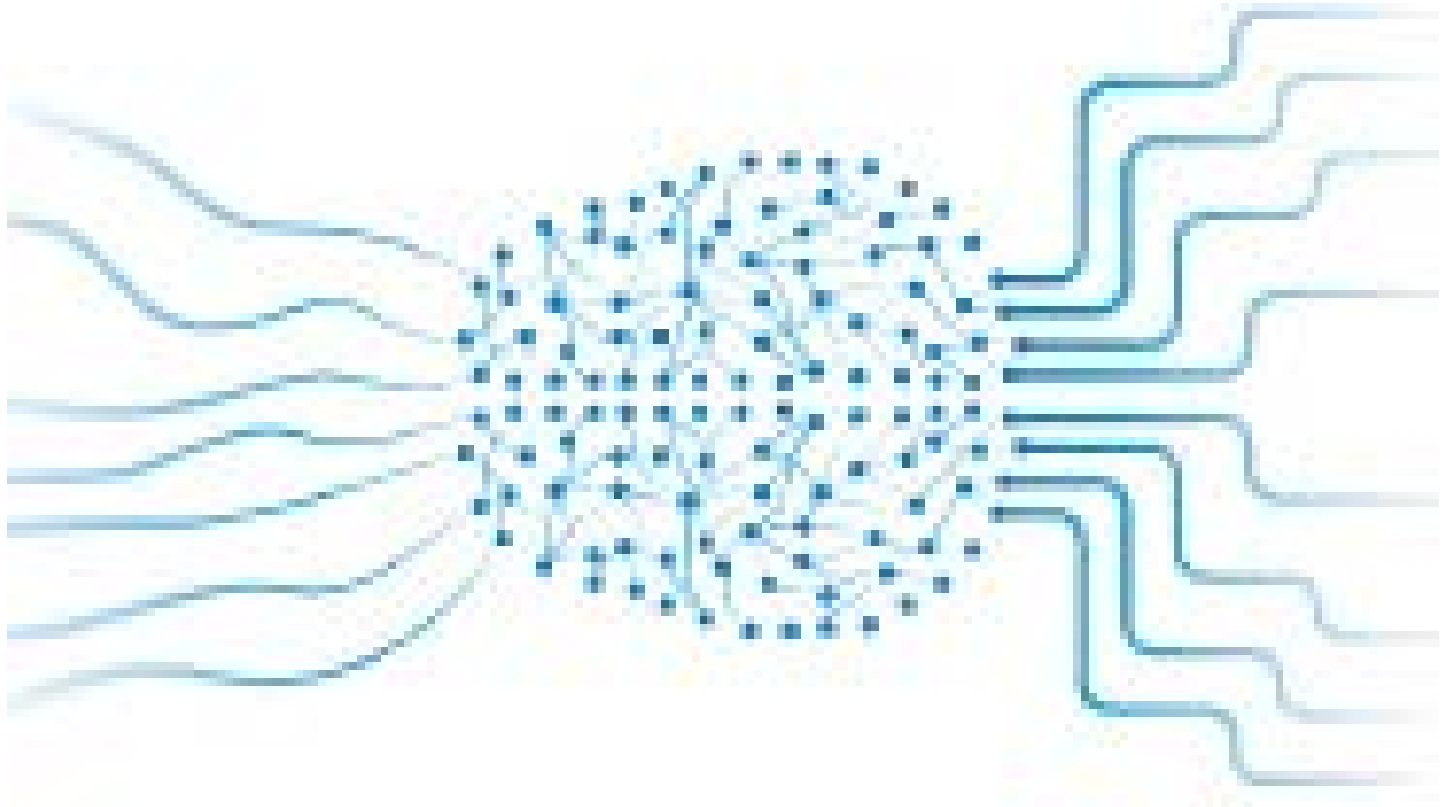
In 2021 we were privileged to chair a task force convened by Frank Doyle, dean of the Harvard Paulson School of Engineering and Applied Sciences (SEAS), to consider ways Harvard University and SEAS could assist the broader community in the responsible use of AI. From our work, we've been able to distill some ideas that we hope will assist corporate directors and management.

THE PROMISE AND PERIL OF AI

Advanced AI has blossomed over the past decade. We're no longer impressed that computers can identify individual faces, anticipate our shopping choices, recognize our handwriting, or read and prepare digests of complicated business contracts. We're not surprised when a call to customer service involves a preliminary chat with a bot that often resolves our issue. No longer does it seem strange that AI systems are able to win consistently against human masters of the ancient games of chess and Go.

The modern use of the term "artificial intelligence" dates to the mid-1950s, when mathematician John McCarthy suggested it to describe the subject of an upcoming conference at Dartmouth College. Most of the conference participants were bullish on the notion that computers, then just beginning to be used widely in business and science, would be able to mimic or even exceed human intelligence. Some participants were certain that we'd see computers exhibiting something approximating general human intelligence within 20 years.

Things didn't work out that way. In fact, AI development proceeded in a halting fashion, through two distinct eras—the first from approximately 1956 to 1979, the second from 1980 to 1993. Each era saw some accomplishments, but generally ended with frustration and disappointment after computer scientists tried first symbol manipulation and then rule-based "expert systems" in pursuit of an intelligent machine.



In the third era, however, computer scientists refocused their efforts from trying to achieve general intelligence to generating solutions to specific problems via probabilistic approaches. By the mid-1990s computer power had become so cheap that it encouraged the use of data- and compute-intensive statistically based machine learning techniques including so-called deep neural networks. (To avoid confusion, artificial intelligence, or AI, refers to the development of machines that exhibit behavior consistent with human intelligence. Machine learning is a sub-field of AI, focusing on developing algorithms that learn from data input to them. Deep neural networks are an important part of machine learning, utilizing algorithms inspired by the neural structure of the human brain.)

The inflection point for the third era, many believe, was a paper published in 2012 by Geoffrey Hinton and his colleagues at the University of Toronto that reported on a deep neural network algorithm that was significantly more effective at accurately identifying images than prior approaches. Since publication of the “cat paper,” as it is often called (image recognition developers are fond of using cat images to illustrate their findings), progress has been dramatic both in image recognition and in many other applications.

Inspired by the structure of the human brain, a deep neural network consists of a set of nodes—that is, artificial neurons—arranged in an input layer that accepts stimuli from the environment, several intermediate “hidden” layers of nodes, and an output layer that emits the result.

To create a deep neural network, a developer constructs a network of multiple layers of interconnected nodes arranged between inputs and outputs. Learning then follows. For “supervised learning,” the network is trained by exposure to an enormous number of inputs (the training set) that have previously been labeled with the correct response. After the training set is exhausted, the model is tested against a comparable data set (the test set), and its accuracy is measured. For “unsupervised learning,” the network is largely left to detect whatever patterns it can among an enormous set of training data. In both cases, as each additional observation in the training set is processed, the weights at each node are adjusted to achieve greater performance.

THE LIMITS OF DEEP LEARNING

Deep learning algorithms have proven very successful in many applications. But since there is no simple rule relating input to output, it’s difficult to state how or why a deep learning algorithm acts as it does in any particular case. Its behavior is not easily explainable: if the algorithm recognizes a picture of a cat, it is only because it has learned that a particular set of features indicates cat, but it is hard for us to know what features the algorithm is attending to. The algorithm may not share our same notion of “cat-ness.”

While current deep learning models have succeeded in providing performance breakthroughs that have driven the reengineering of machine learning pipelines across industries, they have also produced some disappointments, surprises, and, yes, nightmares. Facial recognition—image recognition tying a facial picture to a specific individual—has proven particularly controversial and even, many might say, downright dangerous. Several major technology companies including IBM Corp. and Facebook have discontinued or significantly restricted their facial recognition development programs in response to concerns about privacy and possible misuse of the technology.

One of the most challenging AI application areas has been natural language processing (NLP), that is, equipping a computer to understand and manipulate human language with all its subtlety. The most successful NLP models, including GPT-3 (featured in our first nightmare scenario), make up for the difficulty of this task by using an extremely complicated model with a huge number of parameters (175 billion). GPT-3, like other advanced NLP algorithms, applies a kind of unsupervised learning, thereby making it possible to utilize the immense amount of linguistic data available on the web. The number of words used to train the GPT-3 model amounted to more than 500 billion, most acquired via an extensive web crawl.

NLP models have gotten very good. But they have been criticized for their serious environmental cost (in terms of the energy required to run the computers to construct the model) relative to fairly small increases in accuracy once the model has reached a certain size; for their embodiment of stereotypical prejudices and offensive language (inherent in the use of a training set scraped from the web); for their inability to take into account forward trends in language and social attitude (because the models, by definition, are backward-looking); for their shortcomings with regard to explainability and accountability; and for their complete lack of understanding in the human sense. In a March 2021 paper titled “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” four researchers (including two one-time coleads of Google’s ethical AI team) argued that since current NLP models have no true human understanding, they can introduce serious ethical issues. The more powerful and attractive the algorithms become, the more they become opaque, difficult to predict, and potentially dangerous.

Even the most technically sophisticated companies can be surprised in algorithm development. When Amazon.com developed a recruiting tool to perform the quotidian task of sorting incoming résumés, it utilized as a training set 10 years’ worth of résumés it had received, giving weight to those who had been successfully employed by Amazon. Since these past hires were predominantly male, the algorithm that resulted was woefully

gender-biased. Worse, the tool didn’t do a very good job identifying the better candidates. Amazon scrapped the program before using it.

As Jeannette Wing, a respected computer science researcher at Columbia University, has written, “How . . . can we deliver on the promise of the benefits of AI but address . . . scenarios that have life-critical consequences for people and society?”

OUR BELIEF IS THAT WE CAN
ACHIEVE RESPONSIBLE USE
THROUGH CAREFUL ATTENTION IN
DESIGN, TESTING, AND DEPLOYMENT
TO GET REASONABLY HIGH ASSURANCE
ABOUT HOW A MODEL WILL
OPERATE IN THE REAL WORLD.

THE RESPONSIBLE USE OF AI

As noted earlier, directors will see an increasing use of AI in the companies they oversee. It is up to them to ensure that the AI applications are handled responsibly.

In our task force’s study of the responsible use of AI, we identified five important findings.

1. Most of the ethical principles applicable to AI are familiar; they’ve already arisen in other contexts (especially in the life sciences). Many organizations have proposed ethical principles to govern the development and deployment of AI. Writing in the inaugural issue of the *Harvard Data Science Review*, Luciano Floridi and Josh Cowls, both of the University of Oxford, surveyed several sets of such principles that they then reduced to five basic tenets. Four of the basic tenets aren’t unique to AI:

- **Beneficence.** AI should promote human well-being and dignity and should help sustain the planet.

- **Non-maleficence.** AI should not be used for immoral ends. It should not impinge on privacy, and it should operate securely without giving bad actors the opportunity to co-opt its use.

- **Human autonomy.** Though AI will be used in place of or even to supplant human decision-making, care must be taken not to cede all decision-making to AI. Human autonomy and the ability to reverse bad AI decisions must be preserved.

- **Justice.** AI use should promote prosperity, increase solidarity, and avoid unfairness.

To corporate directors, these tenets should sound familiar. They are very much like the principles motivating the environmental, social, and governance (ESG) movement that is demanding so

much public and shareholder attention and increasingly dominating proxy proposals. In fact, responsible AI itself is seen as an emerging area of focus within ESG.

2. Modern AI introduces a novel set of ethical considerations related to the need to understand and account for the operation of advanced algorithms. Floridi’s and Cowls’ fifth tenet is something they describe as “explicability: enabling the other principles through intelligibility and accountability.”

As discussed above, the operation of trained deep learning AI algorithms can be obscure, even to the developers of the algorithms. This makes it hard to predict what an algorithm will do if conditions change or if it encounters a situation not even remotely in its training set. It also makes it hard to predict when or why an algorithm will produce a result that is wrong, offensive, or downright dangerous.

The problem can become more pronounced when algorithms developed for one task are borrowed or coerced into use for

another. (One suspects that this might have been the case in our third nightmare situation.)

Can we really understand completely and precisely how a particular algorithm works? The answer, at least for some of the very large models, is no. But our belief is that we can achieve responsible use through careful attention in design, testing, and deployment to get reasonably high assurance about how a model will operate in the real world.

3. Culture can affect the application of ethical principles to AI. Ethical principles and their application can vary quite a bit depending on a society’s cultural norms.


A striking example of cultural differences across societies lies in the field of privacy. In the United States, privacy is broadly defined as freedom from government surveillance, but corporate surveillance is largely given a pass (at least currently). In Europe, privacy is defined mostly as consumer freedom from corporate surveillance, with government surveillance given a bit of a pass. In China, government surveillance is promoted as a public good, thus the government lately has been successfully scooping up personal information held by private companies and previously provided by consumers who probably never intended that it be in government hands.

Sensitivity to this is important for any company operating across national borders. Company boards and management will have to think very clearly about how their global strategies and tactics apply in different countries. Other areas where cultural differences are likely to be stark include personal autonomy and the meaning of solidarity and fairness.

4. Responsible AI starts at the bottom . . . Because “third era” AI algorithms can be so powerful for both good and bad, they need to be developed very carefully. Responsible use starts at a very preliminary stage where designers frame the requirements, the use model, and the application of an algorithm. It’s important that the engineering organization engage in a kind of systems thinking that involves careful consideration of where and how an algorithm might be used, and how it might produce unanticipated or undesirable results.

The same care and attention need to be continued through the training phase. Careless selection and preparation of training sets has been at least part of the problem with facial recognition software, among other applications, leading to some truly horrible situations.

The positive news is that technology can be recruited to help engineering teams practice responsible usage. A good example of this is differential privacy, an approach to anonymizing individual data that permits large data sets to be used in research and learning model development without invading individual privacy.



QUESTIONS DIRECTORS SHOULD ASK

There are many things corporate directors can do to engender the responsible use of AI in their organizations and to protect their companies and customers from the consequences of not doing so. Consider asking the following:

- **Where are we in using AI at our company?** What AI is used internally? What AI systems do our customers interact with? Did we develop the technology ourselves or did we acquire it, and if so, from whom?
- **What is our process for testing algorithms we develop or acquire?** How transparent are the providers? How have they tested what they provide to us?
- **What audits of the firm’s AI should be conducted to spot issues of bias or other problems before they become nightmares?**
- **How are we monitoring the regulatory environment relative to AI?**
- **Have our engineering managers taken courses to update themselves on the responsible use of AI?**
- **Have there been any ethically questionable incidents related to AI in our organization?** If so, how did we resolve the issue(s)?

GETTY IMAGES

Several start-up companies are developing digital tools to render AI algorithms more explainable and auditable, complementing efforts underway at most Big Tech companies.

5. . . . And requires strong support from the top. Among Silicon Valley software developers who build software products that will be deployed from the cloud, “move fast and break things” is a popular slogan. Problems can be fixed on the fly unlike, for example, machines in traditional factories that, once built, were difficult to change. Unfortunately, large machine learning algorithms have returned us to an era where things aren’t that easy to fix once deployed. This means it’s incumbent upon corporate leaders to set a clear objective to use the technology responsibly, even if it means moving more slowly.

There are various ways in which companies can address this. At minimum, there should be a clear internal—and even exter-

LARGE MACHINE LEARNING ALGORITHMS HAVE RETURNED US TO AN ERA WHERE THINGS AREN’T THAT EASY TO FIX ONCE DEPLOYED.

nal—statement of company goals with respect to responsible AI. Every company should support a process to vet algorithms developed in-house and to qualify those licensed from elsewhere before they are deployed. The process should include frequent or even continuous monitoring of algorithms after deployment. To be effective, such a process needs to have “teeth,” including the ability to abort deployment of a new algorithm if necessary. Depending on the business, this could very well involve review by a board-level committee empowered to make decisions that are binding for corporate officers. Otherwise, commercial pressure may lead to embarrassing or downright dangerous lapses.

WHAT’S A DIRECTOR TO DO?

Directors, even those with no technical background, should start by familiarizing themselves with the nature, promises, and perils of advanced AI. At minimum, we’d recommend reading a book or two on the subject. Board members (and all corporate managers, not just engineering managers) should also consider taking an executive or continuing education course on AI, its applications, and its potential problems, with particular focus on responsible use themes.


As directors review their companies’ operations, they should focus on processes that encourage responsible use. Chief among

these is explicability, the need for intelligibility and accountability, as mentioned earlier.

It’s especially important to ensure these processes are taken seriously at all levels of the company, particularly in the engineering organization. Machine learning as a separate academic discipline is not much more than 20 years old; the rise of deep learning occurred only about 10 years ago. So, few engineering managers have ever had formal instruction in the technology, and even fewer have attended a course addressing the responsible use of AI.

Directors can push to have company operations, including hiring and human resources, compliance and control functions, and accounting, aligned with responsible AI. Auditing firms are beginning to see AI as an important focus in assessing risks to the enterprise, as are legal teams, business consultants, and ESG analysts. Board members owe it to their shareholders to demand alignment with this trend.

Directors should familiarize themselves with emerging regulations as well. Broad use of AI is relatively recent, and there is no doubt that we will see increasing regulation in this area. AI used in health care and medical systems is already regulated in terms of efficacy and patient privacy. There has been significant regulatory activity in the past few years in the general privacy area, especially in Europe, and in certain US states, notably California. The European Commission is currently considering a comprehensive legal framework for AI regulation.

AI is here and its use will continue to increase. The question for boards and management is not whether to use it, but how to ensure it is used responsibly. We’re convinced that this is possible. But it will require careful study and thought, the adoption of the right business processes, and the need to work through the sometimes-difficult trade-offs involved. Most of all, it requires a commitment by boards and management to engage their organizations in developing and deploying AI that works for all company stakeholders. 

Bob Doris cofounded and heads Accanto Partners, which invests in seed-stage technology start-ups. Previously, Doris cofounded and served for more than 20 years as CEO and chair of Sonic Solutions, a Nasdaq-listed digital media technology company that merged with Rovi Corp. in 2011. Diana Wagner is a portfolio manager and partner at Capital Group Companies, a 90-year-old investment firm that manages the American Funds mutual fund family. Herbert Winokur is chair and CEO of Capricorn Holdings, a private investment firm. Any views expressed in the above article are solely those of the authors and do not reflect the views held by their employers.