https://hbr.org/2021/09/ai-regulation-is-coming

AI And Machine Learning

AI Regulation Is Coming

by François Candelon, Rodolphe Charme di Carlo, Midas De Bondt, and Theodoros Evgeniou

From the Magazine (September–October 2021)

Summary.

For years public concern about technological risk has focused on the misuse of personal data. But as firms embed more and more artificial intelligence in products and processes, attention is shifting to the potential for bad or biased decisions by algorithms—particularly the complex, evolving kind that diagnose cancers, drive cars, or approve loans. Inevitably, many governments will feel regulation is essential to protect consumers from that risk.

This article explains the moves regulators are most likely to make and the three main challenges businesses need to consider as they adopt and integrate AI. The first is ensuring fairness. That requires evaluating the impact of AI outcomes on people's lives, whether decisions are mechanical or subjective, and how equitably the AI operates across varying markets. The second is transparency. Regulators are very likely to require firms to explain how the software makes decisions, but that often isn't easy to unwind. The third is figuring out how to manage algorithms that learn and adapt; while they may be more accurate, they also can evolve in a dangerous or discriminatory way.

Though AI offers businesses great value, it also increases their strategic risk. Companies need to take an active role in writing the rulebook for algorithms.

For most of the past decade, public concerns about digital technology have focused on the potential abuse of personal data. People were uncomfortable with the way companies could track their movements online, often gathering credit card numbers, addresses, and other critical information. They found it creepy to be followed around the web by ads that had clearly been triggered by their idle searches, and they worried about identity theft and fraud.

Those concerns led to the passage of measures in the United States and Europe guaranteeing internet users some level of control over their personal data and images—most notably, the European Union's 2018 General Data Protection Regulation (GDPR). Of course, those measures didn't end the debate around companies' use of personal data. Some argue that curbing it will hamper the economic performance of Europe and the United States relative to less restrictive countries, notably China, whose digital giants have thrived with the help of ready, lightly regulated access to personal information of all sorts. (Recently, however, the Chinese government has started to limit the digital firms' freedom—as demonstrated by the large fines imposed on Alibaba.) Others point out that there's plenty of evidence that tighter regulation has put smaller European companies at a considerable disadvantage to deeper-pocketed U.S. rivals such as Google and Amazon.

But the debate is entering a new phase. As companies increasingly embed artificial intelligence in their products, services, processes, and decision-making, attention is shifting to how data is used by the software—particularly by complex, evolving algorithms that might diagnose a cancer, drive a car, or approve a loan. The EU, which is again leading the way (in its 2020 white paper "On Artificial Intelligence—A European Approach to Excellence and Trust" and its 2021 proposal for an AI legal framework), considers regulation to be essential to the development of AI tools that consumers can trust.

What will all this mean for companies? We've been researching how to regulate AI algorithms and how to implement AI systems that are based on the key principles underlying the proposed regulatory frameworks, and we've been helping companies across industries launch and scale up

AI-driven initiatives. In the following pages we draw on this work and that of other researchers to explore the three main challenges business leaders face as they integrate AI into their decision-making and processes while trying to ensure that it's safe and trustworthy for customers. We also present a framework to guide executives through those tasks, drawing in part on concepts applied to the management of strategic risks.

Unfair Outcomes: The Risks of Using AI

AI systems that produce biased results have been making headlines. One well-known example is Apple's credit card algorithm, which has been accused of discriminating against women, triggering an investigation by New York's Department of Financial Services.

But the problem crops up in many other guises: for instance, in ubiquitous online advertisement algorithms, which may target viewers by race, religion, or gender, and in Amazon's automated résumé screener, which filtered out female candidates. A recent study published in *Science* showed that risk prediction tools used in health care, which affect millions of people in the United States every year, exhibit significant racial bias. Another study, published in the *Journal of General Internal Medicine,* found that the software used by leading hospitals to prioritize recipients of kidney transplants discriminated against Black patients.

> **AI increases the potential scale of bias: Any flaw could affect millions of people, exposing companies to class-action lawsuits.**

In most cases the problem stems from the data used to train the AI. If that data is biased, then the AI will acquire and may even amplify the bias. When Microsoft used tweets to train a chatbot to interact with Twitter users, for example, it had to take the bot down the day after it went live because of its inflammatory, racist messages. But it's not enough to simply eliminate demographic information such as race or gender from training data, because in some situations that data is needed to *correct* for biases.

In theory, it might be possible to code some concept of fairness into the software, requiring that all outcomes meet certain conditions. Amazon is experimenting with a fairness metric called *conditional demographic disparity,* and other companies are developing similar metrics. But one hurdle is that there is no agreed-upon definition of fairness, nor is it possible to be categorical about the general conditions that determine equitable outcomes. What's more, the stakeholders in any given situation may have very different notions of what constitutes fairness. As a result any attempts to design it into the software will be fraught.

In dealing with biased outcomes, regulators have mostly fallen back on standard antidiscrimination legislation. That's workable as long as there are people who can be held responsible for problematic decisions. But with AI increasingly in the mix, individual accountability is undermined. Worse, AI increases the potential scale of bias: Any flaw could affect millions of people, exposing companies to class-action lawsuits of historic proportions and putting their reputations at risk.

What can executives do to head off such problems?

As a first step, prior to making any decision, they should deepen their understanding of the stakes, by exploring four factors:

The impact of outcomes.

Some algorithms make or affect decisions with direct and important consequences on people's lives. They diagnose medical conditions, for instance, screen candidates for jobs, approve home loans, or recommend jail sentences. In such circumstances it may be wise to avoid using AI or at least subordinate it to human judgment.

The latter approach still requires careful reflection, however. Suppose a judge granted early release to an offender against an AI recommendation and that person then committed a violent crime. The judge would be under pressure to explain why she ignored the AI. Using AI could therefore increase human decision-makers' accountability, which might make people likely to defer to the algorithms more often than they should.

That's not to say that AI doesn't have its uses in high-impact contexts. Organizations relying on human decision-makers will still need to control for unconscious bias among those people, which AI can help reveal. Amazon ultimately decided not to leverage AI as a recruiting tool but rather to use it to detect flaws in its current recruiting approach. The takeaway is that the fairness of algorithms relative to human decision-making needs to be considered when choosing whether to use AI.

The nature and scope of decisions.

Research suggests that the degree of trust in AI varies with the kind of decisions it's used for. When a task is perceived as relatively mechanical and bounded—think optimizing a timetable or analyzing images—software is regarded as at least as trustworthy as humans.

But when decisions are thought to be subjective or the variables change (as in legal sentencing, where offenders' extenuating circumstances may differ), human judgment is trusted more, in part because of people's capacity for empathy. This suggests that companies need to communicate very carefully about the specific nature and scope of decisions they're applying AI to and why it's preferable to human judgment in those situations. This is a fairly straightforward exercise in many contexts, even those with serious consequences. For example, in machine diagnoses of medical scans, people can easily accept the advantage that software trained on billions of well-defined data points has over humans, who can process only a few thousand.

On the other hand, applying AI to make a diagnosis regarding mental health, where factors may be behavioral, hard to define, and case-specific, would probably be inappropriate. It's difficult for people to accept that machines can process highly contextual situations. And even when the critical variables have been accurately identified, the way they differ across populations often isn't fully understood—which brings us to the next factor.

Operational complexity and limits to scale.

An algorithm may not be fair across all geographies and markets. For example, one selecting consumers for discounts may appear to be equitable across the entire U.S. population but still show bias when applied to, say, Manhattan residents if consumer behavior and attitudes in Manhattan don't correspond to national averages and aren't reflected in the algorithm's training. Average statistics can mask discrimination among regions or subpopulations, and avoiding it may require customizing algorithms for each subset. That explains why any regulations aimed at decreasing local or small-group biases are likely to reduce the potential for scale advantages from AI, which is often the motivation for using it in the first place.

Adjusting for variations among markets adds layers to algorithms, pushing up development costs. Customizing products and services for specific markets likewise raises production and monitoring costs significantly. All those variables increase organizational complexity and overhead. If the costs become too great, companies may even abandon some markets. Because of GDPR, for example, certain developers, like Gravity Interactive (the maker of Ragnarok and Dragon Saga games), chose to stop selling their products in the EU for some time. Although most will have found a way to comply with the regulation by now (Dragon Saga was relaunched last May in Europe), the costs incurred and the opportunities lost are important.

Compliance and governance capabilities.

To follow the more stringent AI regulations that are on the horizon (at least in Europe and the United States), companies will need new processes and tools: system audits, documentation and data protocols (for traceability), AI monitoring, and diversity awareness training. A number of companies already test each new AI algorithm across a variety of stakeholders to assess whether its output is aligned with company values and unlikely to raise regulatory concerns.

Google, Microsoft, BMW, and Deutsche Telekom are all developing formal AI policies with commitments to safety, fairness, diversity, and privacy. Some companies, like the Federal Home Loan Mortgage Corporation (Freddie Mac), have even appointed chief ethics officers to oversee the introduction and enforcement of such policies, in many cases supporting them with ethics governance boards.

Transparency: Explaining What Went Wrong

Just like human judgment, AI isn’t infallible. Algorithms will inevitably make some unfair—or even unsafe—decisions.

When people make a mistake, there’s usually an inquiry and an assignment of responsibility, which may impose legal penalties on the decision-maker. That helps the organization or community understand and correct unfair decisions and build trust with its stakeholders. So should we require—and can we even expect—AI to explain its decisions, too?

Regulators are certainly moving in that direction. The GDPR already describes “the right…to obtain an explanation of the decision reached” by algorithms, and the EU has identified explainability as a key factor in increasing trust in AI in its white paper and AI regulation proposal.

But what does it mean to get an explanation for automated decisions, for which our knowledge of cause and effect is often incomplete? It was Aristotle who pointed out that when this is the situation, the ability to explain how results are arrived at can be less important than the ability to reproduce the results and empirically verify their accuracy—something companies can do by comparing AI’s predictions with outcomes.

Business leaders considering AI applications also need to reflect on two factors:

The level of explanation required.

With AI algorithms, explanations can be broadly classified into two groups, suited to different circumstances.

Global explanations are complete explanations for all outcomes of a given process and describe the rules or formulas specifying relationships among input variables. They’re typically required when procedural fairness is important—for example, with decisions about the allocation of resources, because stakeholders need to know in advance how they will be made.

> **Should we require—and can we even expect—AI to explain its decisions? Regulators are certainly moving in that direction.**

Providing a global explanation for an algorithm may seem straightforward: All you have to do is share its formula. However, most people lack the advanced skills in mathematics or computer science needed to understand such a formula, let alone determine whether the relationships specified in it are appropriate. And in the case of machine learning—where AI software creates algorithms to describe apparent relationships between variables in the training data—flaws or biases in that data, not the algorithm, may be the ultimate cause of any problem.

In addition, companies may not even have direct insight into the workings of their algorithms, and responding to regulatory constraints for explanations may require them to look beyond their data and IT departments and perhaps to external experts. Consider that the offerings of large software-as-a-service providers, like Oracle, SAP, and Salesforce, often combine multiple AI components from third-party providers. And their clients sometimes cherry-pick and combine AI-enabled solutions. But all an end product's components and how they combine and interconnect will need to be explainable.

Local explanations offer the rationale behind a specific output—say, why one applicant (or class of applicants) was denied a loan while another was granted one. They're often provided by so-called explainable AI algorithms that have the capacity to tell the recipient of an output the grounds for it. They can be used when individuals need to know only why a certain decision was made about them and do not, or cannot, have access to decisions about others.

Local explanations can take the form of statements that answer the question, What are the key customer characteristics that, had they been different, would have changed the output or decision of the AI? For example, if the only difference between two applicants is that one is 24 and the other is 25, then the explanation would be that the first applicant would have been granted a loan if he'd been older than 24. The trouble here is that the characteristics identified may themselves conceal biases. For example, it may turn out that the applicant's zip code is what makes the difference, with otherwise solid applicants from Black neighborhoods being penalized.

The trade-offs involved.

The most powerful algorithms are inherently opaque. Look at Alibaba's Ant Group in China, whose MYbank unit uses AI to approve small business loans in under three minutes without human intervention. To do this, it combines data from all over the Alibaba ecosystem, including information on sales from its e-commerce platforms, with machine learning to predict default risks and maintain real-time credit ratings.

Because Ant's software uses more than 3,000 data inputs, clearly articulating how it arrives at specific assessments (let alone providing a global explanation) is practically impossible. Many of the most exciting AI applications require algorithmic inputs on a similar scale. Tailored payment terms in B2B markets, insurance underwriting, and self-driving cars are only some of the areas where stringent AI explainability requirements may hamper companies' ability to innovate or grow.

Companies will face challenges introducing a service like Ant's in markets where consumers and regulators highly value individual rights—notably, the European Union and the United States. To deploy such AI, firms will need to be able to explain how an algorithm defines similarities between customers, why certain differences between two prospects may justify different treatments, and why similar customers may get different explanations about the AI.

Expectations for explanations also vary by geography, which presents challenges to global operators. They could simply adopt the most stringent explainability requirements worldwide, but doing so could clearly put them at a disadvantage to local players in some markets. Banks following EU rules would struggle to produce algorithms as accurate as Ant's in predicting the likelihood of borrower defaults and might have to be more rigorous about credit requirements as a consequence. On the other hand, applying multiple explainability standards will most likely be more complex and costly—because a company would, in essence, be creating different algorithms for different markets and would probably have to add more AI to ensure interoperability.

There are, however, some opportunities. Explainability requirements could offer a source of differentiation: Companies that can develop AI algorithms with stronger explanatory capabilities will be in a better position to win the trust of consumers and regulators. That could have strategic consequences. If Citibank, for example, could produce explainable AI for small-business credit that's as powerful as Ant's, it would certainly dominate the EU and U.S. markets, and it might even gain a foothold on Ant's own turf. The ability to communicate the fairness and transparency of offerings' decisions is a potential differentiator for technology companies, too. IBM has developed a product that helps firms do this: Watson OpenScale, an AI-powered data analytics platform for business.

The bottom line is that although requiring AI to provide explanations for its decisions may seem like a good way to improve its fairness and increase stakeholders' trust, it comes at a stiff price—one that may not always be worth paying. In that case the only choice is either to go back to striking a balance between the risks of getting some unfair outcomes and the returns from more-accurate output overall, or to abandon using AI.

Learning and Evolving: A Shifting Terrain

One of the distinctive characteristics of AI is its ability to learn; the more labeled pictures of cows and zebras an image-recognition algorithm is fed, the more likely it is to recognize a cow or a zebra. But there are drawbacks to continuous learning: Although accuracy can improve over time, the same inputs that generated one outcome yesterday could register a different one tomorrow because the algorithm has been changed by the data it received in the interim.

In figuring out how to manage algorithms that evolve—and whether to allow continuous learning in the first place—business leaders should focus on three factors:

Risks and rewards.

Customer attitudes toward evolving AI will probably be determined by a personal risk-return calculus. In insurance pricing, for example, learning algorithms will most likely provide results that are better tailored to customer needs than anything humans could offer, so customers will probably have a relatively high tolerance for that kind of AI. In other contexts, learning might not be a concern at all. AI that generates film or book recommendations, for instance, could quite safely evolve as more data about a customer's purchases and viewing choices came in.

But when the risk and impact of an unfair or negative outcome are high, people are less accepting of evolving AI. Certain kinds of products, like medical devices, could be harmful to their users if they were altered without any oversight. That's why some regulators, notably the U.S. Food and Drug Administration, have authorized the use of only "locked" algorithms—which don't learn every time the product is used and therefore don't change—in them. For such offerings, a company can run two parallel versions of the same algorithm: one used only in R&D that continuously learns, and a locked version for commercial use that is approved by regulators. The commercial version could be replaced at a certain frequency with a new version based on the continuously improving one—after regulatory approval.

Regulators also worry that continuous learning could cause algorithms to discriminate or become unsafe in new, hard-to-detect ways. In products and services with which unfairness is a major concern, you can expect a brighter spotlight on evolvability as well.

Complexity and cost.

Deploying learning AI can add to operational costs. First, companies may find themselves running multiple algorithms across different regions, markets, or contexts, each of which has responded to local data and environments. Organizations may then need to create new sentinel roles and processes to make sure that all these algorithms are operating appropriately and within authorized risk ranges. Chief risk officers may have to expand their mandates to include monitoring autonomous AI processes and assessing the level of legal, financial, reputational, and physical risk the company is willing to take on evolvable AI.

Firms also must balance decentralization against standardized practices that increase the rate of AI learning. Can they build and maintain a global data backbone to power the firm's digital and AI solutions? How ready are their own systems for decentralized storage and processing? How prepared are they to respond to cybersecurity threats? Does production need to shift closer to end customers, or would that expose operations to new risks? Can firms attract enough AI-savvy talent in the right leadership positions in local markets? All those questions must be answered thoughtfully.

Human input.

New data or environmental changes can also cause people to adjust their decisions or even alter their mental models. A recruiting manager, for example, might make different decisions about the same job applicant at two different times if the quality of the competing candidates changes—or even because she's tired the second time around. Since there's no regulation to prevent that from happening, a case could be made that it's permissible for AI to evolve as a result of new data. However, it would take some convincing to win people over to that point of view.

> **Regulators worry that continuous learning could cause algorithms to discriminate or become unsafe in new, hard-to-detect ways.**

What people might accept more easily is AI complemented in a smart way by human decision-making. As described in the 2020 HBR article "A Better Way to Onboard AI" (coauthored by Theodoros Evgeniou), AI systems can be deployed as "coaches"—providing feedback and input to employees (for instance, traders in financial securities at an asset management firm). But it's not a one-way street: Much of the value in the collaboration comes from the feedback that humans give the algorithms. Facebook, in fact, has taken an interesting approach to monitoring and accelerating AI learning with its Dynabench platform. It tasks human experts with looking for ways to trick AI into producing an incorrect or unfair outcome using something called *dynamic adversarial data collection*.

When humans actively enhance AI, they can unlock value fairly quickly. In a recent TED Talk, BCG's Sylvain Duranton described how one clothing retailer saved more than $100 million in just one year with a process that allowed human buyers to input their expertise into AI that predicted clothing trends.

. . .

Given that the growing reliance on AI—particularly machine learning—significantly increases the strategic risks businesses face, companies need to take an active role in writing a rulebook for algorithms. As analytics are applied to decisions like loan approvals or assessments of criminal recidivism, reservations about hidden biases continue to mount. The inherent opacity of the complex programming underlying machine learning is also causing dismay, and concern is rising about whether AI-enabled tools developed for one population can safely make decisions about other populations. Unless all companies—including those not directly involved in AI development—engage early with these challenges, they risk eroding trust in AI-enabled products and triggering unnecessarily restrictive regulation, which would undermine not only business profits but also the potential value AI could offer consumers and society.

A version of this article appeared in the September–October 2021 issue of *Harvard Business Review*.

François Candelon is a managing director and senior partner at the Boston Consulting Group and the global director of the BCG Henderson Institute.

Rodolphe Charme di Carlo is a partner in the Paris office of the Boston Consulting Group.

Midas De Bondt is a project leader in the Brussels office of the Boston Consulting Group.

Theodoros Evgeniou is a professor at INSEAD.