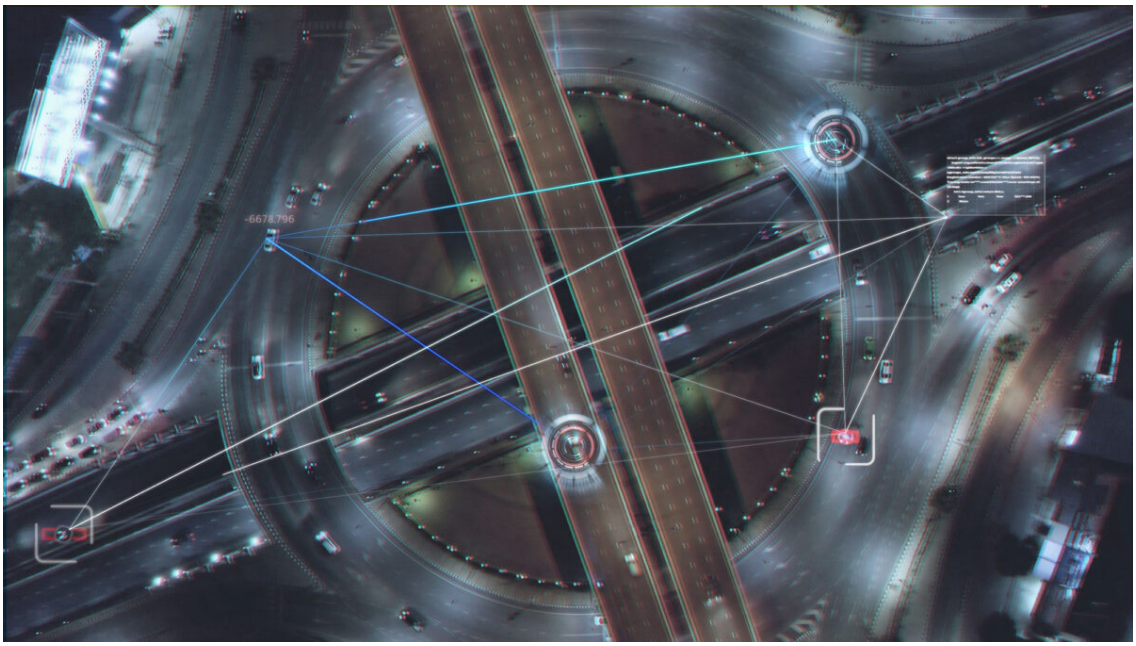
Harvard Business Review

Strategy

A Practical Guide to Building Ethical AI

by Reid Blackman

October 15, 2020



MR.Cole_Photographer/Getty Images

Summary. Companies are quickly learning that AI doesn't just scale solutions — it also scales risk. In this environment, data and AI ethics are business necessities, not academic curiosities. Companies need a clear plan to deal with the ethical quandaries this new tech is... **more**

Companies are leveraging data and artificial intelligence to create scalable solutions — but they're also scaling their reputational, regulatory, and legal risks. For instance, Los Angeles is suing IBM for allegedly misappropriating data it collected with its

ubiquitous weather app. Optum is being investigated by regulators for creating an algorithm that allegedly recommended that doctors and nurses pay more attention to white patients than to sicker black patients. Goldman Sachs is being investigated by regulators for using an AI algorithm that allegedly discriminated against women by granting larger credit limits to men than women on their Apple cards. Facebook infamously granted Cambridge Analytica, a political firm, access to the personal data of more than 50 million users.

Just a few years ago discussions of "data ethics" and "AI ethics" were reserved for nonprofit organizations and academics. Today the biggest tech companies in the world — Microsoft, Facebook, Twitter, Google, and more — are putting together fast-growing teams to tackle the ethical problems that arise from the widespread collection, analysis, and use of massive troves of data, particularly when that data is used to train machine learning models, aka AI.

INSIGHT CENTER

AI and Equality

Designing systems that are fair for all.

These companies are investing in answers to once esoteric ethical questions because they've realized one simple truth: failing to operationalize data and AI ethics is a threat to the bottom line. Missing the mark can expose companies to reputational, regulatory, and legal risks, but that's not the half of it. Failing to operationalize data and AI ethics leads to wasted resources, inefficiencies in product development and deployment, and even an inability to use data to train AI models at all. For example, Amazon engineers reportedly spent years working on AI hiring software, but eventually scrapped the program because they couldn't figure out how to create a model that doesn't systematically discriminate against women. Sidewalk Labs, a subsidiary of Google, faced massive backlash by citizens and local government officials over their plans to build an IoT-

fueled "smart city" within Toronto due to a lack of clear ethical standards for the project's data handling. The company ultimately scrapped the project at a loss of two years of work and USD $50 million.

Despite the costs of getting it wrong, most companies grapple with data and AI ethics through ad-hoc discussions on a per-product basis. With no clear protocol in place on how to identify, evaluate, and mitigate the risks, teams end up either overlooking risks, scrambling to solve issues as they come up, or crossing their fingers in the hope that the problem will resolve itself. When companies have attempted to tackle the issue at scale, they've tended to implement strict, imprecise, and overly broad policies that lead to false positives in risk identification and stymied production. These problems grow by orders of magnitude when you introduce third-party vendors, who may or may not be thinking about these questions at all.

Companies need a plan for mitigating risk — how to use data and develop AI products without falling into ethical pitfalls along the way. Just like other risk-management strategies, an operationalized approach to data and AI ethics must systematically and exhaustively identify ethical risks throughout the organization, from IT to HR to marketing to product and beyond.

What Not to Do

Putting the larger tech companies to the side, there are three standard approaches to data and AI ethical risk mitigation, none of which bear fruit.

First, there is the **academic approach**. Academics — and I speak from 15 years of experience as a former professor of philosophy — are fantastic at rigorous and systematic inquiry. Those academics who are ethicists (typically found in philosophy departments) are

adept at spotting ethical problems, their sources, and how to think through them. But while academic ethicists might seem like a perfect match, given the need for systematic identification and mitigation of ethical risks, they unfortunately tend to ask different questions than businesses. For the most part, academics ask, "Should we do this? Would it be good for society overall? Does it conduce to human flourishing?" Businesses, on the other hand, tend to ask, "Given that we are going to do this, how can we do it without making ourselves vulnerable to ethical risks?"

The result is academic treatments that do not speak to the highly particular, concrete uses of data and AI. This translates to the absence of clear directives to the developers on the ground and the senior leaders who need to identify and choose among a set of risk mitigation strategies.

Next, is the **"on-the-ground" approach**. Within businesses those asking the questions are standardly enthusiastic engineers, data scientists, and product managers. They know to ask the business-relevant risk-related questions precisely because they are the ones making the products to achieve particular business goals. What they lack, however, is the kind of training that academics receive. As a result, they do not have the skill, knowledge, and experience to answer ethical questions systematically, exhaustively, and efficiently. They also lack a critical ingredient: institutional support.

Finally, there are companies (not to mention countries) rolling out **high-level AI ethics principles**. Google and Microsoft, for instance, trumpeted their principles years ago. The difficulty comes in operationalizing those principles. What, exactly, does it mean to be for "fairness?" What are engineers to do when confronted with the dozens of definitions and accompanying metrics for fairness in the computer science literature? Which metric is the right one in any given case, and who makes that

judgment? For most companies — including those tech companies who are actively trying to solve the problem — there are no clear answers to these questions. Indeed, seeming coalescence around a shared set of abstract values actually obscures widespread misalignment.

How to Operationalize Data and AI Ethics

AI ethics does not come in a box. Given the varying values of companies across dozens of industries, a data and AI ethics program must be tailored to the specific business and regulatory needs that are relevant to the company. However, here are seven steps towards building a customized, operationalized, scalable, and sustainable data and AI ethics program.

1. Identify existing infrastructure that a data and AI ethics program can leverage. The key to a successful creation of a data and AI ethics program is using the power and authority of existing infrastructure, such as a data governance board that convenes to discuss privacy, cyber, compliance, and other data-related risks. This allows concerns from those "on the ground" (e.g., product owners and managers) to bubble up and, when necessary, they can in turn elevate key concerns to relevant executives. Governance board buy in works for a few reasons: 1) the executive level sets the tone for how seriously employees will take these issues, 2) a data and AI ethics strategy needs to dovetail with the general data and AI strategy, which is devised at the executive level, and 3) protecting the brand from reputational, regulatory, and legal risk is ultimately a C-suite responsibility, and they need to be alerted when high stakes issues arise.

If such a body does not exist then companies can create one — an ethics council or committee, for example — with ethics-adjacent personnel, such as those in cyber, risk and compliance, privacy, and analytics. It may also be advisable to include external subject matter experts, including ethicists.

2. Create a data and AI ethical risk framework that is tailored to your industry. A good framework comprises, at a minimum, an articulation of the ethical standards — including the ethical nightmares — of the company, an identification of the relevant external and internal stakeholders, a recommended governance structure, and an articulation of how that structure will be maintained in the face of changing personnel and circumstances. It is important to establish KPIs and a quality assurance program to measure the continued effectiveness of the tactics carrying out your strategy.

A robust framework also makes clear how ethical risk mitigation is built into operations. For instance, it should identify the ethical standards data collectors, product developers, and product managers and owners must adhere to. It should also articulate a clear process by which ethical concerns are elevated to more senior leadership or to an ethics committee. All companies should ask whether there are processes in place that vet for biased algorithms, privacy violations, and unexplainable outputs.

Still, frameworks also need to be tailored to a company's industry. In finance, it is important to think about how digital identities are determined and how international transactions can be ethically safe. In health care there will need to be extra protections built around privacy, particularly as AI enables the development of precision medicine. In the retail space, where recommendation engines loom large, it is important to develop methods to detect and mitigate associative bias, where recommendations flow from stereotypical and sometimes offensive associations with various populations.

3. Change how you think about ethics by taking cues from the successes in health care. Many senior leaders describe ethics in general — and data and AI ethics in particular — as "squishy" or "fuzzy," and argue it is not sufficiently "concrete" to be actionable.

Leaders should take inspiration from health care, an industry that has been systematically focused on ethical risk mitigation since at least the 1970s. Key concerns about what constitutes privacy, self-determination, and informed consent, for example, have been explored deeply by medical ethicists, health care practitioners, regulators, and lawyers. Those insights can be transferred to many ethical dilemmas around consumer data privacy and control.

For instance, companies attest to respect the users of their products, but what does that mean in practice? In health care, an essential requirement of demonstrating respect for patients is that they are treated only after granting their informed consent — understood to include consent that, at a minimum, does not result from lies, manipulation, or communications in words the patient cannot understand, such as impenetrable legalese or Latin medical terms. These same kinds of requirements can be brought to bear on how people's data is collected, used, and shared. Ensuring that users are not only informed of how their data is being used, but also that they are informed early on and in a way that makes comprehension likely (for instance, by not burying the information in a long legal document), is one easy lesson to take from health care. The more general lesson is to break down big ethical concepts like privacy, bias, and explainability into infrastructure, process, and practice that realize those values.

4. Optimize guidance and tools for product managers. While your framework provides high-level guidance, it's essential that guidance at the product level is granular. Take, for instance, the oft-lauded value of explainability in AI, a highly valued feature of ML models that will likely be part of your framework. Standard machine-learning algorithms engage in pattern recognition too unwieldy for humans to grasp. But it is common — particularly when the outputs of the AI are potentially life-altering — to want or demand explanations for AI outputs. The problem is that there is often a tension between making outputs explainable, on the

one hand, and making the outputs (e.g. predictions) accurate, on the other.

Product managers need to know how to make that tradeoff, and customized tools should be developed to help product managers make those decisions. For example, companies can create a tool by which project managers can evaluate the importance of explainability for a given product. If explainability is desirable because it helps to ferret out bias in an algorithm, but biased outputs are not a concern for this particular ML application, then that downgrades the importance of explainability relative to accuracy. If the outputs fall under regulations that require explanations — for instance, regulations in the banking industry that require banks to explain why someone has been turned down for a loan — then explainability will be imperative. The same goes for other relevant values, e.g. which, if any, of the dozens of metrics to use when determining whether a product delivers fair or equitable outputs.

5. Build organizational awareness. Ten years ago, corporations scarcely paid attention to cyber risks, but they certainly do now, and employees are expected to have a grasp of some of those risks. Anyone who touches data or AI products — be they in HR, marketing, or operations — should understand the company's data and AI ethics framework. Creating a culture in which a data and AI ethics strategy can be successfully deployed and maintained requires educating and upskilling employees, and empowering them to raise important questions at crucial junctures and raise key concerns to the appropriate deliberative body. Throughout this process, it's important to clearly articulate why data and AI ethics matters to the organization in a way that demonstrates the commitment is not merely part of a public relations campaign.

6. Formally and informally incentivize employees to play a

role in identifying AI ethical risks. As we've learned from numerous infamous examples, ethical standards are compromised when people are financially incentivized to act unethically. Similarly, failing to financially incentivize ethical actions can lead to them being deprioritized. A company's values are partly determined by how it directs financial resources. When employees don't see a budget behind scaling and maintaining a strong data and AI ethics program, they will turn their attention to what moves them forward in their career. Rewarding people for their efforts in promoting a data ethics program is essential.

7. Monitor impacts and engage stakeholders. Creating organizational awareness, ethics committees, informed product managers owners, engineers, and data collectors is all part of the development and, ideally, procurement process. But due to limited resources, time, and a general failure to imagine all the ways things can go wrong, it is important to monitor the impacts of the data and AI products that are on the market. A car can be built with air bags and crumple zones, but that doesn't mean it's safe to drive it at 100 mph down a side street. Similarly, AI products can be ethically developed but unethically deployed. There is both qualitative and quantitative research to be done here, including especially engaging stakeholders to determine how the product has affected them. Indeed, in the ideal scenario, relevant stakeholders are identified early in the development process and incorporated into an articulation of what the product does and does not do.

Operationalizing data and AI ethics is not an easy task. It requires buy-in from senior leadership and cross-functional collaboration. Companies that make the investment, however, will not only see mitigated risk but also more efficient adoption of the technologies they need to forge ahead. And finally, they'll be exactly what their clients, consumers, and employees are looking for: trustworthy.

Reid Blackman is the author of *Ethical*

Machines (Harvard Business Review Press, 2022), the host of a podcast by the same name, and the founder and CEO of Virtue, a digital ethical risk consultancy. He advises the government of Canada on federal AI regulations and corporations on how to implement digital ethical risk programs. He has been a senior adviser to the Deloitte AI Institute, served on Ernst & Young's AI Advisory Board, and volunteers as the chief ethics officer to the nonprofit Government Blockchain Association. Previously he was a professor of philosophy at Colgate University and the University of North Carolina, Chapel Hill.

Recommended For You

Why You Need an AI Ethics Committee

Ethics and AI: 3 Conversations Companies Need to Have

If Your Company Uses AI, It Needs an Institutional Review Board

PODCAST
First He Saved Unilever. Now He Wants to Save Capitalism.