ChangeThis

208.03

ETHICAL ISSUES & AUTOMATION

Dr. Bhaskar Ghosh, Rajendra Prasad, and Gayathri Pallail

Many enterprises today—commercial enterprises, nonprofit organizations, and government agencies—are experimenting with and investing in intelligent automation. Yet they are not achieving the full value of these potentially transformative capabilities. Those that do make the leap to scale will gain an enduring performance advantage—an automation edge.

Resolving to gain that edge, however, is as much a question of embracing the right management principles as it is about investing in the right technologies. The convenient shorthand we offer is that automation solutions must be relevant, resilient, and responsible. As guideposts for decision-making these may sound unsurprising—even obvious, perhaps. But we refer to them as ideals because in most organizations, there is still a long way to go in all three respects.

Here, we want to focus on the third major principle that must guide future human-and-machine pairing, **responsibility:** a much greater level of attention to the ethical issues raised by these powerful new tools. Machines, left to their own devices, have no ethos to guide them. As for people, there is nothing more differentiating about the human race than its capacity for moral action—and nothing more disappointing than someone setting their morals aside.

Among those who are wary of the advancing capabilities of artificial intelligence, the worst fears are about irresponsible applications and dangerous misuses of it. Beyond the threats of job destruction, the possibilities range from “deep fakes” spreading misinformation to unprecedented surveillance tools empowering totalitarian repression.

In China, for example, Baidu's DeepVoice software can "clone" a voice based on just 3.7 seconds of audio input from the original.[1] The appealing side of this is that, for example, you could choose to have a novel read to you in any voice, from the author's own to your late grandmother's. You could have your favorite actor telling you the next turn to take as you use a navigation app. At the same time, the potential for abuse is obvious.

Ethics will come to the fore with increasing violations of privacy, biases in decision-making, and lack of control over automated systems and robots. And solutions to ethics issues will have to be scalable, as intelligent automation becomes ever more widely applied, more deeply embedded in customer solutions, and more responsible for decisions that affect lives—such as medical diagnoses, government benefit payments, and mortgage approvals. Legal scholars are already busy identifying the issues that will inevitably arise, and proposing frameworks and principles for dealing with them consistently.[2] Four of these principles are certain to remain pillars of "responsible automation." To avoid causing reckless or heedless damage, solutions will have to be unbiased, transparent, controllable, and protected.

Ethics will come to the fore with increasing violations of privacy, biases in decision-making, and lack of control over automated systems and robots.

UNBIASED DECISIONS

An intelligent automation solution is only as good as its data. Particularly when artificial intelligence is embedded in an intelligent automation solution, it becomes clear how the characteristics of the data used to train an AI model influence the recommendations and decisions it produces. An infamous example was an AI-powered chatbot named Tay, created by a team of researchers and given its own social media account. Of course, the bot wasn't coded to be racist, a *TechCrunch* journalist reported—it just learned from the other accounts it interacted with. "And naturally, given that this is the Internet, one of the first things online users taught Tay was how to be racist, and how to spout back ill-informed or inflammatory political opinions."[3]

More often, an enterprise AI solution is trained on a data set largely consisting of proprietary records of customer profiles, interactions, and transactions. But this, too, can easily allow the tool to spot and act on patterns that drive its recommendations and decisions in ways that cause harm to groups of people. Sometimes this is because the nature of data is always that it is historical—it may not represent the current reality well enough to predict likely outcomes in the future. Sometimes it is because the output of the decision-making process itself adds up to a pattern of discrimination that the company never intended—and which creates legal exposure for the company and often mortifying reputational damage.

Already there have been widely condemned examples of racial and gender bias patterns in business AI use.

Vivienne Ming describes in the *Financial Times*, for example, the terrible surprise that a technology company's human resources department had when they tried to use AI to find the best candidates in the mountain of resumes the company receives. Having trained it on the backgrounds of people who had succeeded in challenging technology jobs in the past, they were dismayed when the candidates it selected were overwhelmingly male. Avoiding such an outcome isn't just a matter of having unbiased AI developers, or even better data, Ming insists. There also have to be people and processes focused on "de-biasing" the data (as the company learned, and quickly put in place).[4]

One of the oldest and truest sayings in the information systems world is "garbage in, garbage out" (GIGO). The output of a system will never be useful if the input was fatally flawed from the start. In the realm of AI, there is a version of GIGO we might call BIBO: Bias in, bias out. When these high-profile embarrassments occur for companies and the work is done to unpack how the machinery could have come up with such results, usually the problem can be traced to the base data itself. Since most of the data points are collecting the actions of humans in real-world scenarios—choices they have made, experiences they have had—the human biases, often implicit and acting on a sub-conscious level, end up tainting the AI models' training.

The effect, then, of the AI solution's application is to amplify and magnify the bias. To the extent possible, it is imperative to identify and eliminate the potential for bias before training the AI model. Even if care is taken to do this, however, there must also be ex post facto assessments to flag the problem if outputs are unfairly weighted for or against people with characteristics that really should not matter to the decision at hand.

There are statistical methods that can be adopted to minimize the data bias. There are also good management processes to reduce bias in an AI model. Here are several tips to keep in mind:

- Identify the bias vectors an AI model is exposed to, among them the ethical, social, political, and historical biases that could infect its training.

- Gather perspectives from experts in varied fields about possible negative scenarios that should be anticipated and avoided. Establish metrics to monitor any movement toward or away from these scenarios.

- Vet the data for its inclusiveness and ability to represent the full diversity of the population, across gender, race, ethnicity, religion, ideology, and other social lines.

- Know the data so thoroughly that it's easy to focus on and quickly fix any problematic tags. Structure the data well.

- Identify and neutralize any factors that can be foreseen or identified in practice as driving outcomes in prejudiced and discriminatory directions.

- Continuously analyze performance and outcomes, and incorporate feedback from users.

Of course, the point is not to eliminate all bias from AI, even if that were practically possible. AI's ability to find and act on patterns that humans' own cognitive biases or limited perspective have kept hidden is central to its value.

The problem emerges when the biases that AI tends toward are socially unjust based on historical patterns that society has rejected. The infamous examples of bias in AI have all had to do with situations where the AI's output is perpetuating a discriminatory stance, and using the AI would actively undermine the progress that humanity is trying to make. At a level higher than the previous bullet points, we can offer a handful of management approaches that we have seen organizations use to eliminate bias in their intelligent automation solutions.

BREAK MAJOR BUSINESS PROBLEM AREAS DOWN INTO MANAGEABLE SEGMENTS

Addressing any major area that is inherently problematic in terms of bias forces a team to imagine a wide range of scenarios and anticipate all the ways that stereotypes could influence algorithms across multiple system modules. Scoping an initial target for a solution more narrowly will make it easier to comprehend its complexity and trace the origins of any unintended consequences, while also raising the likelihood that it will perform efficiently and satisfy the need it was built to serve.

AI's ability to find and act on patterns that humans' own cognitive biases or limited perspective have kept hidden is central to its value.

EMPATHIZE WITH END USERS IN DEVELOPMENT

A good way to make a model more user-friendly is to adopt the points of view of many kinds of people who will use it and be affected by it. Do this personally and as a team, actively role-playing and challenging decisions as a devil's advocate, to avoid being surprised by an AI bias that end users might face as they interact with the model.

SUBJECT THE MODEL TO DIVERSE TESTERS

At the launch phase of a model, expose it to as diverse a group as possible, and to experts in AI bias avoidance. Understand that an AI response which one person considers to be purely rational and neutral can be perceived by another set of people as deeply biased. The more viewpoints in the room, the more problems can be avoided. As well as detecting potential for bias in this model, the others' input will guide better planning for more bias-free models in the future.

CREATE FEEDBACK SYSTEMS THAT RECOGNIZE DIVERSITY

If a model has a diverse set of people interacting with it and being affected by it, then it would be a lost opportunity if feedback systems managed to erase that diversity by emphasizing only the model's overall performance. Introduce ways of getting and considering more nuanced feedback. More fine-grained information will reveal if certain segments of people are having experiences different from the majority (and possibly becoming frustrated that their voices are not being heard).

ESTABLISH A SYSTEM OF CONTINUOUS IMPROVEMENT BASED ON FEEDBACK

Have a process in place by which the model will continue to be tweaked based on feedback received. This will allow the model to continuously move toward the ideal of unbiased, accurate performance. Remember that, once deployed, AI models are exposed to various scenarios and circumstances. Even with the highest level of due diligence we cannot guarantee complete elimination of bias.

CREATE TRANSPARENT SYSTEMS

In his revelatory book, *Principles*, Bridgewater Associates founder Ray Dalio devotes some words to describing his company's pioneering use of automation in investment decision-making—a process where, he says, "the machine does most of the work and we interact with it in a quality way." He writes:

"One of the great things about algorithmic decision making is that it focuses people on cause-effect relationships and, in that way, helps foster a real idea meritocracy. When everyone can see the criteria algorithms use and have a hand in developing them, they can all agree that the system is fair and trust the computer to look at the evidence, make the right assessments about people, and assign them the right authorities. The algorithms are essentially principles in action on a continuous basis."[5]

What Dalio is describing is what more designers of automation solutions should aim for. In most cases, automation and intelligence technology are more like a black box. People may see the inputs and they can easily see the results, but they don't have a clue about the weightings and calculations of the algorithms being used or the logic of the decision-making process.

This brings us to the concept of explainable AI, often referred to as XAI.[6] These are systems which can explain the steps in their decision-making process, the alternatives involved, and how they arrived at an output. This gives a fair idea of the behavioral patterns of the technology and how its future evolution paths can be mapped. As a result, the technology becomes more transparent and has a built-in trust factor.

In today's world, AI algorithms are being applied to highly sensitive territory, such as in legal affairs and medical diagnostics. In areas like these, where the costs of mistakes can be very high, automated decision-making will be subjected to even greater scrutiny in years to come. Imagine the extreme, hypothetical situation of AI standing in for a judge and jury. Once the AI finds a defendant guilty and renders its decision as to punishment, what if the defendant files an appeal, asking a higher court to reverse the decision? In that case, the higher court's first step is to investigate the decision-making process used in the first round. But if the lower court's AI system is a typical black box, that investigation goes nowhere, and the higher court must reconsider the case by its own logic. Now imagine that, using its own preferred logic, the higher court overturns the verdict. With that reversal, the credibility of the AI system is damaged; it got the answer wrong evidently, and there is no clear way to adjust it to get things more right in the future. Growing distrust eventually leads to abandonment of the AI system.

Automated decision-making will be subjected to even greater scrutiny in years to come.

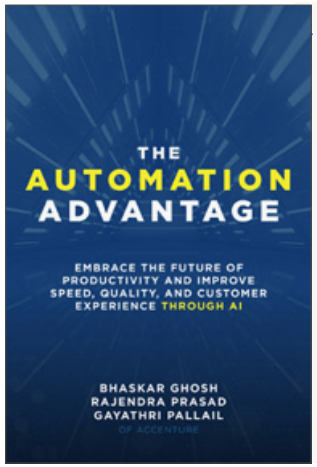
By analogy, this kind of situation is easier to imagine in the realm of medical diagnostics, where AI is already making significant inroads. Any diagnosis indicated by an AI system is carefully reviewed by a physician before a therapy or intervention is prescribed to the patient. If the experienced judgment of the human doctor does not match the data-driven diagnosis of the AI, the doctor gets the final say—but the doctor will want to review why the AI came to a different conclusion. If the AI systems are opaque, doctors will go with their own judgment, both because they have more confidence in it and because they can explain the reasoning to their patients. Again, having done so once, a doctor is less likely in a similar situation to use the AI system for the initial diagnostic having lost faith in its usefulness. No change in behavior will take hold, and whomever decided to invest in this AI diagnostic system will see the implementation fail.

There can be multiple such situations where an explanation might be needed from an AI system to validate its decisions. A car insurance company will probably want to review the decision-making process of a driverless car before settling the insurance claim. A university's admissions department might be under scrutiny if it suddenly rejects the applications of a group of students who all happen to be of a particular ethnic origin. The decision making might have been entirely on merit considerations, but regulatory authorities will probably want to take a deeper look at the decision process. As such, AI systems of the future cannot remain opaque if they have to garner trust from their human users.

Without trust, the implementation of AI systems will remain incomplete and will not be able to reach its full potential.

Excerpt from *The Automation Advantage: Embrace the Future of Productivity and Improve Speed, Quality, and Customer Experience Through AI* by Dr. Bhaskar Ghosh, Rajendra Prasad, and Gayathri Pallail pp. 209-210 and 218-226 (McGraw Hill, January 2022).

Info



Ready to dig deeper into the book?
Buy a copy of
The Automation Advantage.

Want copies for your organization or for an event? We can help:
customerservice@porchlightbooks.com 800-236-7323

ABOUT THE AUTHORS

Dr. Bhaskar Ghosh, PhD, (Bangalore, India) is Accenture's Chief Strategy Officer. In this role, he directs the company's strategy and investments, including ventures and acquisitions, all offerings and assets, and Accenture Research.

Rajendra Prasad (Basking Ridge, NJ) is the Global Automation Lead at Accenture and heads a team that has helped organizations across the globe successfully implement and scale their intelligent automation transformations. He has spent over two decades innovating and defining frameworks for driving efficiency and managing change in software engineering.

Gayathri Pallail (Bangalore, India) is a Managing Director for automation strategy and deployment at Accenture. She has implemented enterprise-wide automation-based solutions and successful change management to enable seamless adoption for over 500 clients across industries.

SHARE THIS

Pass along a copy of this manifesto to others.

SUBSCRIBE

Sign up for e-news to learn when our latest manifestos are available.

Porchlight

Curated and edited by the people of Porchlight, ChangeThis is a vehicle for big ideas to spread. Keep up with the latest book releases and ideas at porchlightbooks.com.

This document was created on January 19, 2022 and is based on the best information available at that time.

Endnotes

1. Bernard Marr, "Artificial Intelligence Can Now Copy Your Voice: What Does That Mean For Humans?," *Forbes*, May 6, 2019.
2. Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation," *Harvard Journal of Law & Technology* 31, no. 2 (Spring 2018).
3. Sarah Perez, "Microsoft Silences Its New A.I. Bot Tay, after Twitter Users Teach It Racism," *TechCrunch*, March 24, 2016.
4. Vivienne Ming, "Human Insight Remains Essential to Beat the Bias of Algorithms," *Financial Times*, December 3, 2019.
5. Ray Dalio, *Principles: Life and Work* (New York: Simon and Schuster, 2017), 100-101.
6. Ron Schmelzer, "Understanding Explainable AI," *Forbes*, July 23, 2019.