Harvard Business Review

Business Ethics

AI's Trust Problem

by Bhaskar Chakravorti

May 03, 2024

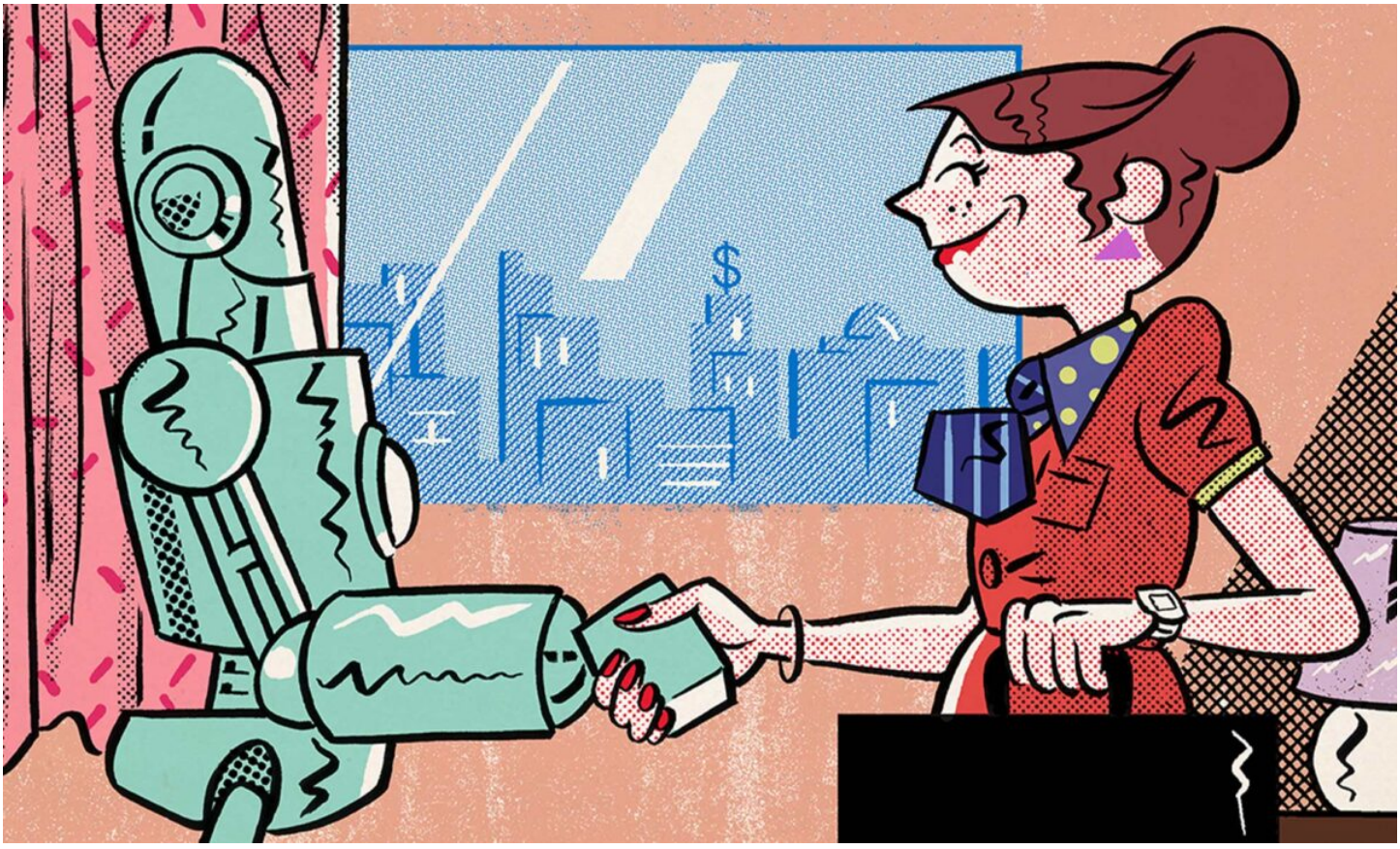


Illustration by Gabriel Corbera

Summary. As AI becomes more powerful, it faces a major trust problem. Consider 12 leading concerns: disinformation, safety and security, the black box problem, ethical concerns, bias, instability, hallucinations in LLMs, unknown unknowns, potential job losses and social inequalities, environmental impact, industry concentration, and state overreach. Each of these issues is complex — and not easy to solve. But there is one consistent approach to addressing the trust gap:

training, empowering, and including humans to manage AI tools. **close**

With tens of billions invested in AI last year and leading players such as OpenAI looking for trillions more, the tech industry is racing to add to the pileup of generative AI models. The goal is to steadily demonstrate better performance and, in doing so, close the gap between what humans can do and what can be accomplished with AI.

There is another gulf, however, that ought to be given equal, if not higher, priority when thinking about these new tools and systems: the AI trust gap. This gap is closed when a person is willing to entrust a machine to do a job that otherwise would have been entrusted to qualified humans. It is essential to invest in analyzing this second, under-appreciated gap — and in what can be done about it — if AI is to be adopted widely.

The AI trust gap can be understood as the sum of the persistent risks (both real and perceived) associated with AI; depending on the application, some risks are more critical. These cover both predictive machine learning and generative AI. According to the Federal Trade Commission, consumers are voicing concerns about AI, while businesses are worried about several near to long term issues. Consider 12 AI risks that are among the most commonly cited across both groups:

- Disinformation
- Safety and security
- The black box problem
- Ethical concerns
- Bias
- Instability

- Hallucinations in LLMs
- Unknown unknowns
- Job loss and social inequalities
- Environmental impact
- Industry concentration
- State overreach

Taken together, the cumulative effect of these risks contribute to broad public skepticism and business concerns about AI deployment. This, in turn, deters adoption. For instance, radiologists hesitate to embrace AI when the black box nature of the technology prevents a clear understanding of how the algorithm makes decisions on medical image segmentation, survival analysis, and prognosis. Ensuring a level of transparency on the algorithmic decision-making process is critical for radiologists to feel they are meeting their professional obligations responsibly — but that necessary transparency is still a long way off. And the black box problem is just one of many risks to worry about. Given similar issues across different application situations and industries, we should expect the AI trust gap to be permanent, even as we get better in reducing the risks.

This has three major implications. First, no matter how far we get in improving AI's performance, AI's adopters — users at home and in businesses, decision-makers in organizations, policymakers — must traverse a persistent trust gap. Second, companies need to invest in understanding the risks most responsible for the trust gap affecting their applications' adoption and work to mitigate those risks. And third, pairing humans with AI will be the most essential risk-management tool, which means we shall always have a need for humans to steer us through the gap — and the humans need to be trained appropriately.

Consider the 12 risks. For each of them, there are four questions: How they undermine trust in AI? What are the options — industry-initiated or required by regulators — for mitigating or managing the risk? Why do the options offer at best a partial remedy that allows the risk to persist? What are the lessons learned and implications? Collectively, these help break down the AI trust gap, why it can be expected to persist, and what can be done about it.

[1]

Disinformation

Online disinformation isn't new, but AI tools have supercharged it. AI-aided deepfakes have accompanied elections from Bangladesh (where an opposition leader was depicted in a bikini) to Moldova (where a fake clip of president supporting pro-Russian party circulated before the election), giving voters a reason to distrust essential information necessary for the functioning of democracies. As of late 2023, 85% of internet users worried about their inability to spot fake content online — a serious problem given 2024's major elections across the globe.

Social media companies are largely failing to address the threat, as most have severely cut back on the human content moderators that are the most successful defense against disinformation. The largest platform company, Meta, for example, drastically reduced content moderation teams, shelved a fact-checking tool that was in development, and cancelled contracts with external content moderators as part of its "year of efficiency" in 2023. Now, the platform is dealing with a flood of bizarre advertising-driven AI generated content, a reminder that social media recommendation algorithms are yet another form of AI that can be manipulated. Meta's retreats were mirrored at YouTube, which cut its content moderation team, and at X, with even more drastic dismantling. (While Tik Tok hasn't experienced the same level of

cutbacks in its content moderation teams, it has to defend itself against a different set of worries: concerns over compromised security and privacy of user data.) The algorithmic/automated content moderation often offered in place of human moderation is far from adequate.

In the absence of company initiated mitigation, the responsibility falls to regulators, who are stepping in to compel companies to act. In the U.S., multiple states have introduced bills targeting elections-related disinformation and deepfakes. The White House has an executive order requiring the "watermarking," or clear labeling, of AI-created content, which is also required by the EU's recently-passed AI regulation. Elsewhere, India's government holds social media companies accountable for content flagged as harmful that has not been taken down.

Such well-intentioned risk-management measures may have unintended consequences, however, as platforms may simply allocate limited moderation resources to markets with the most regulatory pressure rather than invest in more moderation. The U.S. or the EU will get an over-allocation at the expense of the rest of the world — particularly developing world countries where regulatory and commercial demands are lower — even though there are many more users in these locations. There's evidence was already happening before recent cutbacks: *The Wall Street Journal* uncovered that in 2020, 87% of Facebook's content moderation time was spent on posts in the U.S., despite the fact that 90% of Facebook users were non-U.S.

The lesson is that we have to accept that the harsh reality that disinformation will be hard to eliminate. Depending on where you are in the world, it might even increase in volume and — with the growing sophistication of AI-aided deepfakes — in the degree of deceptiveness. Human vigilance and education in "digital hygiene" will be essential.

[2]

Safety and security

The outlook for AI safety and security risks is sobering. In the largest ever survey of AI and machine learning experts, between 37.8% and 51.4% of all respondents placed at least a 10% probability on scenarios as dire as human extinction, with even 48% of net optimists pegging that probability at 5%. It's hard to think of such dire assessments being considered acceptable for any other technology currently in wide adoption. There are, of course, less apocalyptic risks: malicious use cases for AI tools in cyberattacks, being "jailbroken" to follow illegal commands, etc. In the same survey, situations, such as the chances of AI being jailbroken, were given a relatively high chance — the majority of respondents rated it "likely" or "very likely" — even in the year 2043.

Once again, regulations are critical in mitigating such risks. The White House executive order and the EU regulations require generative AI models above a certain risk threshold to publish results from simulated "red-team" attacks to identify vulnerabilities. That said, it isn't clear that such requirements can be effective in eliminating risk. What's worse is that measures such as red-teaming requirements may be encouraging mere "security theater." There are few standards on foolproof red-teaming methods and criteria, and even if regulations force some transparency, it is hard to confirm such efforts were exhaustive. Startups are unlikely to have the resources to do this work in-house or vouch for externally sourced tests, thereby introducing new sources of vulnerability as their products plug into the larger AI ecosystem or the cost burden deters the startups at the very outset.

The larger lesson — as many experts believe — is that AI safety

and security risks are impossible to eliminate in the foreseeable future. This means that awareness and preparedness will be key and for the most critical and life-and-death applications — from national security to health care — it will be important to keep humans in the loop ensuring that decisions are never fully automated; for example, in highly sensitive negotiations among nuclear-armed nations, agreements would have to ensure keeping decisions relating to launching tests or missiles remain in the hands of humans.

[3]

The black box problem

Transparency is essential to trust-building. With AI, that can include informing users when they are interacting with an AI model, being able to explain how it produced a particular output, and being mindful of what information stakeholders need and delivering it in terms they can understand. Key regulations, such as the EU AI Act will enforce certain transparency standards, but the ever-present challenge is that the incentives for the AI companies encourage them to minimize transparency — to preserve competitive advantage and intellectual property, and to prevent malicious hacks, and reduce exposure to lawsuits about copyright. As such, AI is often a black box — it isn't clear why it produces the output it does.

An industry-led approach to transparency is part of the appeal of open-source AI development. But this, too, has limitations. There are far too many inputs into AI models — from the training data to code used to preprocess it and govern the training process, the model architecture of the model, etc. — so much so that experts cannot agree on what really constitutes "open-source." Companies use this ambiguity as cover to make up their own definitions and hide the key component — the training data, including "synthetic" data — from public view. Even companies,

such as Meta, championing open-source models, are becoming less "open" over time: its Llama 2 model is far less transparent than Llama 1. And even Llama 2, an industry standard on transparency, rates only 54 out of 100 in the Stanford Center for Research on Foundation Models' transparency score. Companies, such as IBM, have volunteered "factsheets" for tracking and transparency mechanisms, but unaudited self-disclosures aren't ideal mechanisms for building trust.

Once again, regulations are expected to play a role in mitigating risks of black box systems. Regulation could compel companies to submit to external audits of AI models and publish the results, but that would require auditing criteria, standards, credible auditors, and genuine regulatory enforceability. A New York law requiring employers using automated employment decision tools to audit them for race and gender bias was found to be toothless by a recent Cornell study. The National Institute of Standards and Technology has an AI Risk Management Framework, but without certification, standards, or an audit methodology, it is, as yet, ineffective.

The lesson here is that while there will be progress on transparency, the black box problem of AI will remain. Each application area will need to develop initiatives geared towards building transparency, which will help ease the adoption process. For example, to help build confidence among radiologists noted earlier, "interpretability" of AI — that is, being able understand the cause of a decision made by an algorithm — with radiological applications is a crucial and growing research field to support clinical practice and adoption.

[4]

Ethical concerns

Most users agree that it's critical to ensure that algorithms go

beyond mathematics and data and are paired with guidelines ensuring ethical principles — e.g., that they respect human rights and values, no matter what the mathematics suggests. There have been several attempts at getting AI developers to coalesce around universally accepted ethical criteria: the Asilomar AI principles, for example, embrace "human values," "liberty and privacy," "common good" among other ideals in developing and using AI models. But there are three obstacles to such endeavors.

For one, ethical ideals aren't universal. The two pre-dominant AI nations, U.S. and China, interpret "liberty and privacy" differently: free speech is paramount in the U.S., while in China, unmoderated speech conflicts with the "common good." Even within the U.S., with its bitter culture wars and polarization, pro-life and pro-choice groups differ on "human values." Some want AI to be anti-"woke," while others want AI's decolonization.

Second, apolitical trans-national bodies have limited powers. The UN has ethical AI principles consistent with its charter and UNESCO has brought companies together to commit to building more ethical AI. Given that most of AI development happens in the private sector, the UN's leverage is limited.

Third, AI companies' organizational incentives exacerbate tensions between ethics and other considerations. For example, with a workforce generally leaning left politically, there's a need for political diversity in ethical oversight. This is hard to do in practice: Google's efforts to assemble an AI ethics advisory council fell apart when employees objected to the appointment the president of the right-wing Heritage Foundation. The much-publicized boardroom versus Sam Altman drama at OpenAI, the failed attempt to separate DeepMind from Google's standard business structure after its acquisition, and the implosion of Stability AI's leadership are also recurring reminders of the battle over priorities in the pioneering AI companies: repeatedly,

commercial goals win over AI for "common good" ideals.

The lesson here is that ethical dilemmas are context-dependent and will be a permanent fixture of AI systems; they are especially critical if they give rise to exclusionary or dangerous decisions. Keeping humans, including those assembled as governance or oversight boards and teams of external watchdogs, in the loop will be essential.

[5]

Bias concerns

Biases in AI stem from many sources: biased or limited training data, the limitations of the people involved in the training, and even the usage context. They can erode confidence in AI models when they appear in critical applications, say, when lenders are found to be more likely to deny home loans to people of color by an even higher percentage when AI is used for mortgage approvals. There are several mitigating actions that can be taken, such as enforcing fairness constraints on AI models, adding more diverse sources of data, training the AI developers in recognizing bias, diversifying the AI talent pool, using tools and metrics to test for biases, etc.

Despite these corrective measures, AI may never be reliably bias-free for several reasons. For one, because AI tools are trained in closed environments and may encounter unfamiliar application environments, they can produce surprising biases due to their limited exposure to real-world data. Further, the processes for testing the presence of bias are difficult. Definitions of what constitutes bias and unfairness can vary widely with contexts as different as the West, China, India — the idea of "fairness," for instance, lends itself to 21 different definitions, making it difficult to reach consensus on when an outcome is considered truly unbiased. Even "unlearning" biases can be hazardous, as it could

introduce new unpredictable associations learned by the AI model, making matters worse overall; Google's and Meta's production of faulty ahistorical images offers a stark example of such risks. Besides, AI models also risk running out of new high-quality data to train on and neutralizing biases arising from limited/low-quality datasets.

The lesson here is that we must accept that AI models will be trained with limitations — of data or trainers themselves operating with human limits — and biases will be inevitable. It will be essential to apply human judgment and vigilance along with swift remedial action before they do damage.

[6]

Instability concerns

In some contexts, AI decisions can change drastically when the input is changed slightly and not in a meaningful way, leading to mistakes and small-to-catastrophic differences in outcomes. For instance, autonomous vehicles can be trusted with many functions but at times they fail: say, when a small obstruction on a stop sign causes an AI-assisted car to drive past it. While AI models are constantly being improved upon by adding training data, enhancing testing protocols, and continuous machine learning, academic research on "stability" of AI has found that beyond basic problems, it is mathematically impossible to develop universally stable AI algorithms. This means we can never be sure of AI making sound decisions when there is even a tiny bit of noise in the input data.

The lesson here is that AI systems can be sensitive to small changes, which are inevitable in the real world beyond the training dataset. The presence of alert humans to do a manual correction or over-ride will be critical in these situations.

[7]

Hallucinations in LLMs

AI hallucinations have caused models to do bizarre things — from professing being in love with their users to claiming to have spied on company employees. Many AI producers have developed a range of mitigation techniques. For example, IBM recommends using high-quality training data; setting clear boundaries on the use of the AI model; using data templates to facilitate output consistency; and continuous testing and refining. Regardless of the actions taken, research suggests that there is a statistical lower-bound on hallucination rates, which means that there will always be a chance of hallucinations appearing. Once again, as is logical for probabilistic models, regardless of the quality of the model architecture or dataset, hallucination incidents can be expected to go down but can never be eliminated.

The lesson is to never trust or put into public use any product of a generative AI model — especially in high-stakes scenarios, such as legal documentation — without trained professionals checking it thoroughly. This can help avoid situations such as the one where ChatGPT made up half-dozen fake court cases with bogus quotes and citations while preparing a legal brief.

[8]

Unknown unknowns

AI can act in ways that we humans cannot anticipate. Models can have blind spots, their training data may not align with the environment in which they're being applied, and they can make errors that developers can't understand. Image recognition models confidently identify items but can, inexplicably, be completely wrong. Continuously training the models on new datasets helps cut the chances, but even as the model improves,

there will always be more information beyond its line of sight, and the risks created by such missing elements compound and can evolve in unexpected ways.

The lesson is that unquestioningly applying AI, which itself has blind spots, is a recipe for disaster; it's critical to ensure the human hand in guiding decisions with an awareness of the application context.

[9]

Job losses and social inequalities

Economies with rising productivity should experience rapid wage gains. Expectations of AI's productivity impact vary: McKinsey has projected an optimistic high of 3.3% a year by 2040 due to use of generative AI. Former Google CEO, Eric Schmidt expects that AI will double everyone's productivity. U.S. Federal Reserve chair, Jerome Powell, is more measured about predicting AI's productivity impact and expects little change in the short run.

A natural way to get a firmer grasp of the impact is to turn to history. Unfortunately, in this case, history provides little guidance. U.S. worker productivity growth, in fact, fell when early digital technologies were introduced. Even when it doubled in the late 1990s, at the time of the World Wide Web's launch, the surge was short-lived, with subsequent surges in 2009 during the Great Recession, after the pandemic began in 2020, and then up again to 4.7%, in the third quarter of 2023, too early to be attributed to AI. This offers insufficient evidence to be optimistic about AI's impact on productivity and wages across economies.

Individual businesses, however, are more bullish, which could translate into job losses as AI takes on tasks done by humans. But that would mean AI would increase wages of those employed, while leading to wage losses for those whose jobs are displaced

worsening social inequalities. To counter such fears, some experts anticipate that generative AI can reduce inequalities by giving lower-skilled workers access to tools for upward mobility. History is more useful here, as it suggests that inequalities will rise: wage inequality tended to increase the most in countries in which companies already relied on automation; Black and Hispanic workers were overrepresented in the 30 occupations with the highest exposure to automation and underrepresented in the 30 occupations with the lowest exposure; and women were expected to be disproportionately negatively affected with 79% of working women in occupations vulnerable to labor displacement by generative AI, as compared to 58% of working men being vulnerable to displacement.

The overall lesson is that the shadow of job losses and increased social inequalities hangs over the adoption of AI. Even acknowledging AI adoption can be problematic: UPS' largest layoff in its history was due to AI replacing humans, according to the CEO on an earnings call, but a spokesperson later denied any connection between layoffs and AI. Clearly, the CEO wished to signal to investors that the company was adopting AI to benefit from cost efficiencies of reduced headcount, but it also had a negative public relations fallout; this suggests that the impact on jobs creates friction in wholeheartedly embracing AI. With multiple stakeholder concerns to balance, businesses will hopefully adopt AI judiciously.

[**10**]

Environmental impact

AI's share of data centers' power use worldwide is expected to grow to 10% by 2025. By 2027, with water needed for cooling, AI's use of data centers could remove the equivalent of half the water consumed in the UK each year. Increasingly powerful chips are needed for AI, and they are contributing to one of the fastest-

growing waste streams. None of these trends shows signs of slowing. The increased use of generative AI, especially for producing images, will make matters even worse. One study finds that 1,000 images using Stable Diffusion XL emits as much carbon dioxide as a gas-powered car driving 4.1 miles.

One important consideration is that AI-aided applications may take the place of other environmentally-costly activity, and may help cut emissions and use of resources. Nevertheless, it is necessary to be aware of its impact. Specific actions, such as the Artificial Intelligence Environmental Impacts Act of 2024, introduced in the U.S. senate, are laudable but will be challenging with no standards for measuring or verifying AI-related emissions. Another risk-mitigation approach is to have new data centers be powered by renewable energy, but the demand for them is growing too fast to be entirely renewable-powered. Even with recycling initiatives in place, only 13.8% of the documented electronic waste is formally collected and recycled with an estimated 16% outside the formal system in high and upper-middle income countries. For the foreseeable future, AI's negative environmental impact is inescapable.

The lesson here is that just as several industries, such as the fossil fuels industry or gas-guzzling vehicle manufacturers, have lost trust among many consumers because of their environmental impact, AI could run similar risks. Human judgment is needed to assess whether the benefits of, say, incorporating AI enhancements to products with good enough alternatives is worth the environmental costs.

[11]

Industry concentration

Despite the high priority on AI from political leadership, its development is industry-led. The reasons are structural: AI

development requires several critical inputs, such as talent, data, computational power, and capital — and the private sector is better positioned to have access to them. Moreover, these resources are concentrated among a few firms.

There are two major concentration points in the AI value chain. A handful of dynamic innovators developing AI models draw on another handful of large companies for critical inputs. Nvidia, Salesforce, Amazon, Google, and Microsoft are the biggest investors in the leading AI innovators, while Meta is major source of open-source models.

Besides capital, AI model developers turn to Nvidia for graphics processing units, to cloud providers such as Amazon and Microsoft to run the models, while Google, Meta, and Microsoft are integrating AI to defend their core products. Even with a more competitive layer of AI applications and services tailored for specific uses, the industry's foundation will clearly remain concentrated. The distrust that users feel towards Big Tech's control will be revisited in even more powerful ways as tech is more AI-intensive.

The usual action to mitigate industry concentration risks, i.e. regulatory scrutiny, has come belatedly. The Federal Trade Commission has only recently launched an inquiry into this growing concentration risk. In the meantime, the trends continue: since the inquiry was launched, Microsoft, already the largest investor in OpenAI, absorbed the top team from Inflection and Amazon invested $2.75 billion in Anthropic. And these — OpenAI, Inflection, and Anthropic — are the three most significant AI innovators in the U.S. currently.

The lesson is that concentration of power in a few firms erode trust because consumers feel locked-in, they worry about overpaying, and have privacy concerns about their data cornered

by powerful firms that can exploit it in other areas.

[12]

State overreach

Trends point in the direction of a greater use of AI and related tools to exert control over citizens by governments across the world. To add to this, the share of populations living in political environments designated as "free" by Freedom House has fallen over the past decade and a half. Global internet freedoms have been declining for 13 years in a row, according to Freedom House, and AI has been facilitating that decline in many ways: spreading state propaganda, enabling more efficient censorship, creating citizens' behavioral profiles, and developing predictive analytics and surveillance. As evidence of the last development, consider that at least 75 out of 176 countries globally are actively using AI technologies for surveillance purposes, including 51% of advanced democracies. With citizen data increasingly in the possession of governments, especially with the growth of digital identity systems, the chances of abuse of power are even greater. Concerned experts have proposed several possible checks and balances, but they haven't been adopted widely.

The larger lesson is that the concern about state overreach may lead to rejecting AI's use even when it can be societally beneficial if used with safeguards. Testing the willingness to accept the tradeoffs will be critical to ensuring that citizens are comfortable with states using AI. Consider the use of facial recognition technology by police: cities, such as San Francisco have banned it.

• • •

While much of the attention has been on the impressive gains in AI performance, Americans are also increasingly pessimistic about AI's impact. Worldwide, trust in AI companies has fallen,

and in the U.S., the drop has been even more dramatic. Granted, many tech companies and commentators suggest you can build AI trust quickly and easily, but let's not kid ourselves; a stubborn AI trust gap persists. And it's here to stay.

Even if the trust gap shrinks, it's important to remember that trust does not necessarily follow from a mathematical or logical calculation: even a single door plug blowing out of a plane shakes up our confidence in the entire aviation system, statistically one of the safest modes of transport. The trust deficit will affect adoption for highly sensitive applications in particular, e.g., health care, finance, transportation, or national security. Leaders should recognize which of the 12 risks are most critical to an application and monitor progress in narrowing the gap.

Even as the technology advances and matures, pairing of AI with humans will remain the biggest signal to concerned potential users that the companies deploying this tech are deserving of trust. But the humans that accompany AI need to be prepared, whether it is by having evidence-based conversations, engaging in active citizenship to scrutinize AI-aided outputs, or ensuring diversity among the teams producing the AI itself.

Currently, the focus is on training AI models to become more like us. Let's not forget that we must also train the humans. They must learn to recognize what causes the AI trust deficit, accept that it will remain, and understand how best to step in to fill the void. Put differently, the industry has spent tens of billions in creating AI products, such as Microsoft Copilot. It's time to also invest in the human alongside: the pilot.

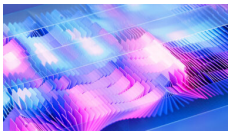
Bhaskar Chakravorti is the Dean of Global

Business at The Fletcher School at Tufts University and founding Executive Director of Fletcher's Institute for Business in the Global Context. He is the author of *The Slow Pace of Fast Change*.

Recommended For You

Research: What Companies Don't Know About How Workers Use AI

4 Types of Gen AI Risk and How to Mitigate Them



PODCAST
Tech at Work: What GenAI Means for Companies Right Now

How to Implement AI - Responsibly