FOREIGN AFFAIRS

Spirals of Delusion

How AI Distorts Decision-Making and Makes Dictators More Dangerous

BY HENRY FARRELL, ABRAHAM NEWMAN, AND JEREMY WALLACE
September/October 2022

Published on August 31, 2022

HENRY FARRELL is the Stavros Niarchos Foundation Agora Professor of International Affairs at Johns Hopkins University.

ABRAHAM NEWMAN is Professor of Government in the Edmund A. Walsh School of Foreign Service at Georgetown University.

JEREMY WALLACE is Associate Professor of Government at Cornell University and the author of *Seeking Truth and Hiding Facts: Information, Ideology, and Authoritarianism in China*.

In policy circles, discussions about artificial intelligence invariably pit China against the United States in a race for technological supremacy. If the key resource is data, then China, with its billion-plus citizens and lax protections against state surveillance, seems destined to win. Kai-Fu Lee, a famous computer scientist, has claimed that data is the new oil, and China the new OPEC. If superior technology is what provides the edge, however, then the United States, with its world class university system and talented workforce, still has a chance to come out ahead. For either country, pundits assume that

superiority in AI will lead naturally to broader economic and military superiority.

But thinking about AI in terms of a race for dominance misses the more fundamental ways in which AI is transforming global politics. AI will not transform the rivalry between powers so much as it will transform the rivals themselves. The United States is a democracy, whereas China is an authoritarian regime, and machine learning challenges each political system in its own way. The challenges to democracies such as the United States are all too visible. Machine learning may increase polarization—reengineering the online world to promote political division. It will certainly increase disinformation in the future, generating convincing fake speech at scale. The challenges to autocracies are more subtle but possibly more corrosive. Just as machine learning reflects and reinforces the divisions of democracy, it may confound autocracies, creating a false appearance of consensus and concealing underlying societal fissures until it is too late.

Early pioneers of AI, including the political scientist Herbert Simon, realized that AI technology has more in common with markets, bureaucracies, and political institutions than with simple engineering applications. Another pioneer of artificial intelligence, Norbert Wiener, described AI as a "cybernetic" system—one that can respond and adapt to feedback. Neither Simon nor Wiener anticipated how machine learning would dominate AI, but its evolution fits with their way of thinking. Facebook and Google use machine learning as the analytic engine of a self-correcting system, which continually updates its understanding of the data depending on whether its predictions succeed or fail. It is this loop between statistical analysis and feedback from the environment that has made machine learning such a formidable force.

What is much less well understood is that democracy and authoritarianism are cybernetic systems, too. Under both forms of rule, governments enact policies and then try to figure out whether these policies have succeeded or failed. In democracies, votes and voices provide powerful feedback about whether a given approach is really working. Authoritarian systems have historically had a much harder time getting good feedback. Before the information age, they relied not just on domestic intelligence but also on petitions and clandestine opinion surveys to try to figure out what their citizens believed.

Now, machine learning is disrupting traditional forms of democratic feedback (voices and votes) as new technologies facilitate disinformation and worsen existing biases—taking prejudice hidden in data and confidently transforming it into incorrect assertions. To autocrats fumbling in the dark, meanwhile, machine learning looks like an answer to their prayers. Such technology can tell rulers whether their subjects like what they are doing without the hassle of surveys or the political risks of open debates and elections. For this reason, many observers have fretted that advances in AI will only strengthen the hand of dictators and further enable them to control their societies.

The truth is more complicated. Bias is visibly a problem for democracies. But because it is more visible, citizens can mitigate it through other forms of feedback. When, for example, a racial group sees that hiring algorithms are biased against them, they can protest and seek redress with some chance of success. Authoritarian countries are probably at least as prone to bias as democracies are, perhaps more so. Much of this bias is likely to be invisible, especially to the decision-makers at the top. That makes it far more difficult to correct, even if leaders can see that something needs correcting.

Contrary to conventional wisdom, AI can seriously undermine autocratic regimes by reinforcing their own ideologies and fantasies at the expense of a finer understanding of the real world. Democratic countries may discover that, when it comes to AI, the key challenge of the twenty-first century is not winning the battle for technological dominance. Instead, they will have to contend with authoritarian countries that find themselves in the throes of an AI-fueled spiral of delusion.

BAD FEEDBACK

Most discussions about AI have to do with machine learning—statistical algorithms that extract relationships between data. These algorithms make guesses: Is there a dog in this photo? Will this chess strategy win the game in ten moves? What is the next word in this half-finished sentence? A so-called objective function, a mathematical means of scoring outcomes, can reward the algorithm if it guesses correctly. This process is how commercial AI works. YouTube, for example, wants to keep its users engaged, watching more videos so that they keep seeing ads. The objective function is designed to maximize user engagement. The algorithm tries to serve up content that keeps a user's eyes on the page. Depending on whether its guess was right or wrong, the algorithm updates its model of what the user is likely to respond to.

Machine learning's ability to automate this feedback loop with little or no human intervention has reshaped e-commerce. It may, someday, allow fully self-driving cars, although this advance has turned out to be a much harder problem than engineers anticipated. Developing autonomous weapons is a harder problem still. When algorithms encounter truly unexpected information, they often fail to make sense of it. Information that a human can easily understand but that machine learning misclassifies—known as "adversarial examples"—can gum up the works badly. For example, black

and white stickers placed on a stop sign can prevent a self-driving car's vision system from recognizing the sign. Such vulnerabilities suggest obvious limitations in AI's usefulness in wartime.

Diving into the complexities of machine learning helps make sense of the debates about technological dominance. It explains why some thinkers, such as the computer scientist Lee, believe that data is so important. The more data you have, the more quickly you can improve the performance of your algorithm, iterating tiny change upon tiny change until you have achieved a decisive advantage. But machine learning has its limits. For example, despite enormous investments by technology firms, algorithms are far less effective than is commonly understood at getting people to buy one nearly identical product over another. Reliably manipulating shallow preferences is hard, and it is probably far more difficult to change people's deeply held opinions and beliefs.

General AI, a system that might draw lessons from one context and apply them in a different one, as humans can, faces similar limitations. Netflix's statistical models of its users' inclinations and preferences are almost certainly dissimilar to Amazon's, even when both are trying to model the same people grappling with similar decisions. Dominance in one sector of AI, such as serving up short videos that keep teenagers hooked (a triumph of the app TikTok), does not easily translate into dominance in another, such as creating autonomous battlefield weapons systems. An algorithm's success often relies on the very human engineers who can translate lessons across different applications rather than on the technology itself. For now, these problems remain unsolved.

Bias can also creep into code. When Amazon tried to apply machine learning to recruitment, it trained the algorithm on data from résumés that

human recruiters had evaluated. As a result, the system reproduced the biases implicit in the humans' decisions, discriminating against résumés from women. Such problems can be self-reinforcing. As the sociologist Ruha Benjamin has pointed out, if policymakers used machine learning to decide where to send police forces, the technology could guide them to allocate more police to neighborhoods with high arrest rates, in the process sending more police to areas with racial groups whom the police have demonstrated biases against. This could lead to more arrests that, in turn, reinforce the algorithm in a vicious circle.

The old programming adage "garbage in, garbage out" has a different meaning in a world where the inputs influence the outputs and vice versa. Without appropriate outside correction, machine-learning algorithms can acquire a taste for the garbage that they themselves produce, generating a loop of bad decision-making. All too often, policymakers treat machine learning tools as wise and dispassionate oracles rather than as fallible instruments that can intensify the problems they purport to solve.

CALL AND RESPONSE

Political systems are feedback systems, too. In democracies, the public literally evaluates and scores leaders in elections that are supposed to be free and fair. Political parties make promises with the goal of winning power and holding on to it. A legal opposition highlights government mistakes, while a free press reports on controversies and misdeeds. Incumbents regularly face voters and learn whether they have earned or lost the public trust, in a continually repeating cycle.

But feedback in democratic societies does not work perfectly. The public may not have a deep understanding of politics, and it can punish governments for things beyond their control. Politicians and their staff may

misunderstand what the public wants. The opposition has incentives to lie and exaggerate. Contesting elections costs money, and the real decisions are sometimes made behind closed doors. Media outlets may be biased or care more about entertaining their consumers than edifying them.

All the same, feedback makes learning possible. Politicians learn what the public wants. The public learns what it can and cannot expect. People can openly criticize government mistakes without being locked up. As new problems emerge, new groups can organize to publicize them and try to persuade others to solve them. All this allows policymakers and governments to engage with a complex and ever-changing world.

Feedback works very differently in autocracies. Leaders are chosen not through free and fair elections but through ruthless succession battles and often opaque systems for internal promotion. Even where opposition to the government is formally legal, it is discouraged, sometimes brutally. If media criticize the government, they risk legal action and violence. Elections, when they do occur, are systematically tilted in favor of incumbents. Citizens who oppose their leaders don't just face difficulties in organizing; they risk harsh penalties for speaking out, including imprisonment and death. For all these reasons, authoritarian governments often don't have a good sense of how the world works or what they and their citizens want.

Such systems therefore face a tradeoff between short-term political stability and effective policymaking; a desire for the former inclines authoritarian leaders to block outsiders from expressing political opinions, while the need for the latter requires them to have some idea of what is happening in the world and in their societies. Because of tight controls on information, authoritarian rulers cannot rely on citizens, media, and opposition voices to provide corrective feedback as democratic leaders can. The result is that they

risk policy failures that can undermine their long-term legitimacy and ability to rule. Russian President Vladimir Putin's disastrous decision to invade Ukraine, for example, seems to have been based on an inaccurate assessment of Ukrainian morale and his own military's strength.

Even before the invention of machine learning, authoritarian rulers used quantitative measures as a crude and imperfect proxy for public feedback. Take China, which for decades tried to combine a decentralized market economy with centralized political oversight of a few crucial statistics, notably GDP. Local officials could get promoted if their regions saw particularly rapid growth. But Beijing's limited quantified vision offered them little incentive to tackle festering issues such as corruption, debt, and pollution. Unsurprisingly, local officials often manipulated the statistics or pursued policies that boosted GDP in the short term while leaving the long-term problems for their successors.

The world caught a glimpse of this dynamic during the initial Chinese response to the COVID-19 pandemic that began in Hubei Province in late 2019. China had built an internet-based disease-reporting system following the 2003 SARS crisis, but instead of using that system, local authorities in Wuhan, Hubei's capital, punished the doctor who first reported the presence of a "SARS-like" contagion. The Wuhan government worked hard to prevent information about the outbreak from reaching Beijing, continually repeating that there were "no new cases" until after important local political meetings concluded. The doctor, Li Wenliang, himself succumbed to the disease and died on February 7, triggering fierce outrage across the country.

Beijing then took over the response to the pandemic, adopting a "zero COVID" approach that used coercive measures to suppress case counts. The policy worked well in the short run, but with the Omicron variant's

tremendous transmissibility, the zero-COVID policy increasingly seems to have led to only pyrrhic victories, requiring massive lockdowns that have left people hungry and the economy in shambles. But it remained successful at achieving one crucial if crude metric—keeping the number of infections low.

Data seem to provide objective measures that explain the world and its problems, with none of the political risks and inconveniences of elections or free media. But there is no such thing as decision-making devoid of politics. The messiness of democracy and the risk of deranged feedback processes are apparent to anyone who pays attention to U.S. politics. Autocracies suffer similar problems, although they are less immediately perceptible. Officials making up numbers or citizens declining to turn their anger into wide-scale protests can have serious consequences, making bad decisions more likely in the short run and regime failure more likely in the long run.

IT'S A TRAP?

The most urgent question is not whether the United States or China will win or lose in the race for AI dominance. It is how AI will change the different feedback loops that democracies and autocracies rely on to govern their societies. Many observers have suggested that as machine learning becomes more ubiquitous, it will inevitably hurt democracy and help autocracy. In their view, social media algorithms that optimize engagement, for instance, may undermine democracy by damaging the quality of citizen feedback. As people click through video after video, YouTube's algorithm offers up shocking and alarming content to keep them engaged. This content often involves conspiracy theories or extreme political views that lure citizens into a dark wonderland where everything is upside down.

By contrast, machine learning is supposed to help autocracies by facilitating greater control over their people. Historian Yuval Harari and a host of other scholars claim that AI "favors tyranny." According to this camp, AI centralizes data and power, allowing leaders to manipulate ordinary citizens by offering them information that is calculated to push their "emotional buttons." This endlessly iterating process of feedback and response is supposed to produce an invisible and effective form of social control. In this account, social media allows authoritarian governments to take the public's pulse as well as capture its heart.

But these arguments rest on uncertain foundations. Although leaks from inside Facebook suggest that algorithms can indeed guide people toward radical content, recent research indicates that the algorithms don't themselves change what people are looking for. People who search for extreme YouTube videos are likely to be guided toward more of what they want, but people who aren't already interested in dangerous content are unlikely to follow the algorithms' recommendations. If feedback in democratic societies were to become increasingly deranged, machine learning would not be entirely at fault; it would only have lent a helping hand.

There is no good evidence that machine learning enables the sorts of generalized mind control that will hollow out democracy and strengthen authoritarianism. If algorithms are not very effective at getting people to buy things, they are probably much worse at getting them to change their minds about things that touch on closely held values, such as politics. The claims that Cambridge Analytica, a British political consulting firm, employed some magical technique to fix the 2016 U.S. presidential election for Donald Trump have unraveled. The firm's supposed secret sauce provided to the Trump campaign seemed to consist of standard psychometric targeting

techniques—using personality surveys to categorize people—of limited utility.

Indeed, fully automated data-driven authoritarianism may turn out to be a trap for states such as China that concentrate authority in a tiny insulated group of decision-makers. Democratic countries have correction mechanisms—alternative forms of citizen feedback that can check governments if they go off track. Authoritarian governments, as they double down on machine learning, have no such mechanism. Although ubiquitous state surveillance could prove effective in the short term, the danger is that authoritarian states will be undermined by the forms of self-reinforcing bias that machine learning facilitates. As a state employs machine learning widely, the leader's ideology will shape how machine learning is used, the objectives around which it is optimized, and how it interprets results. The data that emerge through this process will likely reflect the leader's prejudices right back at him.

As the technologist Maciej Ceglowski has explained, machine learning is "money laundering for bias," a "clean, mathematical apparatus that gives the status quo the aura of logical inevitability." What will happen, for example, as states begin to use machine learning to spot social media complaints and remove them? Leaders will have a harder time seeing and remedying policy mistakes—even when the mistakes damage the regime. A 2013 study speculated that China has been slower to remove online complaints than one might expect, precisely because such griping provided useful information to the leadership. But now that Beijing is increasingly emphasizing social harmony and seeking to protect high officials, that hands-off approach will be harder to maintain.

Chinese President Xi Jinping is aware of these problems in at least some policy domains. He long claimed that his antipoverty campaign—an effort to eliminate rural impoverishment—was a signature victory powered by smart technologies, big data, and AI. But he has since acknowledged flaws in the campaign, including cases where officials pushed people out of their rural homes and stashed them in urban apartments to game poverty statistics. As the resettled fell back into poverty, Xi worried that "uniform quantitative targets" for poverty levels might not be the right approach in the future. Data may indeed be the new oil, but it may pollute rather than enhance a government's ability to rule.

This problem has implications for China's so-called social credit system, a set of institutions for keeping track of pro-social behavior that Western commentators depict as a perfectly functioning "AI-powered surveillance regime that violates human rights." As experts on information politics such as Shazeda Ahmed and Karen Hao have pointed out, the system is, in fact, much messier. The Chinese social credit system actually looks more like the U.S. credit system, which is regulated by laws such as the Fair Credit Reporting Act, than a perfect Orwellian dystopia.

More machine learning may also lead authoritarian regimes to double down on bad decisions. If machine learning is trained to identify possible dissidents on the basis of arrest records, it will likely generate self-reinforcing biases similar to those seen in democracies—reflecting and affirming administrators' beliefs about disfavored social groups and inexorably perpetuating automated suspicion and backlash. In democracies, public pushback, however imperfect, is possible. In autocratic regimes, resistance is far harder; without it, these problems are invisible to those inside the system, where officials and algorithms share the same prejudices.

Instead of good policy, this will lead to increasing pathologies, social dysfunction, resentment, and, eventually, unrest and instability.

WEAPONIZED AI

The international politics of AI will not create a simple race for dominance. The crude view that this technology is an economic and military weapon and that data is what powers it conceals a lot of the real action. In fact, AI's biggest political consequences are for the feedback mechanisms that both democratic and authoritarian countries rely on. Some evidence indicates that AI is disrupting feedback in democracies, although it doesn't play nearly as big a role as many suggest. By contrast, the more authoritarian governments rely on machine learning, the more they will propel themselves into an imaginary world founded on their own tech-magnified biases. The political scientist James Scott's classic 1998 book, *Seeing Like a State*, explained how twentieth-century states were blind to the consequences of their own actions in part because they could see the world through only bureaucratic categories and data. As sociologist Marion Fourcade and others have argued, machine learning may present the same problems but at an even greater scale.

This problem creates a very different set of international challenges for democracies such as the United States. Russia, for example, invested in disinformation campaigns designed to sow confusion and disarray among the Russian public while applying the same tools in democratic countries. Although free speech advocates long maintained that the answer to bad speech was more speech, Putin decided that the best response to more speech was more bad speech. Russia then took advantage of open feedback systems in democracies to pollute them with misinformation.

One rapidly emerging problem is how autocracies such as Russia might weaponize large language models, a new form of AI that can produce text or images in response to a verbal prompt, to generate disinformation at scale. As the computer scientist Timnit Gebru and her colleagues have warned, programs such as Open AI's GPT-3 system can produce apparently fluent text that is difficult to distinguish from ordinary human writing. Bloom, a new open-access large language model, has just been released for anyone to use. Its license requires people to avoid abuse, but it will be very hard to police.

These developments will produce serious problems for feedback in democracies. Current online policy-comment systems are almost certainly doomed, since they require little proof to establish whether the commenter is a real human being. Contractors for big telecommunications companies have already flooded the U.S. Federal Communications Commission with bogus comments linked to stolen email addresses as part of their campaign against net neutrality laws. Still, it was easy to identify subterfuge when tens of thousands of nearly identical comments were posted. Now, or in the very near future, it will be trivially simple to prompt a large language model to write, say, 20,000 different comments in the style of swing voters condemning net neutrality.

Artificial intelligence–fueled disinformation may poison the well for autocracies, too. As authoritarian governments seed their own public debate with disinformation, it will become easier to fracture opposition but harder to tell what the public actually believes, greatly complicating the policymaking process. It will be increasingly hard for authoritarian leaders to avoid getting high on their own supply, leading them to believe that citizens tolerate or even like deeply unpopular policies.

SHARED THREATS

What might it be like to share the world with authoritarian states such as China if they become increasingly trapped in their own unhealthy informational feedback loops? What happens when these processes cease to provide cybernetic guidance and instead reflect back the rulers' own fears and beliefs? One self-centered response by democratic competitors would be to leave autocrats to their own devices, seeing anything that weakens authoritarian governments as a net gain.

Such a reaction could result in humanitarian catastrophe, however. Many of the current biases of the Chinese state, such as its policies toward the Uyghurs, are actively malignant and might become far worse. Previous consequences of Beijing's blindness to reality include the great famine, which killed some 30 million people between 1959 and 1961 and was precipitated by ideologically driven policies and hidden by the unwillingness of provincial officials to report accurate statistics. Even die-hard cynics should recognize the dangers of AI-induced foreign policy catastrophes in China and elsewhere. By amplifying nationalist biases, for instance, AI could easily reinforce hawkish factions looking to engage in territorial conquest.

Perhaps, even more cynically, policymakers in the West may be tempted to exploit the closed loops of authoritarian information systems. So far, the United States has focused on promoting Internet freedom in autocratic societies. Instead, it might try to worsen the authoritarian information problem by reinforcing the bias loops that these regimes are prone to. It could do this by corrupting administrative data or seeding authoritarian social media with misinformation. Unfortunately, there is no virtual wall to separate democratic and autocratic systems. Not only might bad data and crazy beliefs leak into democratic societies from authoritarian ones, but terrible authoritarian decisions could have unpredictable consequences for democratic countries, too. As governments think about AI, they need to

realize that we live in an interdependent world, where authoritarian governments' problems are likely to cascade into democracies.

A more intelligent approach, then, might look to mitigate the weaknesses of AI through shared arrangements for international governance. Currently, different parts of the Chinese state disagree on the appropriate response to regulating AI. China's Cyberspace Administration, its Academy of Information and Communications Technology, and its Ministry of Science and Technology, for instance, have all proposed principles for AI regulation. Some favor a top-down model that might limit the private sector and allow the government a free hand. Others, at least implicitly, recognize the dangers of AI for the government, too. Crafting broad international regulatory principles might help disseminate knowledge about the political risks of AI.

This cooperative approach may seem strange in the context of a growing U.S.-Chinese rivalry. But a carefully modulated policy might serve Washington and its allies well. One dangerous path would be for the United States to get sucked into a race for AI dominance, which would extend competitive relations still further. Another would be to try to make the feedback problems of authoritarianism worse. Both risk catastrophe and possible war. Far safer, then, for all governments to recognize AI's shared risks and work together to reduce them.
