CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

JUNE 2020

Shaping the Terrain of AI Competition

AUTHOR
Tim Hwang

CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Established in January 2019, the Center for Security and Emerging Technology (CSET) at Georgetown's Walsh School of Foreign Service is a research organization focused on studying the security impacts of emerging technologies, supporting academic work in security and technology studies, and delivering nonpartisan analysis to the policy community. CSET aims to prepare a generation of policymakers, analysts, and diplomats to address the challenges and opportunities of emerging technologies. During its first two years, CSET will focus on the effects of progress in artificial intelligence and advanced computing.

CSET.GEORGETOWN.EDU | CSET@GEORGETOWN.EDU

JUNE 2020

Shaping the Terrain of AI Competition

AUTHOR
Tim Hwang

ACKNOWLEDGMENTS

The author would like to acknowledge Greg Allen, Tarun Chhabra, Matt Daniels, Richard Danzig, Andrew Imbrie, Matt Mahoney, Jason Matheny, Maura McCarthy, Igor Mikolic-Torreira, Dewey Murdick, Michael Page, Paul Scharre, Helen Toner, and Lynne Weil for their invaluable feedback on earlier drafts of this paper.

PRINT AND ELECTRONIC DISTRIBUTION RIGHTS

Cover photo: Kmls/AdobeStock

Contents

Executive Summary

The concern that China is well-positioned to overtake current U.S. leadership in artificial intelligence in the coming years has prompted a simply stated but challenging question. How should democracies effectively compete against authoritarian regimes in the AI space?

Policy researchers and defense strategists have offered possible paths forward in recent years, but the task is not an easy one. Particularly challenging is the possibility that authoritarian regimes may possess structural advantages over liberal democracies in researching, designing, and deploying AI systems. Authoritarian states may enjoy easier access to data and an ability to coerce adoption of technologies that their democratic competitors lack. Authoritarians may also have stronger incentives to prioritize investments in machine learning, as the technology may significantly enhance surveillance and social control. No policy consensus has emerged on how the United States and other democracies can overcome these burdens without sacrificing their commitments to rights, accountability, and public participation.

This paper offers one answer to this unsettled question in the form of a "terrain strategy." It argues that the United States should leverage the malleability of the AI field and shape the direction of the technology to provide structural advantages to itself and other democracies. This effort involves accelerating the development of certain areas within machine learning (ML)—the core technology driving the most dramatic advances in AI—to alter the global playing field.

This "terrain" approach is somewhat novel in literature on AI, national strategy, and geopolitical competition. However, the framework presented

here adopts an established, time-tested approach taken by the U.S. government in structuring investments in other domains of scientific research. From the National Science Foundation to Project Apollo, the United States has played a major role in advancing progress on scientific problems relevant to the national interest but lacking sufficient support from industry.

An examination of these issues through a "terrain" lens suggests democracies should invest their resources in three critical domains:

- **Reducing a dependence on data.** Authoritarian regimes may have structural advantages in marshaling data for ML applications when compared to their liberal democratic adversaries. To ensure better competitive parity, democracies should invest in techniques that reduce the scale of real-world data needed for training effective ML systems.
- **Fostering techniques that support democratic legitimacy.** Democracies may face greater friction in deploying ML systems relative to authoritarian regimes due to their commitments to public consent. Enhancing the viability of the technology in a democratic society will require investing in ML subfields, including work in interpretability, fairness, and privacy.
- **Challenging the social control uses of ML.** Recent advances in AI appeal to authoritarian regimes in part because they promise to enhance surveillance and other mechanisms of control. Democracies should advance research eroding the usefulness of these applications.

This paper makes the case for the terrain strategy in two parts. The first proposes a strategic framework for thinking about global competition in AI and argues for shaping the research field to give the United States and other democracies an advantage in deploying the technology. The second fleshes out this framework, recommending specific, promising technical domains that the United States should accelerate in order to execute on this strategy.

Introduction

Leadership in technological innovation has long been a crucial national security asset to the United States. From atomic energy and stealth technology to the internet and genetic engineering, the United States has led the way in nearly all the major breakthroughs of the last decades. This has served its interests not just on the battlefield but economically as well.

The potential loss of this national technological lead serves as a powerful motivating force and locus of discussion in national security circles. The Soviet Union's launch of Sputnik in 1957 triggered a major surge of investment and coordination activity to ensure the United States was not left behind in aerospace technology. Numerous, less popularly known examples pop up in other domains, including security concerns around a perceived loss of superiority in developing supercomputers[1] and green technology,[2] as well as more conventional fears over new weapons technology such as hypersonic missiles.[3] Commentators worry too about the decline of science and technology education in the United States and its long-term impact on the nation's global dominance.[4]

In recent years, China has stoked these fears perhaps most among U.S. security analysts. Chinese technology firms seem well poised to compete, if not outcompete, their U.S. counterparts. The Chinese government has made several major announcements signaling an aggressive campaign to invest in and quickly advance a range of critical technologies. To the extent that U.S. political and economic dominance hinges on technological dominance, the two global powers seem poised to settle in for an extended period of competition.

Recent breakthroughs in artificial intelligence—specifically in the subfield of AI known as machine learning (ML)—have become wrapped up in this broader concern around a loss of the U.S. technological edge. China has made AI a key priority, announcing a raft of new initiatives that signal major investment and state interest in becoming a global leader in ML. In July 2017, the Chinese government announced its "Next Generation Artificial Intelligence Development Plan," a detailed and specific agenda designed to position China as the world's premier AI innovation center by 2030.[5] Chinese cities and states have pledged billions of dollars to support AI development in their regions. Chinese universities train massive numbers of engineers and researchers in the field, and Chinese products powered by AI have proven wildly successful both at home and abroad. By October 5, 2018, China boasted 14 "unicorn" AI companies—private companies valued at $1 billion or more.[6] Mobile application unicorn ByteDance, which owns TikTok, credits its successful expansion beyond the Chinese market to the broad applicability of user data collected and processed with technologies developed in ByteDance's AI lab.[7] Commercial drone manufacturer DJI, based in Shenzhen, has partnered with Microsoft to build out a suite of AI capabilities to enhance its hugely popular remote-controlled robots.[8]

The United States, which has made AI a core part of its defense strategy and whose leading companies have made AI a key differentiator in their products and services, perceives these developments as a major national security risk. Some commentators have dubbed the competition between the United States and China a new "arms race" in AI.[9] Congress has also taken action, creating and funding in 2018 the National Security Commission on Artificial Intelligence, tasked with reviewing developments in AI to address the national and economic security needs of the United States and seek out "opportunities to advance U.S. leadership."[10]

The concern that China is well positioned to overtake current U.S. leadership in AI has prompted policymakers to question how the United States should most effectively compete in the AI space. Policy researchers and defense strategists have offered a number of different possible paths forward.[11]

But this has not been an easy task. One particularly thorny challenge has been the possibility that authoritarian regimes may possess structural advantages over liberal democracies in researching, designing, and deploying AI systems. Authoritarian states may enjoy easier access to data and more effective tools to coerce adoption of technologies that their democratic competitors do not. Authoritarians may also have stronger incentives to prioritize investments in ML as the technology may significantly enhance their systems of surveillance and social control.

This paper offers one answer in the form of a "terrain strategy." It argues that the United States should leverage the malleability of the field of AI and work to shape the direction of research in ways that provide structural advantages to itself and other democracies. This involves accelerating the development of certain areas within ML—the core technology driving the most dramatic advances in AI—to alter the global playing field of the technology. Such an approach would prioritize investments in research that reduces dependence on large corpora of real-world training data, improves the technology's democratic viability, and attacks the social control uses of high value to authoritarian states.

In essence, the terrain strategy seeks to empower democracies to effectively compete in the technology by reshaping the nature of the technology itself. Competing in AI without engaging with the technology in this way will leave the United States in particular and democracies in general at a structural disadvantage against their authoritarian competitors. Even worse, democracies may be in the unenviable position of having to compromise on core values like privacy in an effort to preserve their technological lead. The terrain strategy offers one path whereby democracies might effectively compete without having to make such sacrifices.

It is easy to assume that AI is a monolithic, single technology whose applications, research ecosystem, and future directions are already settled. This picture does not reflect reality. ML is both highly multi-faceted and deeply malleable. Rather than a single monolith, the technology is better understood as a broad family of related but distinct techniques, each with unique strengths and weaknesses.

The competitive strategy around AI must take these characteristics into account. It is not just a matter of whether or not to invest in ML, but specifically *what* provides the greatest national benefit within the domain of ML. Failing to examine these details may prevent the national security community from identifying important opportunities that enable the United States to retain its edge and even outpace competitors like China in the mid- to long-term.

The first part of this paper proposes a strategic framework for thinking about global competition in AI and argues for shaping the research field to give the United States and its allies the advantage in the technology. The second fleshes out this framework, recommending specific, promising technical domains that the United States should accelerate in order to execute on this strategy.

1 The Terrain Strategy

The current state of play in ML favors authoritarian societies over democratic ones. To win, democracies must work to re-shape the competitive dynamics of ML in order to retain their leadership. This section outlines why these structural advantages exist and how AI might be reshaped to rectify this imbalance, then argues that democratic governments should play a role in addressing private underinvestment in certain areas of ML research.

ACCESS TO INPUTS DEFINE COMPETITIVE ADVANTAGE IN ML

Viewing ML as a technology whose success depends on the availability of a specific set of resources enables us to think concretely about the kinds of actors and entities that are best positioned to obtain the benefits of these technologies.

It is worthwhile to take a step back from the intense hype around ML to think for a moment about what the technology actually is. Simply stated, ML is the subfield of AI studying computer systems that improve through processing data. This improvement process is called "training." The training process generates a piece of code—known in the field as a "model"—which ideally can then accomplish the trained task.

Consider the example of teaching a computer to recognize a cat in a photo. ML requires a large corpus of training data to do this: images of cats that are manually labeled by annotators as depicting a cat. These are processed through a learning algorithm, whereby a model is generated that associates the visual of the cat with the label "cat." This training process, in effect, is a large number of mathematical operations enabling the machine to infer a set of rules that accomplish the task: detecting a

cat in an image effectively. If the process is successful, the resulting model can then accomplish this task with novel images it has not seen before.

From even this rudimentary example, it is clear that the successful use of ML relies on a few core inputs.[12] Specifically, it requires (1) *training data*, which are the examples that the algorithm learns from; (2) *learning algorithms*, the algorithms that execute the training process; (3) *computational power*, the computers necessary to run the many calculations needed to generate a model; and (4) *talent*, the human expertise necessary to set up these systems and assess the quality of the resulting model. For the vast portion of ML applications used today, the absence of one of these resources will result in poor quality models or an inability to create ML systems at all.

It is possible to look to these inputs to make concrete predictions about the relative advantages and disadvantages that different actors bring to the competitive landscape. Imagine for a moment the marketplace for cat detection technologies. Who in the market is well poised to offer an ML-driven cat detection system? Will it be one of the existing, established players in the space, or an upstart? Among upstarts and incumbents, who will have the strongest chance at building the highest-performance systems?

The business running an online cat lover community may already have access to a large number of cat images; this access places it at a cost advantage against a business that needs to purchase this training dataset from a third party or hire photographers to go out and collect many photos of cats. Similarly, an upstart company with extensive expertise building ML models in other domains might come to the market with an established team of experts, allowing it to outcompete an incumbent cat photo sorting giant with a lower capacity to recruit specialists to build the technology.

Of course, none of these assessments is definitive in determining who might emerge as the dominant business in a sector. A wily competitor able to market effectively their subpar AI systems might still prevail over a company offering a technically superior product. Whether or not a competitor with more relevant data but less technical expertise triumphs over a rival with less data but more technical expertise will be highly dependent on the context. Non-technical factors, such as organizational effectiveness in integrating the technology into existing processes, the costs of training personnel, and the quality of software development practices will play a major role.[13] Inputs set the stage, but there is no law that makes them determine the final outcome.

Nonetheless, it is indisputable that the relative distribution of access to ML inputs exerts a powerful influence over the competitive dynamics between actors and shapes the strategies they might bring to the field. Such a lens allows us to make

some structured conjectures around the structural advantages and disadvantages that different actors face in attempting to leverage the benefits of ML.

AN AUTHORITARIAN ADVANTAGE

The relationship between inputs and competitive dynamics in ML can be generalized beyond commercial competition between businesses in next-generation cat detection. The effectiveness of an ML system is based on the ability to marshal relevant data, computing power, learning architectures, and expertise. That applies regardless of whether the task is to create ML systems for sorting cat images or piloting a military drone.

This analysis can be expanded to ask a far broader question: are certain societies or governance structures more or less able to leverage the benefits of ML? More to the point, are centralized autocracies better or worse than liberal democracies in building, training, and deploying AI, particularly for forwarding the interests of a state?

There are three reasons to believe that authoritarian regimes have a relative advantage. For one, they may have an easier time acquiring data for training ML applications,[14] in part because they may already maintain an existing infrastructure for ubiquitous surveillance that enables easy data collection. Moreover, there may be no strong legal mechanisms to protect citizen privacy or prevent the state from compelling companies to provide access to their data.

In contrast, liberal democracies may impede the collection of data through relatively robust privacy regimes. To the extent that these societies have large-scale systems of data collection, these datasets may be centralized within private corporations with legal protections, rather than in an institution immediately accessible by the state.

Second, authoritarian regimes can more effectively force the deployment of novel ML technologies, allowing these systems to be rolled out and fine-tuned through subsequent data collection without needing to obtain the consent of the general public.

Democratic commitments to public consent mean citizens have comparatively more mechanisms for resisting unwanted deployments of ML technology. Individuals and civil society organizations can protest or bring lawsuits to prevent the adoption of certain ML systems. Media freedoms allow journalists to expose and rally public opinion against objectionable uses of the technology. Key communities of technical specialists can refuse to work on certain applications of ML, and discourage their employers from doing so, as well.[15]

Third, the surveillance and control interests that authoritarians bring to their investments in ML may also make it challenging for liberal democracies to make the comparable financial commitments necessary to retain their lead on the cutting

edge of the technology. For authoritarian regimes, refining and perfecting ML may go directly to a core priority of sustaining and protecting the state, whereas democratic societies may face more conflicting incentives around whether or not to prioritize and coordinate their investment.

Authoritarian regimes, therefore, may be able to train, deploy, and improve ML systems more effectively than democratic societies. Facial recognition technology provides a concrete illustration of these authoritarian advantages. Companies specializing in building facial recognition models such as SenseTime have benefitted significantly from collaborations with the Chinese government on surveillance applications.[16] These collaborations have yielded access to extensive face data for training ML models and provided opportunities to test their technology at scale without the need for public consent.

Training and deploying facial recognition systems has not been as easy in the United States. Academic researchers have criticized companies marketing these technologies, highlighting the gender and racial biases that facial recognition systems might perpetuate.[17] Journalists have aggressively exposed unscrupulous practices that some startups have engaged in to gather training data.[18] Civil society organizations have successfully lobbied for municipal bans on the use of the technology throughout the country.[19] Industry leaders in ML such as Google have spoken out in favor of a moratorium on the use of facial recognition technologies.[20] Driven by the public concern around facial recognition, Congress is considering laws that would require consent before collecting or sharing face data.[21]

The result is that the U.S. government and companies contend with a more difficult environment for collecting data to train facial recognition capabilities and have less freedom to deploy these systems once trained. While this resistance works to protect civil liberties in the United States, the outcome is that Chinese companies will likely continue to lead in ML-based facial recognition technologies for the foreseeable future.

These structural advantages do not automatically determine the winner in AI competition. Autocratic rollouts of ML-driven surveillance and social control mechanisms have occasionally been far less successful than some commentators have made them out to be.[22] Moreover, citizens of autocratic regimes have found numerous ways to defend their privacy and subvert state surveillance even in the absence of formal legal mechanisms.[23] Governments may acquire a cutting-edge technology, only to find that bureaucratic politics thwarts its usefulness in practice.[24] The specifics will matter. But all else being equal, autocracies have an edge when it comes to building and deploying AI systems.

This raises an important question: given the structural advantages that more autocratic regimes bring to competition around AI, how can democracies retain and expand their edge in the technology?

The national security community has leaned heavily on a relatively small set of tools to answer this question. Frequently cited proposals include streamlining the process for government funding of non-defense companies, increasing AI R&D budgets, developing private sector incentives—namely generous tax breaks—and recruiting talent from other countries.[25] While these policies would improve democratic competitiveness in AI, they do not directly address the core structural advantages authoritarian regimes can bring to the table in advancing the state of the art in ML. Democracies are left fighting an uphill battle, contending with headwinds their competitors do not face.

Perhaps more troubling, democracies such as the United States may choose to compromise on their values in order to compete effectively in ML. Executive orders have already been issued to lower barriers to accessing citizen data in order to increase the availability of training data for ML systems.[26] Policies like these erode privacy protections for the sake of accelerating AI development. Democracies should seek to compete effectively while preserving their core values, rather than move in a more autocratic direction to preserve some semblance of technological parity.

Democracies can do better than this. The shape of AI is not fixed, but in flux. The United States can strategically rewrite the underlying competitive dynamics of the technology, working to offset the structural advantages that autocracies enjoy while mitigating the structural challenges that democracies face.

THE MALLEABILITY OF ARTIFICIAL INTELLIGENCE

Most popular reporting on breakthroughs in ML tends to emphasize the advancement of the technology: a new performance milestone passed, a new level of investment reached, or a new application of the technology discovered. Less frequently spotlighted is that the practice of ML itself has rapidly morphed over the past decade of rapid progress. Democracies can take advantage of this malleability to offset autocratic advantage.

ML has not merely improved; it has fundamentally changed. Some of these changes concern the practical ecosystem of engineering in ML systems. For one, the high-profile nature of the technology has encouraged a massive influx of talent into the field, from highly specialized researchers working in the field and advancing the state of the art to an increasing pool of yeoman software engineers familiar with the basics of the technology.[27] The universe of practical techniques, tools, and resources available for building ML systems has also expanded. This includes the development of open-source software packages for using ML, such as TensorFlow and PyTorch, and a range of training and educational resources for learning techniques in the field.[28]

Other changes concern the research field of ML itself: the last decade of activity has led to shifts in what is being done with the technology—and how. Progress has

been made in the subdomain of research focusing on "few shot" learning—the challenge of developing ML systems that can perform a task effectively with significantly less training data than typically required.[29] Adversarial examples—the subdomain of research focusing on the creation of and defense against seemingly innocuous inputs that can manipulate the behavior of ML systems—have also become a major area of activity in the past few years.[30]

These changes are important because they modify the strategic landscape around AI. For example, the rapidly increasing pool of available global talent working in ML makes it more challenging for an actor to gain a robust strategic advantage by monopolizing key personnel. In the early 2010s, by contrast, ML had not yet received mainstream attention and the number of leading researchers focused on these technical problems was comparatively small.[31]

Overcoming specific technical hurdles may be even more strategically significant. For example, major breakthroughs in "one shot" and "few shot" learning might reconfigure the competitive landscape. These techniques lower the requisite quantity of data needed to train a high-performance ML system. Democracies could leverage these techniques to produce effective AI systems despite having more limited access to data than their authoritarian counterparts. In this sense, "one shot" and "few shot" learning can work to offset a key authoritarian advantage in AI competition.

The reasons for these technical breakthroughs are not mysterious: progress on a given problem will depend in part on researchers and technologists deciding to focus their energies on solving that problem. This makes the field and the shape of ML malleable insofar as an actor can influence how the technical community prioritizes its efforts.

Democracies can use this malleability to their advantage. They can work to encourage progress on technical challenges that, if resolved, would mitigate the structural advantages that authoritarians currently bring to AI competition. Failing to do so may mean that democracies perpetually compete on uneven terrain.

THE ROLE OF GOVERNMENT IN SHAPING THE AI TERRAIN

ML is multifaceted, and a state might potentially invest in a wide range of targets in an attempt to shape the strategic terrain of the technology. How should governments identify the opportunities that will yield the greatest impact?

Governments do not act alone in investing in ML. Quite the opposite, the funding and support flowing from states is just one part of a vast ecosystem of corporations and other funders that are financing AI around the world. This funding constitutes a set of market forces pushing and pulling the field of ML in different directions.

Some of these directions might make it more likely for liberal democracies to excel in the technology, while others might make it less so.

For each potential area of investment, one can ask if the existing ecosystem of financial and human capital will work to systematically shift the field toward ensuring that democracies can compete on even-footing or at an advantage. Is the field making it easier over time for democracies to overcome their limitations in obtaining training data? Is the field expanding the ability for ML engineers to meet requirements for public consent in the rollout of these systems?

Where the market produces underinvestment in certain problem areas, the state can shape incentives to change the level of activity in that domain. For areas with sufficient private investment, governments can allow the market to shape the field. In other words, governments might act as an effective "gap filler" in the space by accelerating and fostering work on problems that otherwise would receive insufficient support due to the market incentives facing the private sector.

This method does not require a substantial change in how liberal democracies engage in scientific research. It instead applies an existing, long-standing approach. Vannevar Bush, whose seminal 1945 report *Science—the Endless Frontier* "entrenched the concept of government patronage of scientific research in policy discourse" and inspired the creation of the National Science Foundation, recommended a similar framework.[32] As he wrote, "[t]here are areas of science in which the public interest is acute but which are likely to be cultivated inadequately if left without more support than will come from private sources. These areas ... should be advanced by active Government support."[33] This should continue to be an organizing principle even in today's dynamic ML research environment.

This "gap filler" approach suggests that democracies should place their resources into advancing ML along three dimensions: data efficiency, democratic viability, and subversion of social control.

First, democracies should invest in solving a set of technological problems that reduce or replace the reliance of ML systems on massive datasets. Liberal democracies will face a comparatively harder environment for acquiring data than authoritarian regimes due to their commitments to privacy and private enterprise. Reducing the data requirements needed to create effective ML systems will help eliminate a hard tradeoff between acquiring data and maintaining democratic values. It will also allow democracies to better compete at parity with their authoritarian competitors in training a range of AI systems. Leading industrial labs often assume plentiful access to data, limiting their incentives to make solving these types of data-scarce technical challenges a top priority. Investing in this area may offer great promise.

Second, democracies should invest in advancing state-of-the-art techniques to ensure the technology is democratically viable. Liberal democracies will face challenges in unilaterally imposing technologies on their publics and in forcing collaborations between private industry and the state. Advancing methods and know-how that improve transparency, ensure fairness, and protect privacy will raise the possibility of consensual adoption of ML systems. This will work to offset the authoritarian advantage of being able to deploy AI systems unilaterally. Democracies can rally fragmented research efforts on these topics, helping to speed innovation and foster global norms around values that should be embedded in ML systems.

Finally, democracies should invest in a set of methods that actively undermine and raise the risks of using these technologies for social control. This prong focuses on eroding the advantages authoritarian regimes may enjoy from developments in ML, rather than eliminating the structural disadvantages that liberal democracies may face. Investing in technologies that attack these applications will reduce their value to authoritarian regimes, potentially reducing investment by these regimes in the field over time and allowing democratic investment to more easily keep pace. This space is likely to see under-investment from the private sector. Although corporations have strong interests in building defenses against attacks on ML systems, they do not have similar incentives to commoditize and distribute tools to enable those attacks in the first place.

The second part of this report examines each of these prongs in greater detail. For each, this paper argues the case for these priorities, highlights a series of specific technical areas in which investment might make a major impact, and examines the strategic implications if successful.

2 Shaping the Terrain

Democratic societies should invest in advancing the field of ML in ways that offset authoritarian advantages in creating and deploying AI systems. Three research areas are likely to produce the highest impact on this front: improvement of data efficiency, enhancement of democratic viability, and subversion of social control.

REDUCING DATA DEPENDENCE

ML has traditionally relied upon access to a large corpus of data, which is used to train the statistical models solving the problem at hand. Detecting objects in images requires access to many images that have already been tagged with the objects of interest. Creating a translation system between two languages requires access to a large existing body of bilingual texts.[34] Conversely, failure to acquire data relevant to the training task has imposed a hard limit on the level of performance of ML systems.

This technical hurdle advantages authoritarian societies over democratic ones. Privacy protections and a fragmented data landscape in democracies may make it more challenging to acquire the datasets needed to produce effective ML systems. Even when private companies—platforms such as Google or Facebook—possess access to massive amounts of data, governments in democratic societies may have difficulty accessing this data freely. Authoritarians therefore possess a "data advantage"—not in the aggregate amount of data they hold, but in their ease of accessing plentiful training data when needed.

Liberal democracies could take a heavy-handed approach of working to eliminate these privacy protections and expand the ease of government and corporate access to training data. This is a high-cost, time-consuming

endeavor, in some cases requiring a sacrifice of strongly held democratic values to obtain the benefits of the technology.

Such compromises may not be necessary. Rather than concede that ML will forever require large, real-world datasets, targeted investments in the technical field could potentially erode the authoritarian benefit of relatively frictionless data collection. This may allow democratic societies to better benefit from the technology and produce ML systems that perform at similar parity without expansive data collection, significantly leveling the playing field.

This area may be a particularly promising one for acceleration and support in part because it is likely that the market will systematically underinvest in their development over time. One of the main motivations underlying major investments in ML and AI by companies like Google and Facebook is that these technologies leverage the massive datasets that these businesses already possess to enable next-generation products and services.[35] To that end, many industrial labs have invested aggressively toward advancing ML techniques where an abundance of data is presumed.

Some investment into research on making ML methods work in data-limited environments does exist. Indeed, government research agencies like DARPA already support such research.[36] However, pressures to prioritize this work will remain limited so long as the primary use case of the technology remains situations in which data is plentiful.

Investment Opportunities

Three subdomains of technical research could reduce the dependence of ML techniques on access to massive corpora of data: "few shot" or "one shot" learning, simulation learning, and self-play. These are promising areas for investment.

First, "few shot" or "one shot" learning techniques seek to enable ML systems to effectively train on a task with a significantly smaller quantity of data than typically necessary. These improvements can occur in many places throughout the design of an ML system, restructuring everything from the training data to the model itself in an effort to make training more efficient.[37] More ambitious in this context, "zero shot" learning seeks to enable an ML system to generalize to a task it has not previously been trained on.[38]

Second, simulation learning focuses not on reducing the overall amount of data necessary for the training process, but on sourcing that data virtually and thus potentially more cheaply. A classic illustration of this approach's data advantage is in the problem of training a robot arm to successfully grasp an object. Applying ML to this task could mean physically assembling a bank of robotic arms repeatedly attempting to grasp different objects.[39] While this method does enable robotic arms

to learn to manipulate three-dimensional objects, the physical costs of setting up and maintaining such a system to collect data can be expensive.

Simulation bypasses the cost and complexity of maintaining physical robots. Instead, a simulated robotic arm in virtual space repeatedly attempts to pick up virtual objects. With an accurate simulation of the physical world, the machine is able to learn how to grasp an object in a real robot.[40] This reduces the need to draw the data from the real world, and in some contexts, may offer cheaper sourcing of relevant data than physical collection.

Finally, self-play refers to methods in which reinforcement learning agents are set to compete with one another to improve their performance. In some contexts, this can reduce or eliminate entirely the need to draw on real world data to train a system. One striking illustration of this technique is in the transition from AlphaGo—DeepMind's system for playing the game of Go, which bested 18-time world champion Lee Sedol in 2016—to AlphaZero, a successor system introduced in late 2017. The earlier AlphaGo system was trained initially on 30 million moves of Go drawn from games played by humans to gain an understanding of the rules.[41] Its successor, AlphaZero, entirely eliminated the need for any real-world data, replacing it with successive rounds of play against itself to achieve expert performance.[42] The result is similar to simulation learning: the creation of highly proficient systems of AI without a reliance on real-world data.

These research areas are noteworthy because they reduce the potential advantage of a competitor who either enjoys low-cost access to relevant data, or who already possesses significant endowments of data. Advancements in basic research could reduce the barrier to entry for training "data scarce" ML systems with comparable performance to those trained on larger datasets. This development may aid democratic governments in keeping pace with more autocratic competitors and may even help businesses and researchers in democratic societies keep up with counterparts in competitor nations.

Strategic Implications: Narrow Parity

Reducing the dependence of ML systems on immense, real-world datasets would encourage competitive parity between democratic states and authoritarian ones. However, there is an important caveat to this analysis.

Even if democracies make strong investments in reorienting ML toward research on achieving high performance in low data situations, they are likely to see only a narrow form of parity with their authoritarian competitors. Many of the identified areas of opportunity are effective only in particular circumstances. Even with major breakthroughs around some of these research challenges, existing trends do not suggest that these developments would cause the data barrier to entry to fall across all potential applications of ML at once.

Simulation learning is a good example of this limited form of strategic parity. The success of simulation learning depends on the level of similarity between the simulation and the eventual real-world operational environment an ML system will be deployed in. For example, training an ML agent to fly a drone in a virtual environment without wind may lead to poor performing systems in actual flight if windy conditions exist.

Simulation learning is particularly applicable to environments where many of the underlying principles are understood. Training robots to move through physical space and complete tasks like grasping demonstrates some of the most impressive applications of simulation learning because these simulations approximate real-world conditions. Simulation learning may prove less effective in creating models to accomplish tasks in domains where the underlying rules are less well understood or less conducive to quantification. Developing accurate proxy simulations for the expression of human emotions or group behavior, for instance, may prove challenging.[43] In these conditions, it may be challenging to train as successfully models on virtual data as ones trained on real-world data.

To that end, investments to accelerate progress in these data reduction technologies may enable democracies to more effectively compete at parity with authoritarian regimes, but not categorically across all potential applications. Instead, a narrow parity will be achieved under the specific circumstances where these techniques significantly change the data needed to achieve a high level of performance.

Moreover, realizing the full benefits of advancing these data-minimization techniques may require more than simply accelerating basic research. One common theme among many "few shot" simulation learning and self-play techniques is that they reduce the need for collecting real-world data while simultaneously increasing the need for computational power.

Rather than a physical robot navigating the world and collecting data from numerous trials, simulation learning generates this data from the actions of a virtual robot in virtual space. Self-play generates training data through the interaction of an agent with itself, rather than with some outside environment. Simulation learning and self-play methods free the designer of an ML system from the cost and complexity of real-world data acquisition in exchange for running simulations or contests between agents. These tasks all require access to high-performance computational power.

Democratic societies aiming to reduce the level of real-world data required to create competitive ML systems must ensure access to high-performance computational power. Universities, companies, and other actors advancing the state of the art in this domain will require plentiful and affordable computational power to make progress on these research problems and in the training and deployment of actual systems. This may require securing affordable access to corporate clouds, which can

provide computational infrastructure as a service, and a greater investment in creating secure computational infrastructure for training models on certain applications.

ENSURING DEMOCRATIC VIABILITY

Once ML systems are built, democratic commitments to public consent and civic participation require a core set of concerns to be addressed before the technology can be deployed in the field.

Although ML has enabled computers to accomplish an impressive range of tasks that they previously were unable to do, the technology still falls short in several important respects. For one, many modern ML systems lack what is referred to in the field as *interpretability*. Researchers have a somewhat limited understanding of how and why these systems produce the outputs that they do.[44] While it is clear, for instance, that a computer vision system can successfully recognize an object like a cat in an image, the process by which it arrives at this outcome is not always so clear.[45] Even more challenging, in some cases, attempts to make models more explainable will also reduce their performance.[46]

Second, ML systems are prone to learning spurious correlations during the training process. Such correlations cause them to exhibit "algorithmic bias," producing discriminatory and inequitable results when applied to people. For instance, in 2018, Amazon discovered an ML model the company used for filtering through resumes to identify promising job candidates discriminated against women.[47] The model was trained on historical resumes and hiring decisions disproportionately representing men. As a result, the model "penalized resumes that included the word 'women's' as in 'women's chess club captain,' and downgraded graduates of two all-women's colleges."[48]

Finally, many ML applications are privacy invasive. Training high-performance ML systems requires large datasets, and that data must be centralized to enable effective training. This tends to place power in the hands of a single entity such as a government or corporation, who might use this data for purposes beyond its original intent.

Lack of interpretability, algorithmic bias, and data centralization are all potent reasons for democratic publics to reject the deployment of AI systems. For example, consider how interpretability shapes public debate around the potential use of an ML system designed to predict crime and more efficiently allocate law enforcement resources.[49] Limits around interpretability can create practical issues in explaining how such a system actually comes to predict that crime will happen in a specific location. ML may only have limited tools available to conduct effective audits and identify the areas in which the system might make an error, or when it fails to work

at all. Moreover, a lack of understanding around how the system renders decisions can make it challenging to fix when errors emerge.

ML is adoptable only with a murky understanding of its inner-workings, reducing trust and making public consensus around the technology harder to build. This limitation hinders support for the use of ML in applications not just in the context of state functions like national defense, but throughout society. Legal requirements mandating a citizen's right to access a decision explanation may also prevent the technology's adoption.[50] At the very least, this opacity continually opens ML systems up to claims of bias, regardless of the actual impact of these systems in practice.[51]

Authoritarian states enjoy the relative advantage of being able to mandate the deployment of ML technologies throughout the economy and society in spite of these problems. Democratic states, in their preservation of an independent civil society and commitment to civic participation, have numerous mechanisms for citizens to push back on use of undesired ML technologies.

As a result, autocratic states can be "first to market" by forcing the deployment of these technologies through society. This advantage may provide economic and social benefits of ML systems earlier than states that must wait for buy-in from a hesitant public.

Similar to the challenge of immense datasets, one way to circumvent the limitations of democratic norms is to enact legal and policy reforms that smooth adoption of ML. This might include allowing the state and companies to deploy ML-driven technologies more aggressively without requiring public assent and adoption. But democracies should think twice before doing so: public consent may impose practical costs on the speed of adoption of ML technologies, but these costs are desirable in a broader sense. Like the erosion of privacy protections to make ML systems easier to train, these reforms would compromise democratic values in some degree to enable better competition.

An alternative approach is to consider ways in which ML may be made more democratically viable, such that requiring public consent does not hold back the technology. Might ML be reshaped to make it more amenable to public accountability and review? Might ML systems be deployed in ways that can ensure fairness, or avoid the problems of data centralization? Enabling ML to earn public consent and the political legitimacy that consent confers will be critical to ensuring that democracies can benefit from these technological breakthroughs. The approach of forcing AI-driven systems on the public or hiding its deployment risks a backlash that may hinder the implementation of the technology.

Questions of political legitimacy might seem entirely beyond the scope of technical ML research. Certainly, achieving a democratically viable form of ML will

require non-technical reforms in governance and regulation that offer the public a say in the development, design, and deployment of the technology. These reforms will play a major role in facilitating adoption of ML in democratic societies at large, even if the public chooses at moments to reject particular applications.

At the same time, technical details will exert a major influence over the democratic legitimacy of AI. In certain high-stakes arenas like healthcare or government administration, current ML systems may simply not be able to provide the degree of interpretability, fairness, and privacy protection demanded by the public in a democratic society. This poses a dilemma for the adoption of certain applications of AI, as they require either the wholesale rejection of the technology or its acceptance with a host of significant drawbacks.

Investment Opportunities

Democratic adoption will depend on advancing the ML field to meet public demands for interpretability, fairness, and privacy protection. Significant progress in these research areas would supply ways of meeting these concerns that were not previously possible. The end result would be to improve the practice of ML to better win the consent of a democratic public.

First, democratic governments should invest in the subfield of ML focused on the problem of interpretability. Broadly speaking, interpretability research seeks to uncover the internal processes governing how ML systems make decisions while retaining a high level of performance. More ambitiously, researchers seek to translate these internal processes into something useable by a non-expert.[52]

Expanding the range and sophistication of interpretability techniques would provide the public a better window into the processes by which ML systems automate decision-making. This would improve trust and create oversight options that may speed adoption. Government involvement could take a number of forms. Most obviously, it might take a role in directing funding toward research in interpretability, significantly augmenting existing programs working to advance "explainable artificial intelligence."[53] States could also play an important role in spearheading public challenges that raise the profile of certain key problems and incentivize targeted work on them.

Second, democratic governments should invest in advancing the subfield of fairness in ML, which seeks in part to find technical means of ensuring that AI systems can be designed to avoid the discriminatory and inequitable behaviors that may emerge during the training process.[54] For example, an ML system used to make credit and lending decisions may learn spurious correlations from the historical data the model is trained on.[55] The resulting model may systematically disfavor borrowers from certain minority groups based on racially discriminatory patterns of previous lending.

Researchers have developed a range of algorithmic definitions of fairness that can be subsequently implemented into ML models to avoid these kinds of discriminatory behaviors.[56] ML fairness research has also explored the application of a family of techniques known as causal modeling, which seek to extract causal relationships from observational data. Rather than simply predict, say, the recidivism of an offender in the criminal justice system, causal systems may suggest interventions to reduce the rate of crime going forwards—an arrangement more amenable to democratic publics and consistent with protecting the rights of the individual.[57]

Finally, democratic societies may also demand that the benefits of ML be obtained without undue harm to privacy. This might be made possible by reducing the overall data required for training through advancing the "few shot," simulation-based, and self-play learning methods previously described. But democracies might also secure privacy by advancing the state of the art of privacy preserving ML, which seeks to train systems without the need to directly access the raw data itself.

A few promising techniques are emerging in the research field for accomplishing this task. Researchers have been exploring a group of techniques known as homomorphic encryption that would enable ML systems to be effectively trained even on fully encrypted data.[58] This method might make democratic publics more willing to consent to their data being used, secure in the knowledge that their personal information remains inaccessible to those training models on that data.

More radically, the subfield of research into federated learning seeks to effectively train ML systems on data that is widely dispersed across many devices and anonymized through a method known as differential privacy.[59] Breakthroughs in this research may make ML more acceptable to democracies by unlocking different architectures for the technology which do not require centralizing massive corpora of data in a single location, regardless of whether or not it is encrypted.

Such techniques enhance the democratic viability of ML by providing the public and policymakers with an array of concrete options to ensure these technologies reflect democratic values.

Strategic Implications: Democratically Viable Artificial Intelligence

Resources are flowing to the research areas discussed above, driven by a mixture of commercial interest, public pressure, and researcher priorities. Interest in applying ML in high-stakes arenas has expanded the need to create more interpretable systems.[60] For instance, ML use in medicine promises to provide an early warning system to doctors about life-threatening conditions like kidney failure.[61] However, medical professionals have hesitated to adopt these systems in practice due to their "black box" nature, prompting companies to invest in interpretability research.[62]

High-profile failures of ML systems that highlight their tendency to reproduce bias have also encouraged companies to invest in creating more diverse datasets and to support research on fairness in the technology. In one recent incident, IBM received significant public criticism for its ML-based facial recognition system, which was found to significantly underperform on faces with darker skin tones.[63] This prompted IBM to release a "Diversity in Faces" (DiF) dataset in 2019.[64] DiF provided a "more balanced distribution and broader coverage of facial images compared to previous datasets," enabling ML researchers and engineers to address these problems in building future facial recognition technologies.

Despite this work, no organized attempt has emerged to marshal work on the problems of interpretability, fairness, and privacy toward a broader geopolitical objective. This vacuum provides an opportunity for democratic states to build alliances and lead.[65] States can play a role in articulating a unified vision of "democratically viable" AI and funding work in support of this research agenda, helping organize otherwise disparate efforts among researchers and encouraging greater collaboration. Companies may be particularly keen to engage in such an agenda, as it creates an outlet to engage with the government on ML development while avoiding more controversial work in the national security and military domains. Democracies also could play a role in fostering the development of training materials, which help to more rapidly percolate know-how about these techniques throughout the research and engineering communities.

Significant advances in these areas would help level the playing field with authoritarian regimes in the multi-faceted competition around AI. These tools would aid in the creation of high-performance ML systems with the capability to meet demands for transparency, accountability, and protection of social values. Citizens and businesses are more likely to trust these systems and consider them legitimate, enabling speedier integration of the technology and acquisition of its benefits by democratic societies.

Beyond the direct benefit of enabling easier ML adoption, fostering a democratically viable approach to AI offer collateral strategic impacts from a geopolitical perspective. Importantly, this strategy may raise expectations of publics globally that ML will be interpretable, fair, and privacy protecting.

Much of the competition around AI will be for market share: private companies based in various countries will compete aggressively to acquire users for their products and services both domestically and abroad. Widespread awareness about technical breakthroughs in subfields like interpretability, for instance, could cause consumers to demand that ML-driven products and services incorporate certain

transparency features as a default. This is especially the case if these advances overcome the long-standing sacrifice of system performance required to create systems with higher levels of interpretability.

Companies in markets that expect features like interpretability might enjoy a competitive edge over those in markets without the same demands. In this sense, advancing techniques for interpretability, fairness, and privacy might raise the barriers to entry into the markets of democracies, creating a kind of bulwark that favors home-grown products and services.

Moreover, large markets in democratic states with strong preferences for "democratically viable" ML products might force companies and states to invest more in these features than they otherwise would to access these users. Even markets with weak commitments to democratic values may prefer to work with foreign companies that can deliver transparent AI, as these systems may be easier to monitor and trust. The result may impose a cost on authoritarian states—who otherwise do not benefit strongly from advancements in these techniques—while creating know-how beneficial to the accelerated adoption of AI in democratic societies.

Liberal democracies face an unpalatable choice should they fail to invest in research on interpretability, fairness, and privacy. On one hand, democracies may need to slow technological adoption, delaying the economic and social gains that might otherwise be obtained through ML. This approach may put them behind authoritarian competitors that can more speedily force deployment and reap the benefit. On the other, democracies can proceed without regard for public consent, sparking significant resistance from citizens and sacrificing their values in the process.

SUBVERTING SOCIAL CONTROL APPLICATIONS

Recent breakthroughs in ML may be particularly attractive to authoritarian regimes in part because many applications seem well suited for expanding surveillance and enabling systems of social control. Democracies should invest in technologies that will hinder and thwart the use of ML for these purposes.

AI is well suited for applications that suppress dissent and sustain autocratic regimes. Advances in computer vision enable ever better tracking of individuals and relationships through surveillance footage.[66] Algorithms that predict social behavior may eventually work to enhance the effectiveness of "social credit" arrangements incentivizing and disincentivizing the public in ways that align with a regime's motives.[67] Natural language processing models can be leveraged to streamline and enhance regimes of censorship.[68] Even when these techniques replicate existing tools of surveillance and punishment, ML may efficiently automate and otherwise lower the costs of effectively implementing these systems of control.

These high-tech applications will soon "trickle down" to less technologically sophisticated autocrats. Leading companies in the deployment of ML for domestic social control purposes are beginning to market their services to aligned regimes around the world.[69] These tools are becoming commodified, lowering the financial resources and technical expertise needed to wield them.

Technologies for social control are critical to authoritarians, as such systems help neutralize dissent and enforce their policies. Autocrats will prioritize investments in developing and deploying ML in part because the technology promises to significantly enhance social control, thereby promoting regime stability and longevity. In contrast, democracies may not have such clear and urgent needs to advance AI, making it more difficult for them to keep pace over time.

How might democracies erode the strong authoritarian incentives to prioritize investment in ML? Democratic states may have an opportunity to go on the offensive by undermining the effectiveness of ML systems when authoritarian regimes deploy them. ML enhances states' ability to exert control over their populations, but ML research can simultaneously play a role in the creation of techniques that subvert these applications and make them more difficult to use effectively.

By giving populations knowledge about and tools for defeating the social control applications of AI, democratic societies could work to raise the floor of resistance to these uses globally. If successful, surveillance and social control uses would prove not just ineffectual but risky for a regime because citizens could easily exploit the vulnerabilities of the technology to evasion and manipulation.

Eroding the usefulness of AI for social control purposes would counter an important benefit the technology confers upon autocratic regimes. Accordingly, this strategy would work to reduce the attractiveness for authoritarian regimes to invest monetary and human capital in advancing the technology. These dynamics would aid democratic societies in maintaining their lead in the technology and in disproportionately benefiting from its development. While the economic opportunities ML offers mean that it is unlikely that these regimes would entirely halt their investments, this strategy may work to significantly suppress investment.

In the very least, authoritarian regimes might be compelled to allocate a greater portion of their resources toward ensuring that their systems are sufficiently resilient against attacks and advancing research on the topic. This would impose a cost they might not otherwise face, and potentially advance the field by improving the safety and security of the technology generally.

Investment Opportunities

Ongoing research has revealed the extent to which various ML systems are vulnerable to attacks causing them to produce faulty or otherwise undesired outputs.

Perhaps the most dramatic and widely reported example of this vulnerability has been the phenomenon of "adversarial examples." These inputs, when fed to an ML system, cause it to render inaccurate outputs. In one prominent example, a computer vision system was fooled into recognizing a panda as a gibbon.[70] These inputs can look innocuous to a trained eye since the attacks involve changing only a few pixels in an image, raising the possibility that these kinds of manipulations might happen under the noses of human operators. These attacks also take place elsewhere in the "lifecycle" of an ML system. For example, researchers have highlighted the problem of "data poisoning," in which the data leveraged for training a system introduces systematic biases subsequently exploitable by an attacker.[71] Adversarial examples have also been used to manipulate the tools researchers and engineers rely on to diagnose issues and examine the internal functioning of ML systems, potentially allowing attackers to hamper attempts to fix faulty systems in the future.[72]

This research presents a tantalizing opportunity for democracies to undermine the social control applications authoritarians stand to gain from leveraging ML. A surveillance system designed to analyze video footage for political dissidents might be tricked into missing a person of interest, or erroneously identifying regime loyalists as malefactors. Adversarial examples might make automated censorship systems highly "leaky," allowing prohibited news stories and political expression to flow through. Democracies should invest in accelerating the discovery of such vulnerabilities to erode or eliminate the usefulness of this technology in exerting social control.

Researchers have been working on a variety of approaches to fix or mitigate the risk from these potential vulnerabilities as they have been exposed. The most prominent approach involves a technique known as adversarial training, which incorporates adversarial inputs into the training process to enable the system to successfully overcome these attacks.[73] Other approaches include "feature squeezing," a set of processes applied to suspected adversarial inputs that can help identify the use of these tactics.[74]

Still, no categorical solution to the problem of adversarial examples appears to be on the horizon. Authoritarian regimes could not "patch" their systems to create resiliency against adversarial examples. As the authors of one survey article summarized the current state of the art, a "theoretical model of the adversarial example crafting process is very difficult to construct…[it is hard] to make any theoretical argument that a particular defense will rule out a set of adversarial examples."[75] Similar to developments in the realm of computer security, the cat-and-mouse game of discovering vulnerabilities and then repairing them will likely continue for the foreseeable future, if not indefinitely.

Strategic Implications: Managing Unintended Consequences

What specific role might the government take in attacking the effectiveness of social control applications of ML? How might democracies avoid having these methods of attack harm desirable uses of AI?

There is a great deal of ongoing research in this space. Companies deploying new products and services driven by AI have strong interests in securing these technologies from manipulation by malicious actors. Industrial labs accordingly have invested in recruiting top talent around these challenges, and the technical community at large has rallied around the topic. The number of prominent workshops and papers presented at major technical conferences testifies to the prioritization of this topic in the collective research agenda.[76]

In this respect, it is unclear if democratic states can play a strong role in significantly accelerating basic research activity in the space. Considerable energy on many of these topics within the ML field already exists. As a result, this may be an arena in which no serious underinvestment exists—government agencies, industrial labs, and academic institutions readily prioritize research on these topics. Vulnerabilities in these systems will be exposed rapidly, particularly as ML systems are increasingly trialed in high-stakes domains.

But this does not necessarily mean the state has no role to play. Quite the opposite, there may be a few arenas in which democratic governments might make a major impact in mobilizing the burgeoning field of ML security toward subverting the social control applications of the technology.

The primary challenge is that most vulnerabilities can be used to undermine both desirable and undesirable uses of ML. The techniques for subverting a computer vision system designed to suppress political dissidents can be identical to those recognizing cancerous cells.

The question, then, is whether it is possible to limit the unintended negative consequences. Democracies will benefit only insofar as they can ensure that the distribution of vulnerabilities hinders social control applications while mitigating the negative impact on other applications. Democratic governments can work to do this in two ways.

First, governments might simply work to highlight discoveries in this subdomain of security research, using their platform to bring attention to the fragile state of many of the ML systems that could be used to exert social control. This low-cost information campaign may erode the perceived value for autocratic regimes of developing and deploying these systems for surveillance and other purposes. Highlighting high-profile failures and the existence of persistent vulnerabilities might reduce regimes' trust in vendors attempting to sell them AI solutions. In the best pos-

sible case, the host of potential vulnerabilities may convince regimes to retain their higher-cost, legacy security infrastructure rather than bet on an untested technology.

Second, democratic governments might invest in shepherding technical discoveries about ML vulnerabilities into practical software that citizens could subsequently turn on oppressive systems of AI. While certain technical vulnerabilities might be demonstrated in a research setting, they may not be easily leveraged by populations likely to be subject to these technologies in autocratic regimes. The knowledge about these vulnerabilities remains largely relegated to specialized conferences or contained in dry technical literature that is inaccessible to the non-expert reader. There are also relatively low incentives for researchers and consumer web companies leading in ML to "productize" these vulnerabilities into user-friendly applications. The result is that those who are potentially most affected by ML enhanced surveillance or control cannot practically access the discoveries within the research community.

This gap may provide a means by which the impact of these discoveries may be more effectively targeted against specific social control uses of the technology. For example, democratic states might adapt research about adversarial examples into freely available clothing patterns that defeat gait and facial recognition systems, eroding the effectiveness of ML enhanced surveillance.[77] Similarly, democracies might invest in creating easy-to-use software that allows citizens to spread faulty data about themselves online, making it more challenging to create ML systems that accurately predict their behavior.

Beyond the creation of software, spreading know-how and building connections globally may be equally important. Democratic governments might set up knowledge sharing and similar programs to spread awareness about these techniques and build ties between the ML research community and activists resisting regimes on the ground. The coordination of activists to specifically thwart ML-driven surveillance in recent protests in Hong Kong suggests both awareness and demand for these practices among grassroots groups.[78]

By shaping global perceptions around the reliability of the technology for surveillance and control, and helping to translate these discoveries from the lab to actual use, democratic societies can take the lead in influencing the outcome of basic research in ML security as it diffuses more widely. This type of selective involvement might help to mitigate some of the double-edged nature of this research, and places governments in a position to complement existing activity in the research field.

Conclusion

How can democracies effectively compete against authoritarian regimes in the development and deployment of AI systems? The present state of ML imposes limitations on the ability of liberal democracies to do so effectively. Restrictions limiting the collection of training data, commitments to an independent civil society and press, and requirements for public consent all constrain how quickly ML systems can be trained and deployed. Authoritarian competitors do not face such constraints, and have strong incentives to prioritize ML investment in order to enhance systems of social control.

These factors give authoritarians an edge over democracies in AI competition. To keep up, democracies need to find a way to compete on more even footing without slowing technological adoption or sacrificing their values.

This paper offers one way through this dilemma. Democratic societies should work to achieve technical breakthroughs that mitigate the structural disadvantages that they face, while attacking the benefits that authoritarian regimes stand to gain with ML. This includes investments in three domains:

- **Reducing a dependence on data.** Authoritarian regimes may have structural advantages in marshaling data for ML applications when compared to their liberal democratic adversaries. To ensure better competitive parity, democracies should invest in techniques that reduce the scale of real-world data needed for training effective ML systems.
- **Fostering techniques that support democratic legitimacy.** Democracies may face greater friction in deploying ML systems

relative to authoritarian regimes due to their commitments to public participation, accountability, and rights. This requires them to invest in subfields of ML that enhance the viability of the technology in a democratic society, including work in interpretability, fairness, privacy, and causality.

- **Challenging the social control uses of ML.** Recent advances in AI appeal to authoritarian regimes in part because they promise to enhance surveillance and other mechanisms of control. Democracies should advance research eroding the usefulness of these applications, seeking to suppress investment by autocratic regimes and encourage research on robustness.

Targeted investments in these technical areas will work to level the playing field between democracies and their authoritarian competitors. Proposals to simply increase funding, expand the pool of talent, or free up datasets are laudable, but apply resources to AI as a broad, undifferentiated category. Without a close examination of the technical trends within ML, these strategies will expend significant resources ineffectually, redundantly, or most likely, in a diffuse way that fails to make any real difference from a geopolitical standpoint.

In shaping the terrain of AI competition, democracies may be able to achieve the best of both worlds: a future where democratic societies are able to fully obtain the social and economic benefits of AI while simultaneously preserving their core values in a rapidly changing technological landscape.

Endnotes

1. Steve Lohr, "Move Over, China: U.S. Is Again Home to World's Speediest Supercomputer," *New York Times*, June 8, 2018, https://www.nytimes.com/2018/06/08/technology/supercomputer-china-us.html.

2. Alexandra Zavi, "Going green for national security," *Los Angeles Times*, April 26, 2009, https://www.latimes.com/archives/la-xpm-2009-apr-26-me-army-green26-story.html.

3. R. Jeffrey Smith, "Hypersonic Missiles Are Unstoppable. And They're Starting a New Global Arms Race," *New York Times*, June 19, 2019, https://www.nytimes.com/2019/06/19/magazine/hypersonic-missiles.html.

4. Harold O. Levy and Jonathan Plucker, "Brains, Not Brawn," *U.S. News & World Report*, June 5, 2015, https://www.usnews.com/news/the-report/articles/2015/06/05/lack-of-stem-students-is-bad-for-national-security.

5. Cade Metz, "As China Marches Forward on A.I., the White House Is Silent," *New York Times*, February 12, 2018, https://www.nytimes.com/2018/02/12/technology/china-trump-artificial-intelligence.html.

6. Nina Xiang, "China's AI Industry Has Given Birth To 14 Unicorns: Is It A Bubble Waiting To Burst?," *Forbes*, October 5, 2018, https://www.forbes.com/sites/ninaxiang/2018/10/05/chinas-ai-industry-has-given-birth-to-14-unicorns-is-it-a-bubble-waiting-to-pop/.

7. Eva Xiao, "China's Most Addictive News App Toutiao Eyes World Domination with AI Feeds," *Tech in Asia*, December 5, 2017, https://www.techinasia.com/bytedance-overseas-expansion-strategy-breakdown/.

8. Masha Borak, "DJI and Microsoft Are Working on AI Drones That Can Recognize Objects," *Tech in Asia*, January 22, 2019, https://www.techinasia.com/dji-microsoft-working-ai-drones-recognize-objects.

9. Julian E. Barnes and Josh Chin, "The New Arms Race in AI," *Wall Street Journal*, March 2, 2018, https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261.

10. National Security Commission on Artificial Intelligence Act of 2018, S. 2806, 115th Cong. (2018), https://www.congress.gov/bill/115th-congress/senate-bill/2806/text.

11. "Artificial Intelligence and Global Security," Center for a New American Security, accessed April 7, 2020, https://www.cnas.org/artificial-intelligence-and-global-security/.

12. This example is a demonstration of *supervised learning*, one commonly used method for training ML systems. Other methods work somewhat differently, though they all depend on the availability of data, learning algorithms, computational power, and technical talent.

13. D. Sculley et al., *Machine Learning: The High Interest Credit Card of Technical Debt*, 2014, https://ai.google/research/pubs/pub43146.

14. "The Algorithm Kingdom; Artificial intelligence," *Economist*, July 15, 2017, https://www.economist.com/business/2017/07/15/china-may-match-or-beat-america-in-ai.

15. Richard Wike, Laura Silver and Alexandra Castillo, "Many People Around the World Are Unhappy With How Democracy Is Working," *Pew Research Center*, April 29, 2019, https://www.pewresearch.org/global/2019/04/29/many-across-the-globe-are-dissatisfied-with-how-democracy-is-working/.

16. Paul Mozur, "One Month, 500,000 Face Scans: How China is Using A.I. to Profile a Minority," *New York Times*, April 14, 2019, https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html/.

17. Larry Hardesty, "Study finds gender and skin-type bias in commercial artificial-intelligence systems," *MIT News*, February 11, 2018, https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212.

18. Kashmir Hill, "The Secretive Company That Might End Privacy as We Know It," *New York Times*, January 18, 2020, https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

19. Rachel Metz, "Beyond San Francisco, more cities are saying no to facial recognition," *CNN Business*, July 17, 2019, https://www.cnn.com/2019/07/17/tech/cities-ban-facial-recognition/index.html.

20. Javier Espinoza and Madhumita Murgia, "Sundar Pichai supports calls for moratorium on facial recognition," *Financial Times*, January 20, 2020, https://www.ft.com/content/0e19e81c-3b98-11ea-a01a-bae547046735.

21. Makena Kelly, "New facial recognition bill would require consent before companies could share data," *The Verge*, March 14, 2019, https://www.theverge.com/2019/3/14/18266249/facial-recognition-bill-data-share-consent-senate-commercial-facial-recognition-privacy-act.

22. Shazeda Ahmed, "The Messy Truth About Social Credit," *Logic*, May 1, 2019, https://logicmag.io/china/the-messy-truth-about-social-credit/.

23. Yarno Ritzen, "Laser beams, Twitter war: The tech side of Hong Kong protests," *Al Jazeera*, August 16, 2019, https://www.aljazeera.com/news/2019/08/laser-beams-twitter-war-tech-side-hong-kong-protests-190815113759044.html.

24. Andrew Imbrie, "Artificial Intelligence Meets Bureaucratic Politics," *War on the Rocks*, August 1, 2019, https://warontherocks.com/2019/08/artificial-intelligence-meets-bureaucratic-politics/.

25. *China's Pursuit of Emerging and Exponential Technologies, Before the House Armed Services Comm.*, 115th Cong., 2nd session. (2018), https://armedservices.house.gov/hearings?ID=EE29B55E-2DBE-4273-A084-EB76E2A56239.

26. Exec. Order No. 13,859, 84 Fed. Reg. 3967 (February 11, 2019), https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence.

27. Yoav Shoham et al., *AI Index 2018 Annual Report* (Stanford, CA: Human-Centered AI Initiative, 2018), http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf.

28. Recent examples include Fast.ai (https://www.fast.ai/) and Deeplearning.ai (https://www.deeplearning.ai/).

29. For an example of one-shot learning techniques, see Fei-Fei Li, Rob Fergus, and Pietro Perona, "One-shot learning of object categories." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (2006): 594-611, http://vision.stanford.edu/documents/Fei-FeiFergusPerona2006.pdf.

30. Alexey Kurakin, Ian Goodfellow and Samy Bengio. "Adversarial examples in the physical world." Preprint, submitted July 8, 2016. http://arxiv.org/abs/1607.02533.

31. Shoham, "AI Index", 26.

32. Roger Pielke, "In Retrospect: Science—The Endless Frontier." *Nature* 466 (2010): 922–923, https://www.nature.com/articles/466922a.

33. Vannevar Bush, "Science—the Endless Frontier," *National Science Foundation*, last accessed April 7, 2020, https://www.nsf.gov/od/lpa/nsf50/vbush1945.htm.

34. Gideon Lewis-Kraus, "The Great A.I. Awakening," *New York Times*, December 14, 2016, https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html.

35. Aäron van den Oord and Sander Dieleman, "WaveNet: A Generative Model for Raw Audio," DeepMind, last modified September 8, 2016, https://deepmind.com/blog/article/wavenet-generative-model-raw-audio.

36. Bruce Draper, "Learning with Less Labels (LwLL)," Defense Advanced Research Projects Agency (DARPA), last accessed April 7, 2020, https://www.darpa.mil/program/learning-with-less-labels.

37. For a general overview of these methods see Yaqing Wang et al. "Generalizing from a Few Examples: A Survey on Few-Shot Learning." Preprint, submitted April 10, 2019. http://arxiv.org/abs/1904.05046.

38. Christoph H. Lampert, Hannes Nickisch and Stefan Harmeling, "Learning to detect unseen object classes by between class attribute transfer," Conference on Computer Vision and Pattern Recognition 2009, last accessed April 7, 2020, https://dblp.org/rec/conf/cvpr/LampertNH09.html.

39. Shixiang Gu et al. "Deep Reinforcement Learning for Robotic Manipulation with Asynchronous Off-Policy Updates." Preprint, submitted October 3, 2016. http://arxiv.org/abs/1610.00633.

40. Xue Bin Peng, et al., "Generalizing from Simulation," OpenAI, last modified October 19, 2017, https://openai.com/blog/generalizing-from-simulation/; Ilge Akkaya, et al., "Solving Rubik's Cube with a Robot Hand," OpenAI, last modified October 15, 2019, https://openai.com/blog/solving-rubiks-cube/.

41. Cade Metz, "Inside the Epic Go Tournament Where Google's AI Came to Life," *Wired*, May 19, 2016, https://www.wired.com/2016/05/google-alpha-go-ai/.

42. David Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm." Preprint, submitted December 5, 2017. http://arxiv.org/abs/1712.01815.

43. For a discussion of our limited understanding of how and when expressions reflect underlying emotion, see Angela Chen, "Computers can't tell if you're happy when you smile," *MIT Technology Review*, July 26, 2019, https://www.technologyreview.com/s/614015/emotion-recognition-technology-artifical-intelligence-inaccurate-psychology/.

44. Finale Doshi-Velez and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning." Preprint, submitted February 28, 2017. https://arxiv.org/abs/1702.08608v2.

45. Chris Olah et al., "The Building Blocks of Interpretability," *Distill*, March 6, 2018, https://distill.pub/2018/building-blocks/.

46. Leilani H. Gilpin et al., "Explaining Explanations: An Overview of Interpretability of Machine Learning." *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* 80–89, https://ieeexplore.ieee.org/document/8631448/ ("The most accurate explanations are not easily interpretable to people; and conversely the most interpretable descriptions often do not provide predictive power.")

47. Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, October 9, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

48. Dastin, "Amazon scraps."

49. For a review of these systems and their failures, see Jessica Saunders, "Pitfalls of Predictive Policing," RAND, last modified October 11, 2016, https://www.rand.org/blog/2016/10/pitfalls-of-predictive-policing.html.

50. Danielle Keats Citron, "Technological Due Process." *Washington University Law Review*, 85 (2007): 1249-1313, https://ssrn.com/abstract=1012360; Andrew D. Selbst and Julia Powles "Meaningful Information and the Right to Explanation." *International Data Privacy Law 7*, no. 4: 233-242, https://ssrn.com/abstract=3039125.

51. Megan T. Stevenson, "Assessing Risk Assessment in Action," *Minnesota Law Review* 303 (2019): 103, https://ssrn.com/abstract=3016088 (finding that algorithmic risk assessment systems had limited impact, potentially because judges simply ignored these systems in practice).

52. For an exploration of the many facets of improving interpretability in machine learning, see Zachary Lipton. "The Mythos of Model Interpretability." Preprint, submitted June 10, 2016. https://arxiv.org/abs/1606.03490.

53. Matt Turek, "Explainable Artificial Intelligence," Defense Advanced Research Projects Agency (DARPA), https://www.darpa.mil/program/explainable-artificial-intelligence.

54. Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning* (2019), https://fairmlbook.org/.

55. Andreas Fuster, et al. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." Preprint, submitted November 17, 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3072038.

56. For a more in-depth discussion of these fairness definitions, see Arvid Narayanan, "Tutorial: 21 Fairness Definitions and Their Politics," YouTube, https://www.youtube.com/watch?v=jIXIuYdnyyk.

57. Chelsea Barabas et al. "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment." Preprint, submitted December 21, 2017. http://arxiv.org/abs/1712.08238.

58. Thore Graepel, Kristin Lauter and Michael Naehrig, "ML Confidential: Machine Learning on Encrypted Data" in *Information Security and Cryptology*, eds/ Taekyoung Kwon et al. (Berlin: Springer, 2013), 1-21; Raphael Bost et al., "Machine Learning Classification over Encrypted Data," Proceedings 2015 Network and Distributed System Security Symposium, https://www.ndss-symposium.org/ndss2015/ndss-2015-programme/machine-learning-classification-over-encrypted-data/.

59. Jakub Konečný et al. "Federated Learning: Strategies for Improving Communication Efficiency." Preprint, submitted October 18, 2016. http://arxiv.org/abs/1610.05492.

60. Judea Pearl. "Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution." Preprint, submitted January 11, 2018. http://arxiv.org/abs/1801.04016.

61. Mustafa Suleyman, "Using AI to give doctors a 48-hour head start on life-threatening illness," *DeepMind* (blog), July 31, 2019, https://deepmind.com/blog/article/predicting-patient-deterioration.

62. Suleyman, "Using AI."

63. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77-91, http://proceedings.mlr.press/v81/buolamwini18a.html.

64. Sara Robinson "Building ML models for everyone," Google Cloud, last modified September 25, 2019, https://cloud.google.com/blog/products/ai-machine-learning/building-ml-models-for-everyone-

understanding-fairness-in-machine-learning/; John R. Smith, "IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems," IBM Research, last modified January 29, 2019, https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/.

65. This may complement related strategies focused on building alliances between democracies to enhance AI competitiveness. Andrew Imbrie et al., *Agile Alliances: How the United States and Its Allies Can Deliver a Democratic Way of AI* (Washington, DC: Center for Security and Emerging Technology, 2020), http://cset.georgetown.edu/agile-alliances/; Martijn Rasser et al., *The American AI Century: A Blueprint for Action* (Washington, DC: The Center for a New American Security, 2019), https://www.cnas.org/publications/reports/the-american-ai-century-a-blueprint-for-action.

66. Paul Mozur, "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority," *New York Times*, April 14, 2019, https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

67. Louise Matsakis, "How the West Got China's Social Credit System Wrong," *Wired*, July 29, 2019, https://www.wired.com/story/china-social-credit-score-system/.

68. For a discussion of some of the potential implications of these systems on expression, see Natasha Duarte, et al., "Mixed Messages: The Limits of Automated Social Media Content Analysis," Center for Democracy and Technology, last modified November 13 2017, https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf.

69. Steve Feldstein, *The Global Expansion of AI Surveillance* (Washington, DC: Carnegie Endowment for International Peace, 2019), https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847.

70. Ian Goodfellow et al., "Attacking Machine Learning with Adversarial Examples," OpenAI, last modified February 24, 2017, https://openai.com/blog/adversarial-example-research/.

71. Jacob Steinhardt et al. "Certified Defenses for Data Poisoning Attacks." Preprint, submitted June 9, 2017. https://arxiv.org/abs/1706.03691.

72. Pieter-Jan Kindermans et al. "The (Un)reliability of Saliency Methods." Preprint, submitted November 2, 2017. http://arxiv.org/abs/1711.00867; Amirata Ghorbani, Abubakar Abid and James Zou. "Interpretation of Neural Networks is Fragile." Preprint, submitted October 29, 2017. http://arxiv.org/abs/1710.10547.

73. Florian Tramèr et al. "Ensemble Adversarial Training: Attacks and Defenses." Preprint, submitted May 19, 2017. http://arxiv.org/abs/1705.07204.

74. Anirban Chakraborty et al. "Adversarial Attacks and Defences: A Survey." Preprint, submitted September 28, 2018. http://arxiv.org/abs/1810.00069.

75. Chakraborty, "Adversarial Attacks".

76. "NeurIPS 2018 Workshop on Security in Machine Learning," Github, last modified December 7, 2018, https://secml2018.github.io/; "Defending Against Adversarial Artificial Intelligence," Defense Advanced Research projects Agency (DARPA), last modified February 6, 2019, https://www.darpa.mil/news-events/2019-02-06.

77. "Face masks to decoy t-shirts: The rise of anti-surveillance fashion," *Reuters*, September 26, 2019, https://www.reuters.com/article/us-britain-tech-fashion-feature-idUSKBN1WB0HT.

78. Paul Mozur, "In Hong Kong Protests, Faces Become Weapons," *New York Times*, July 26, 2019, https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html.

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

CSET.GEORGETOWN.EDU | CSET@GEORGETOWN.EDU