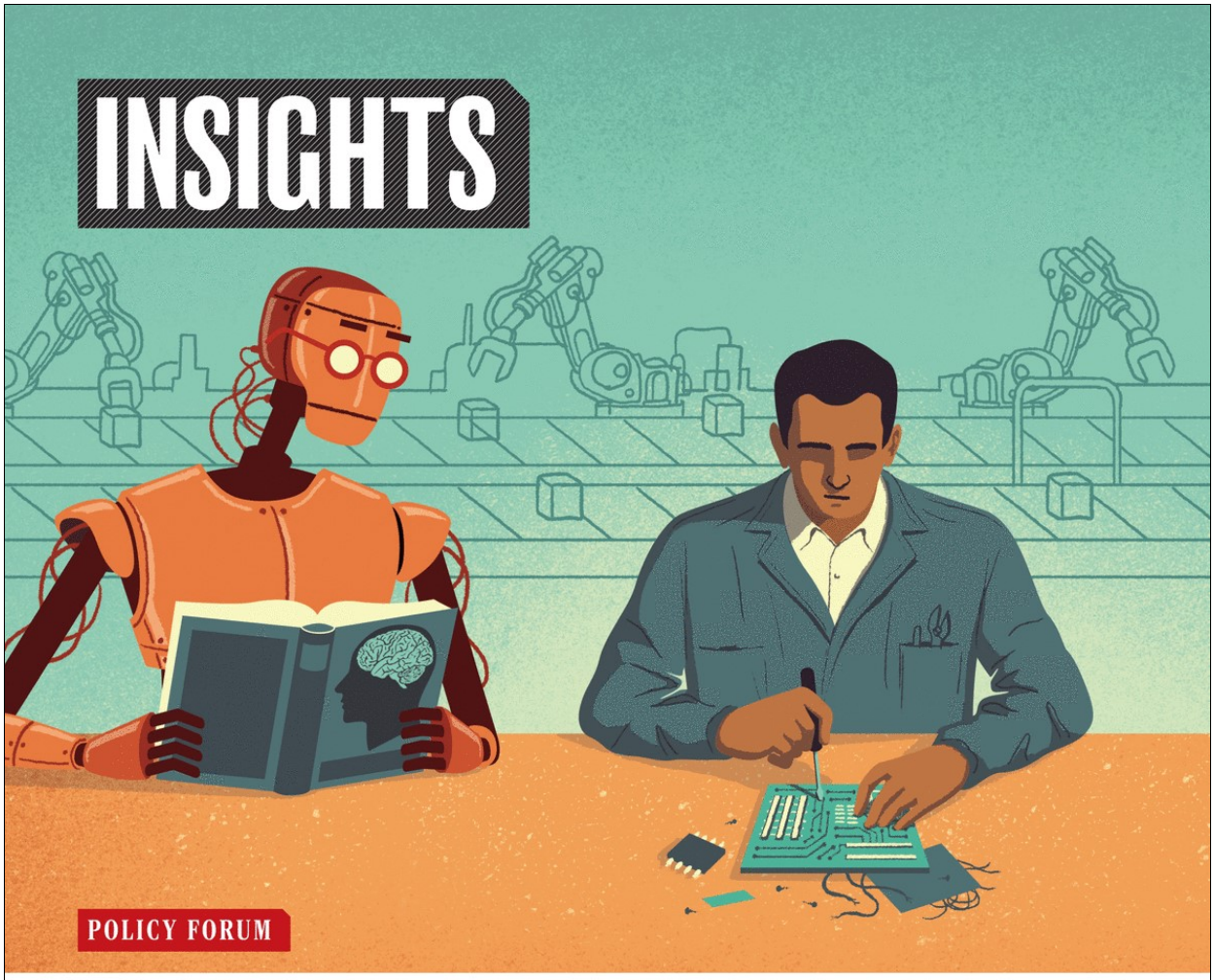


TECHNOLOGY AND THE ECONOMY

What can machine learning do? Workforce implications

Profound change is coming, but roles for humans remain

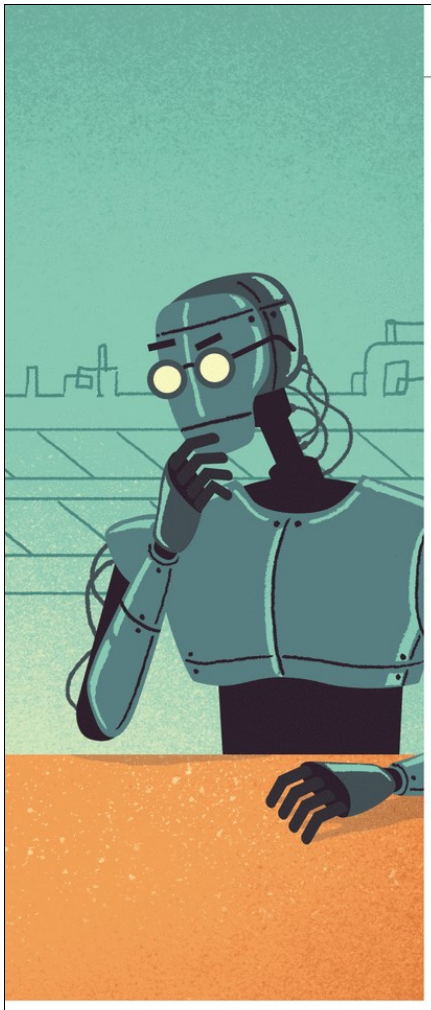
By **Erik Brynjolfsson**[1,2] *and* **Tom Mitchell**[3]

Digital computers have transformed work in almost every sector of the economy over the past several decades (*1*). We are now at the beginning of an even larger and more rapid transformation due to recent advances in machine learning (ML), which is capable of accelerating the pace of automation itself. However, although it is clear that ML is a "general purpose technology," like the steam engine and electricity, which spawns a plethora of additional innovations and capabilities (*2*), there is no widely shared agreement on the tasks where ML systems excel, and thus little agreement on the specific expected impacts on the workforce and on the economy more broadly. We discuss what we see to be key implications for the workforce, drawing on our rubric of what the current generation of ML systems can and cannot do [see the supplementary materials (SM)]. Although parts of many jobs may be "suitable for ML" (SML), other tasks within these same jobs do not fit the criteria for ML well; hence, effects on employment are more complex than the simple replacement and substitution story emphasized by some. Although economic effects of ML are relatively limited today, and we are not facing the imminent "end of work" as is sometimes proclaimed, the implications for the economy and the workforce going forward are profound.

Any discussion of what ML can and cannot do, and how this might affect the economy, should first recognize two broad, underlying considerations. We remain very far from artificial general intelligence (*3*). Machines cannot do the full range of tasks that humans can do (*4*). In addition, although innovations

[1]*Sloan School of Management, Massachussetts Institute of Technology, Cambridge, MA 02139, USA.* [2]*National Bureau of Economic Research, Cambridge, MA 02138, USA.* [3]*Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: erikb@ mit.edu*

ILLUSTRATION: DAVIDE BONAZZI/@SALZMANART



generally have been important for overall improvements in income and living standards, and the first wave of pre-ML information technology (IT) systems in particular has created trillions of dollars of economic value, "The case that technological advances have contributed to wage inequality is strong" [see (*1*), a report from a committee we recently cochaired for the U.S. National Academies of Science, Engineering and Medicine]. Although there are many forces contributing to inequality, such as increased globalization, the potential for large and rapid changes due to ML, in many cases within a decade, suggests that the economic effects may be highly disruptive, creating both winners and losers. This will require considerable attention among policy-makers, business leaders, technologists, and researchers.

As machines automate some of the tasks that are SML in a particular job or process, the remaining tasks that are non-SML may become more valuable. In other cases, machines will augment human capabilities and make possible entirely new products, services, and processes. Therefore, the net effect on the demand for labor, even within jobs that are partially automated, can be either negative or positive. Although broader economic effects can be complex, labor demand is more likely to fall for tasks that are close substitutes for capabilities of ML, whereas it is more likely to increase for tasks that are complements for these systems. Each time an ML system crosses the threshold where it becomes more cost-effective than humans on a task, profit-maximizing entrepreneurs and managers will increasingly seek to substitute machines for people. This can have effects throughout the economy, boosting productivity, lowering prices, shifting labor demand, and restructuring industries.

WE KNOW MORE THAN WE CAN TELL

As the philosopher Polanyi observed, we know more than we can tell (*5*). Recognizing a face, riding a bike, and understanding speech are tasks humans know very well how to do, but our ability to reflect on how we perform them is poor. We cannot codify many tasks easily, or perhaps at all, into a set of formal rules. Thus, prior to ML, Polanyi's paradox limited the set of tasks that could be automated by programming computers (*6*). But today, in many cases, ML algorithms have made it possible to train computer systems to be more accurate and more capable than those that we can manually program.

Until recently, creating a new computer program involved a labor-intensive process of manual coding. But this expensive process is increasingly being augmented or replaced by a more automated process of running an existing ML algorithm on appropriate training data. The importance of this shift is twofold. In a growing subset of applications, this paradigm can produce more accurate and reliable programs than human programmers (e.g., face recognition and credit card fraud detection). Second, this paradigm can dramatically lower costs for creating and maintaining new software. This lowered cost reduces the barrier to experiment with and explore potential computerization of tasks, and encourages development of computer systems that will automatically automate many types of routine workflows with little or no human intervention.

Such progress in ML has been particularly rapid in the past 6 to 8 years due in large part to the sheer volume of training data available for some tasks, which may be large enough to capture highly valuable and previously unnoticed regularities—perhaps impossibly large for a person to examine or comprehend, yet within the processing ability of ML algorithms. When large enough training data sets are available, ML can sometimes produce computer programs that outperform the best humans at the task (e.g., dermatology diagnosis, the game of Go, detecting potential credit card fraud).

Also critical to ML progress has been the combination of improved algorithms, including deep neural networks (DNNs) and considerably faster computer hardware. For example, Facebook switched from phrase-based machine translation models to DNNs for more than 4.5 billion language translations each day. DNNs for image recognition have driven error rates on ImageNet, a large data set of more than 10,000 labeled images (*7*), down from more than 30% in 2010 to less than 3% today. Similarly, DNNs have helped improve error rates from 8.4% to 4.9% in voice recognition since July 2016. The 5% threshold for image recognition and speech is important because that is roughly the error rate of humans when given similar data.

AUTOMATING AUTOMATION

To produce a well-defined learning task to which we can apply a ML algorithm, one must fully specify the task, performance metric, and training experience. In most practical applications, the task to be learned corresponds to some target function, such as a function from input medical patient health records to output patient diagnoses, or a function from the current sensor inputs of a self-driving car to the correct next steering command. The most common type of training experience is data consisting of input-output pairs for the target function (e.g., medical records paired with the correct diagnoses). Obtaining ground-truth training data can be difficult in many domains, such as psychiatric diagnosis, hiring decisions, and legal cases.

Key steps in a successful commercial application typically include efforts to identify precisely the function to be learned; collect and cleanse data to render it useable for training the ML algorithm; engineer data features to choose which are likely to be helpful in predicting the target output, and perhaps to collect new data to make up for shortfalls in the original features collected; experiment with different algorithms and parameter settings to optimize the accuracy of learned classifiers; and embed the resulting learned system into routine business operations in a way that improves productivity and, if possible, in a way that captures additional training examples on an ongoing basis.

One approach that is particularly relevant to gauging the rate of future automation is the "learning apprentice" (sometimes called the "human in the loop") approach (*8*), in which the artificial intelligence (AI) program acts as an apprentice to assist the

human worker, while also learning by observing the human's decisions and capturing these as additional training examples. This approach has led to new kinds of business models.

Training a learning apprentice to mimic human-generated decisions offers the potential for machines to learn from the combined data of multiple people it assists, perhaps leading to outperforming each individual on the team that trains it. Still, its learned expertise may be limited by the skill level of the human team and by the online availability of relevant decision variables. However, in cases where the computer can also access independent data to determine the optimal decision (ground truth), it may be possible to improve on human decisions and then to help the human improve their own performance. For example, in medical diagnosis of skin cancer from dermatological images, using the results of subsequent biopsies as a gold standard for training can produce computer programs with even higher diagnostic accuracies than human doctors (*9*).

MOST SUITABLE TASKS

Although recent advances in the capabilities of ML systems are impressive, they are not equally suitable for all tasks. The current wave of successes draw particularly heavily on a paradigm known as supervised learning, typically using DNNs. They can be immensely powerful in domains that are well suited for such use. However, their competence is also dramatically narrower and more fragile than human decision-making, and there are many tasks for which this approach is completely ineffective. Of course, advances in ML continue, and other approaches are likely to be better suited for different types of tasks. We identify eight key criteria that help distinguish SML tasks from tasks where ML is less likely to be successful, at least when using the currently dominant ML paradigm (see the SM for a more detailed, 21-item rubric).

1. **Learning a function that maps well-defined inputs to well-defined outputs**
   Among others, these include classification (e.g., labeling images of dog breeds or labeling medical records according to the likelihood of cancer) and prediction (e.g., analyzing a loan application to predict the likelihood of future default). Although ML may learn to predict the Y value associated with any given input X, this is a learned statistical correlation that might not capture causal effects.

2. **Large (digital) data sets exist or can be created containing input-output pairs**
   The more training examples are available, the more accurate the learning. One of the remarkable characteristics of DNNs is that performance in many domains does not seem to asymptote after a certain number of examples (*10*). It is especially important that all of the relevant input features be captured in the training data. Although in principle any arbitrary function can be represented by a DNN (*11*), computers are vulnerable to mimicking and perpetuating unwanted biases present in the training data and to missing regularities that involve variables that they cannot observe. Digital data can often be created by monitoring existing processes and customer interactions, by hiring humans to explicitly tag or label portions of the data or create entirely new data sets, or by simulating the relevant problem setting.

3. **The task provides clear feedback with clearly definable goals and metrics**
   ML works well when we can clearly describe the goals, even if we cannot necessarily define the best process for achieving those goals. This contrasts with earlier approaches to automation. The ability to capture input-output decisions of individuals, although it might allow learning to mimic those individuals, might not lead to optimal system-wide performance because the humans themselves might make imperfect decisions. Therefore, having clearly defined system-wide metrics for performance (e.g., to optimize traffic flow throughout a city rather than at a particular intersection) provides a gold standard for the ML system. ML is particularly powerful when training data are labeled according to such gold standards, thereby defining the desired goals.

4. **No long chains of logic or reasoning that depend on diverse background knowledge or common sense**
   ML systems are very strong at learning empirical associations in data but are less effective when the task requires long chains of reasoning or complex planning that rely on common sense or background knowledge unknown to the computer. Ng's "one-second rule" (*4*) suggests that ML will do well on video games that require quick reaction and provide instantaneous feedback but less well on games where choosing the optimal action depends on remembering previous events distant in time and on unknown background knowledge about the world (e.g., knowing where in the room a newly introduced item is likely to be found) (*12*). Exceptions to this are games such as Go and chess, because these nonphysical games can be rapidly simulated with perfect accuracy, so that millions of perfectly self-labeled training examples can be automatically collected. However, in most real-world domains, we lack such perfect simulations.

A heat exchanger was designed by a machine using generative design.

PHOTO: AUTODESK

5. **No need for detailed explanation of how the decision was made**
   Large neural nets learn to make decisions by subtly adjusting up to hundreds of millions of numerical weights that interconnect their artificial neurons. Explaining the reasoning for such decisions to humans can be difficult because DNNs often do not make use of the same intermediate abstractions that humans do. While work is under way on explainable AI systems (*13*), current systems are relatively weak in this area. For example, whereas computers can diagnose certain

types of cancer or pneumonia as well as or better than expert doctors, their ability to explain why or how they came up with the diagnosis is poor when compared with human doctors. For many perceptual tasks, humans are also poor at explaining, for example, how they recognize words from the sounds they hear.

6. **A tolerance for error and no need for provably correct or optimal solutions**
Nearly all ML algorithms derive their solutions statistically and probabilistically. As a result, it is rarely possible to train them to 100% accuracy. Even the best speech, object recognition, and clinical diagnosis computer systems make errors (as do the best humans). Therefore, tolerance to errors of the learned system is an important criterion constraining adoption.

7. **The phenomenon or function being learned should not change rapidly over time**
In general, ML algorithms work well only when the distribution of future test examples is similar to the distribution of training examples. If these distributions change over time, then retraining is typically required, and success therefore depends on the rate of change, relative to the rate of acquisition of new training data (e.g., email spam filters do a good job of keeping up with adversarial spammers, partly because the rate of acquisition of new emails is high compared to the rate at which spam changes).

8. **No specialized dexterity, physical skills, or mobility required**
Robots are still quite clumsy compared with humans when dealing with physical manipulation in unstructured environments and tasks. This is not so much a shortcoming of ML but instead a consequence of the state of the art in general physical mechanical manipulators for robots.

WORKFORCE IMPLICATIONS

The main effects of pre-ML IT have been on a relatively narrow swath of routine, highly structured and repetitive tasks (*14*). This has been a key reason that labor demand has fallen for jobs in the middle of the skill and wage spectrum, like clerks and factory workers, whereas demand at the bottom (e.g., janitor or home health aide) and top (e.g., physicians) has held up in most advanced countries (*15*). But a much broader set of tasks will be automated or augmented by machines over the coming years. This includes tasks for which humans are unable to articulate a strategy but where statistics in data reveal regularities that entail a strategy. Although the framework of routine versus nonroutine tasks did a very effective job of describing tasks suitable for the last wave of automation (*14*), the set of SML tasks is often very different. Thus, simply extrapolating past trends will be misleading, and a new framework is needed.

Jobs typically consist of a number of distinct but interrelated tasks. In most cases, only some of these tasks are likely to be suitable for ML, and they are not necessarily the ones that were easy to automate with previous technologies. For instance, when we apply our 21-question SML rubric to various occupations, we find that a ML system can be trained to help lawyers classify potentially relevant documents for a case but would have a much harder time interviewing potential witnesses or developing a winning legal strategy (*16*). Similarly, ML systems have made rapid advances in reading medical images, outperforming humans in some applications (*17*). However, the more unstructured task of interacting with other doctors, and the potentially emotionally fraught task of communicating with and comforting patients, are much less suitable for ML approaches, at least as they exist today.

That is not to say that all tasks requiring emotional intelligence are beyond the reach of ML systems. One of the surprising implications of our rubric is that some aspects of sales and customer interaction are potentially a very good fit. For instance, transcripts from large sets of online chats between salespeople and potential customers can be used as training data for a simple chatbot that recognizes which answers to certain common queries are most likely to lead to sales (*18*). Companies are also using ML to identify subtle emotions from videos of people.

Another area where the SML rubric departs from the conventional framework is in tasks that may involve creativity. In the old computing paradigm, each step of a process needed to be specified in advance with great precision. There was no room for the machine to be "creative" or figure out on its own how to solve a particular problem. But ML systems are specifically designed to allow the machine to figure out solutions on its own, at least for SML tasks. What is required is not that the process be defined in great detail in advance but that the properties of the desired solution be well specified and that a suitable simulator exists so that the ML system can explore the space of available alternatives and evaluate their properties accurately. For instance, designing a complex new device has historically been a task where humans are more capable than machines. But generative design software can come up with new designs for objects like the heat exchanger (see photo) that meet all the requirements (e.g., weight, strength, and cooling rate) more effectively than anything designed by a human, and with a very different look and feel (*18*).

Is it "creative"? That depends on what definition one uses. But some "creative" tasks that were previously reserved for humans will be increasingly automatable in the coming years. This approach works well when the final goal can be well specified and the solutions can be automatically evaluated as clearly right or wrong, or at least better or worse. As a result, we can expect such tasks to be increasingly subject to automation. At the same time, the role of humans in more clearly defining goals will become more important, suggesting an increased role for scientists, entrepreneurs, and those making a contribution by asking the right questions, even if the machines are often better able to find the solutions to those questions once they are clearly defined.

SIX ECONOMIC FACTORS

There are many nontechnological factors that will affect the implications of ML for the workforce. Specifically, the total effect of ML on labor demand and wages can be written as a function of six distinct economic factors:

1. **Substitution**
Computer systems created by ML will directly substitute for some tasks, replacing the human and reducing labor demand for any given level of output

2. **Price elasticity**
Automation via machine learning may lower prices for tasks. This can lead to lower or higher total spending, depending on the price elasticity of demand. For instance, if elasticity is less than –1, then a decrease in price leads to a more than proportional increase in quantity purchased, and total spending (price times quantity) will increase. By analogy, as technology reduced the price of air travel after 1903, total spending on this type of travel increased, as did employment in this industry.

3. **Complementarities**
Task B may be an important, or even indispensable, complement to another task A that is automated. As the price of A falls, the demand for B will increase. By analogy, as calculation became automated, the demand for human programmers increased. Skills can also be complementary to other skills. For instance, interpersonal skills are increasingly complementary to analytical skills (*19*).

4. **Income elasticity**
   Automation may change the total income for some individuals or the broader population. If income elasticity for a good is nonzero, this will in turn change demand for some types of goods and the derived demand for the tasks needed to produce those goods. By analogy, as total income has increased, Americans have spent more of their income on restaurant meals.

5. **Elasticity of labor supply**
   As wages change, the number of people working on the task will respond. If there are many people who already have the requisite skills (for example, driving a car for a ride-hailing service), then supply will be fairly elastic and wages will not rise (or fall) much, if at all, even if demand increases (or falls) a lot. In contrast, if skills are more difficult to acquire, such as becoming a data scientist, then changes in demand will mainly be reflected in wages, not employment.

6. **Business process redesign**
   The production function that relates any given set of different types and quantities of labor, capital, and other inputs to output is not fixed. Entrepreneurs, managers, and workers constantly work to reinvent the relevant processes. When faced with new technologies, they will change the production process, by design or through luck, and find more efficient ways to produce output (*20*). These changes can take time and will often economize on the most expensive inputs, increasing demand elasticity. Similarly, over time, individuals can make a choice to respond to higher wages in some occupations or places by investing in developing the new skills required for work or moving to a new location, increasing the relevant supply elasticity. Thus, according to Le Chatelier's principle (*21*), both demand and supply elasticities will tend to be greater in the long run than in the short run as quasi-fixed factors adjust.

Adoption and diffusion of technologies often take years or decades because of the need for changes in production processes, organizational design, business models, supply chains, legal constraints, and even cultural expectations. Such complementarities are as ubiquitous in modern organizations and economies as they are subtle and difficult to identify, and they can create considerable inertia, slowing the implementation of even—or especially—radical new technologies (*22*). Applications that require complementary changes on many dimensions will tend to take longer to affect the economy and workforce than those that require less redesign of existing systems. For instance, integration of autonomous trucks onto city streets might require changes in traffic laws, liability rules, insurance regulations, traffic flow, and the like, whereas the switch from talking to a human assistant to a virtual assistant in a call center might require relatively little redesign of other aspects of the business process or customer experience.

Over time, another factor becomes increasingly important: New goods, services, tasks, and processes are always being invented. These inventions can lead to the creation of altogether new tasks and jobs (*23*) and thus can change the magnitudes and signs of the above relationships. Historically, as some tasks have been automated, the freed-up labor has been redeployed to producing new goods and services or new, more effective production processes. Such innovations have been more important than increased capital, labor, or resource inputs as a force for raising overall incomes and living standards. ML systems may accelerate this process for many of the tasks that fit the criteria above by partially automating automation itself.

"Applications that require... changes on many dimensions will tend to take longer to affect the economy..."

As more data come online and are pooled and as we discover which tasks should be automated by ML, we will collect data even more rapidly to create even more capable systems. Unlike solutions to tasks mastered by humans, many solutions to tasks automated by ML can be disseminated almost instantly worldwide. There is every reason to expect that future enterprise software systems will be written to embed ML in every online decision task, so that the cost of attempting to automate will come down even further.

The recent wave of supervised learning systems have already had considerable economic impact. The ultimate scope and scale of further advances in ML may rival or exceed that of earlier general-purpose technologies like the internal combustion engine or electricity. These advances not only increased productivity directly but, more important, triggered waves of complementary innovations in machines, business organization, and even the broader economy. Individuals, businesses, and societies that made the right complementary investments—for instance, in skills, resources, and infrastructure—thrived as a result, whereas others not only failed to participate in the full benefits but in some cases were made worse off. Thus, a better understanding of the precise applicability of each type of ML and its implications for specific tasks is critical for understanding its likely economic impact. ■

REFERENCES AND NOTES

1. National Academies of Sciences, Engineering, and Medicine, *Information Technology and the U.S. Workforce: Where Are We and Where Do We Go from Here?* (National Academies Press, Washington, DC, 2017).
2. E. Brynjolfsson, D. Rock, C. Syverson, "Artificial Intelligence and the Modern Productivity Paradox: A Class of Expectations and Statistics," NBER Working Paper 24001 (National Bureau of Economic Research, Cambridge, MA, 2017).
3. S. Legg, M. Hutter, *Frontiers in Artificial Intelligence and Applications* **157**, 17 (2007).
4. A. Ng, What artificial intelligence can and can't do right now, *Harvard Business Rev.* (9 November 2016).
5. M. Polanyi, *The Tacit Dimension* (University of Chicago Press, Chicago, 1966).
6. D. Autor, Polanyi's paradox and the shape of employment growth; presentation to the Federal Reserve Bank of Kansas City's Jackson Hole Central Banking Conference (2014).
7. J. Deng *et al.*, Imagenet: A large-scale hierarchical image database. Computer Vision and Pattern Recognition, 2009. IEEE Conference on. IEEE, 2009. [ImageNet (most recent competition): http://image-net.org/challenges/LSVRC/2017/results]
8. T. Mitchell, S. Mahadevan, L. Steinberg, LEAP: A learning apprentice for VLSI design, in *ML: An Artificial Intelligence Approach*, vol. III, Y. Kodratoff, R. Michalski, Eds. (Morgan Kaufmann Press, 1990).
9. A. Esteva *et al.*, *Nature* **542**, 115 (2017).
10. A. Coates *et al.*, Deep learning with COTS HPC systems. In International Conference on ML (2013), pp. 1337–1345.
11. G. Cybenko, *Mathematics of Control, Signals, and Systems* **2**, 303 (1989).
12. V. Mnih *et al.*, *Nature* **518**, 529 (2015).
13. D. Gunning, "Explainable artificial intelligence (xai)." Defense Advanced Research Projects Agency, DARPA/I20 (DARPA, 2017).
14. D. H. Autor, F. Levy, R. J. Murnane (2003). *Q. J. Econ.* **118**, 1279 (2003).
15. D. H. Autor, D. Dorn, *Am. Econ. Rev.* **103**, 1553 (2013).
16. D. Remus, F. S. Levy, Can robots be lawyers?, *Georgetown J. Legal Ethics* (Summer 2017), p. 501.
17. G. Litjens *et al.*, A survey on deep learning in medical image analysis. arXiv preprint; arXiv:1702.05747 [cs.CV] (19 Feb 2017).
18. E. Brynjolfsson, A. McAfee, The business of artificial intelligence, *Harvard Business Rev.* (July 2017).
19. D. J. Deming, *Q. J. Econ.* **132**, 1593 (2017).
20. J. Manyika et al., *A Future that Works: Automation, Employment, and Productivity* (McKinsey Global Institute, 2017).
21. P. Milgrom, J. Roberts, *Am. Econ. Rev.* **86**, 173 (1996).
22. E. Brynjolfsson, P. Milgrom, in *The Handbook of Organizational Economics*, R. Gibbons, J. Roberts, Eds. (Princeton Univ. Press, 2013), pp. 11–55.
23. D. Acemoglu, P. Restrepo, NBER Working Paper 22252 (National Bureau of Economic Research, 2016).

ACKNOWLEDGMENTS

The authors acknowledge helpful discussions and comments from D. Acemoglu, D. Autor, S. Benzell, Y. LeCun, F. Levy, A. Ng, T. Poggio, D. Rus, G. Saint-Jacques, Y. Shavit, C. Syverson, S. Thrun, and especially D. Rock. E.B. acknowledges funding from the MIT Initiative on the Digital Economy for this research. T.M. acknowledges support from the U.S. Air Force Office of Scientific Research.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/358/6370/1530/suppl/DC1

10.1126/science.aap8062