NEXUS

EDUCATION MARKETPLACE LEADERSHIP VALUES

The Future of Artificial Intelligence

Andrew Boyarsky, MSM, PMP
Clinical Associate Professor, and Academic Director of the MS in Enterprise Risk Management, Katz School of Graduate and Professional Studies

**“Every major player is working on this technology of artificial intelligence.
As of now, it's benign... but I would say that the day is not far off when
artificial intelligence as applied to cyber warfare becomes a threat to everybody.”**

-Ted Bell, Bestselling novelist and Writer-in-Residence at Cambridge University

If you use technology today, a smartphone, a computer, or any connected electronic device, then you are bound to be using artificial intelligence as part of the embedded software of that technology. Here are just a few commonplace examples:

- Google search
- Voice recognition (Siri, Alexa, Cortana, again Google)
- Netflix or Amazon predictive analytics as to purchases
- Navigation apps like Waze (owned by Google), Google Maps, MapQuest (yes, remember this one?)
- Modern video games: Super Mario Bros., NBA 2K (my son's favorite), Call of Duty, etc.
- Fraud detection
- Translation software

Behind all of this technology is a complex set of algorithms, central processing units (CPUs), and computer servers with increasing levels of sophistication that are designed to accelerate that input and output of information, increasing the quality, complexity, and volume of interactions with our devices, and enhance our quality of life.

We are already living in the world of AI. Read the resources below in order to engage the past, present, and future of this world from a variety of perspectives.

So, what comes next? Where are we headed with AI, and what level of responsibility do the designers and providers have with managing AI technology? Will we control AI technology or will it control us? How do we handle the economic ramifications that are likely to be added to the existing stresses in our local, national, and global communities?

To come back to the title of this edition of Nexus, can we make AI, robots, etc. into the useful tools we intend them to be, with mensch-like attributes; or will it be a Golom that will rage beyond our control? I very much look forward to our community discussion on this topic.

Table of Contents

Curated Sources

The Basics: What is AI?

What's the Difference between Artificial Intelligence, Machine Learning, and Deep Learning? Michael Copeland, Nvidia Blog

Artificial intelligence is the future. Artificial intelligence is science fiction. Artificial intelligence is already part of our everyday lives. All those statements are true, it just depends on what flavor of AI you are referring to.

For example, when Google DeepMind's AlphaGo program defeated South Korean Master Lee Se-dol in the board game Go earlier this year, the terms AI, machine learning, and were used in the media to describe how DeepMind won. And all three are part of the reason why AlphaGo trounced Lee Se-Dol. But they are not the same things.

The easiest way to think of their relationship is to visualize them as concentric circles with AI — the idea that came first — the largest, then machine learning — which blossomed later, and finally deep learning — which is driving today's AI explosion — fitting inside both.

From Bust to Boom

AI has been part of our imaginations and simmering in research labs since a handful of computer scientists rallied around the term at the Dartmouth Conferences in 1956 and birthed the field of AI. In the decades since, AI has alternately been heralded as the key to our civilization's brightest future, and tossed on technology's trash heap as a harebrained notion of over-reaching propeller heads. Frankly, until 2012, it was a bit of both.

Over the past few years AI has exploded, and especially since 2015. Much of that has to do with the wide availability of GPUs that make parallel processing ever faster, cheaper, and more powerful. It also has to do with the simultaneous one-two punch of practically infinite storage and a flood of data of every stripe (that whole Big Data movement) – images, text, transactions, mapping data, you name it.

Let's walk through how computer scientists have moved from something of a bust — until 2012 — to a boom that has unleashed applications used by hundreds of millions of people every day.

Artificial Intelligence — Human Intelligence Exhibited by Machines

King me: computer programs that played checkers were among the earliest examples of artificial intelligence, stirring an early wave of excitement in the 1950s.

Back in that summer of '56 conference the dream of those AI pioneers was to construct complex machines — enabled by emerging computers — that possessed the same characteristics of human intelligence. This is the concept we think of as "General AI" — fabulous machines that have all our senses (maybe even more), all our reason, and think just like we do. You've seen these machines endlessly in movies as friend — C-3PO — and foe — The Terminator. General AI machines have remained in the movies and science fiction novels for good reason; we can't pull it off, at least not yet.

What we can do falls into the concept of “Narrow AI.” Technologies that are able to perform specific tasks as well as, or better than, we humans can. Examples of narrow AI are things such as image classification on a service like Pinterest and face recognition on Facebook.

Those are examples of Narrow AI in practice. These technologies exhibit some facets of human intelligence. But how? Where does that intelligence come from? That get us to the next circle, Machine Learning.

Machine Learning — An Approach to Achieve Artificial Intelligence

Spam free diet: machine learning helps keep your inbox (relatively) free of spam.

Machine learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world. So rather than hand-coding software routines with a specific set of instructions to accomplish a particular task, the machine is “trained” using large amounts of data and algorithms that give it the ability to learn how to perform the task.

Machine learning came directly from minds of the early AI crowd, and the algorithmic approaches over the years included decision tree learning, inductive logic programming, clustering, reinforcement learning, and Bayesian networks among others. As we know, none achieved the ultimate goal of General AI, and even Narrow AI was mostly out of reach with early machine learning approaches.

As it turned out, one of the very best application areas for machine learning for many years was computer vision, though it still required a great deal of hand-coding to get the job done. People would go in and write hand-coded classifiers like edge detection filters so the program could identify where an object started and stopped; shape detection to determine if it had eight sides; a classifier to recognize the letters “S-T-O-P.” From all those hand-coded classifiers they would develop algorithms to make sense of the image and “learn” to determine whether it was a stop sign.

Good, but not mind-bendingly great. Especially on a foggy day when the sign isn’t perfectly visible, or a tree obscures part of it. There’s a reason computer vision and image detection didn’t come close to rivaling humans until very recently, it was too brittle and too prone to error.

Time, and the right learning algorithms made all the difference.

Deep Learning — A Technique for Implementing Machine Learning

Herding cats: Picking images of cats out of YouTube videos was one of the first breakthrough demonstrations of deep learning.

Another algorithmic approach from the early machine-learning crowd, Artificial Neural Networks, came and mostly went over the decades. Neural Networks are inspired by our understanding of the biology of our brains – all those interconnections between the neurons. But, unlike a biological brain where any neuron can connect to any other neuron within a certain physical distance, these artificial neural networks have discrete layers, connections, and directions of data propagation.

You might, for example, take an image, chop it up into a bunch of tiles that are inputted into the first layer of the neural network. In the first layer individual neurons, then passes the data to a second layer. The second layer of neurons does its task, and so on, until the final layer and the final output is produced.

Each neuron assigns a weighting to its input — how correct or incorrect it is relative to the task being performed. The final output is then determined by the total of those weightings. So think of our stop sign example. Attributes of a stop sign image are chopped up and "examined" by the neurons — its octagonal shape, its fire-engine red color, its distinctive letters, its traffic-sign size, and its motion or lack thereof. The neural network's task is to conclude whether this is a stop sign or not. It comes up with a "probability vector," really a highly educated guess, based on the weighting. In our example the system might be 86% confident the image is a stop sign, 7% confident it's a speed limit sign, and 5% it's a kite stuck in a tree, and so on — and the network architecture then tells the neural network whether it is right or not.

Even this example is getting ahead of itself, because until recently neural networks were all but shunned by the AI research community. They had been around since the earliest days of AI, and had produced very little in the way of "intelligence." The problem was even the most basic neural networks were very computationally intensive, it just wasn't a practical approach. Still, a small heretical research group led by Geoffrey Hinton at the University of Toronto kept at it, finally parallelizing the algorithms for supercomputers to run and proving the concept, but it wasn't until GPUs were deployed in the effort that the promise was realized.

If we go back again to our stop sign example, chances are very good that as the network is getting tuned or "trained" it's coming up with wrong answers — a lot. What it needs is training. It needs to see hundreds of thousands, even millions of images, until the weightings of the neuron inputs are tuned so precisely that it gets the answer right practically every time — fog or no fog, sun or rain. It's at that point that the neural network has taught itself what a stop sign looks like; or your mother's face in the case of Facebook; or a cat, which is what Andrew Ng did in 2012 at Google.

Ng's breakthrough was to take these neural networks, and essentially make them huge, increase the layers and the neurons, and then run massive amounts of data through the system to train it. In Ng's case it was images from 10 million YouTube videos. Ng put the "deep" in deep learning, which describes all the layers in these neural networks.

Today, image recognition by machines trained via deep learning in some scenarios is better than humans, and that ranges from cats to identifying indicators for cancer in blood and tumors in MRI scans. Google's AlphaGo learned the game, and trained for its Go match — it tuned its neural network — by playing against itself over and over and over.

Thanks to Deep Learning, AI Has a Bright Future

Deep Learning has enabled many practical applications of Machine Learning and by extension the overall field of AI. Deep Learning breaks down tasks in ways that makes all kinds of machine assists seem possible, even likely. Driverless cars, better preventive healthcare, even better movie recommendations, are all here today or on the horizon. AI is the present and the future. With Deep Learning's help, AI may even get to that science fiction state we've so long imagined. You have a C-3PO, I'll take it. You can keep your Terminator.

Next Steps: How is AI Impacting Society?

Rabbinic Artificial Intelligence? Babylonian Talmud; Sanhedrin 65b

אמר רבא אי בעו צדיקי ברו עלמא שנאמר כי עונותיכם היו מבדילים וגו'

Rava says: If the righteous wish to do so, they can create a world, as it is stated: "But your iniquities have separated between you and your God." In other words, there is no distinction between God and a righteous person who has no sins, and just as God created the world, so can the righteous.

רבא ברא גברא שדריה לקמיה דר' זירא הוה קא משתעי בהדיה ולא הוה קא מהדר ליה אמר ליה מן חבריא את הדר לעפריך

Indeed, Rava created a man, a golem, using forces of sanctity. Rava sent his creation before Rabbi Zeira. Rabbi Zeira would speak to him but he would not reply. Rabbi Zeira said to him: You were created by one of the members of the group, one of the Sages. Return to your dust.

How Judaism Predicted the First Humanoid Robot, Mark Goldfeder, CNN

A column for CNN written by YC and RIETS graduate Professor Mark Goldfeder which uses the Babylonian Talmud in Sanhedrin 65b to examine how Jewish law determines the definition of being human.

The Robot Revolution Will Be the Quietest One, Liu Cixin, The New York Times

Turning Point**:** *Though the first fatal crash involving an autonomous car took place in July 2016, self-driving vehicles have been adopted around the world.*

In 2016, self-driving cars made inroads in several countries, many of which rewrote their laws to accommodate the new technology. As a science-fiction writer, it's my duty to warn the human race that the robot revolution has begun — even if no one has noticed yet.

When a few autonomous test cars appeared on the roads over the last few years, we didn't think of them as robots because they didn't have the humanoid shape that science-fiction movies taught us to expect. In 2016, they were adopted widely: as buses in the United Arab Emirates and the Netherlands, taxis in Singapore and private cars in the United States and China. There was a fatal accident in Florida involving an autonomous car, which caused some concerns, but this did not significantly affect our embrace of this technology.

Instead of arming ourselves against this alien presence, as some of my fellow science-fiction writers have fearfully suggested, we gawked as the vehicles pulled up to the curb. The driverless vehicles, some of which had no steering wheels or gas pedals, merged into traffic and stopped at stop signs, smoothly taking us to our destinations. We lounged in comfort, occasionally taking selfies.

Machine learning has been an important tool for autonomous car companies as they develop the systems that pilot their vehicles. Instead of rigidly following programming as an app on your phone does, an A.I. system can try to learn to do a

task itself, using techniques borrowed from human learning, like pattern recognition and trial and error, and may use hardware modeled on the architecture of a human brain. Currently, the responsibilities of artificial intelligence are mostly limited to tasks like translating texts, helping with medical diagnoses and writing simple articles for media companies. But we can expect to see unimaginable progress in this field in future — and the widespread use of the autonomous car is going to accelerate that process as automobile and technology companies invest ever more resources in its development.

Let's try to envision that future. As during every other technological revolution, the robots will first transform our economy. People who drive for a living will lose their jobs — around 3 million in the United States alone. E-commerce may experience further booms because of automation, and car ownership is likely to become nearly obsolete as more targeted car sharing and public transportation systems are developed. Eventually, the robot cars could be integrated with other transportation systems. Say that you live in New York City and want to go to China's Henan Province: You will enter the address into an app, a car will take you to your plane at the airport, and after you land, another will take you directly to your destination.

Robots will begin to creep into other areas of our lives — serving as busboys or waiters, for example — as our investments in robotic transport improve their prowess in areas such as environmental detection and modeling, hyper-complex problem solving and fuzzy-logic applications. With every advance, the use of A.I.-powered robots will expand into other fields: health care, policing, national defense and education.

There will be scandals when things go wrong and backlash movements from the new Luddites. But I don't think we'll protest very much. The A.I. systems that drive our cars will teach us to trust machine intelligence over the human variety — car accidents will become very rare, for example — and when given an opportunity to delegate a job to a robot, we will placidly do so without giving it much thought.

In all previous technological revolutions, people who lost their jobs mostly moved to new ones, but that will be less likely when the robots take over. A.I. that can learn from experience will replace many accountants, lawyers, bankers, insurance adjusters, doctors, scientific researchers and some creative professionals. Intelligence and advanced training will no longer mean job stability.

Gradually the A.I. era will transform the essence of human culture. When we're no longer more intelligent than our machines, when they can easily outthink and outperform us, making the sort of intuitive leaps in research and other areas that we currently associate with genius, a sort of learned helplessness is likely to set in for us, and the idea of work itself may cease to hold meaning.

As A.I. takes over, the remaining jobs may dwindle to a fraction of what they were, employing perhaps 10 percent or even less of the total population. These may be highly creative or complex jobs that robots can't do, such as senior management, directing scientific research or nursing and child care.

In the dystopian scenario, as jobless numbers rise across the globe, our societies sink into prolonged turmoil. The world could be engulfed by endless conflicts between those who control the A.I. and the rest of us. The technocratic 10 percent could end up living in a gated community with armed robot guards.

There is a second, utopian scenario, where we've anticipated these changes and come up with solutions beforehand. Those in political power have planned a smoother, gentler transition, perhaps using A.I. to help them anticipate and modulate the strife. At the end of it, almost all of us live on social welfare

How we will spend our time is hard to predict. "He who does not work, neither shall he eat" has been the cornerstone of civilizations through the ages, but that will have vanished. History shows that those who haven't had to work — aristocrats, say — have often spent their time entertaining and developing their artistic and sporting talents while scrupulously observing elaborate rituals of dress and manners.

In this future, creativity is highly valued. We sport ever more fantastic makeup, hairstyles and clothing. The labor of past ages seems barbaric.

But the aristocrats ruled nations; in the A.I. era, machines are doing all the thinking. Because, over the decades, we've gradually given up our autonomy, step by step, allowing ourselves to be transformed into A.I.'s docile, fabulously pampered pets. As A.I. whisks us from place to place — visits to family members, art galleries and musical events — we will look out the windows, as unaware of its plans for us as a poodle on its way to the groomer's.

This isn't crying wolf: Machines will take white-collar jobs during the next administration, Martin Ford, Linkedin

In this series, professionals provide advice for the next U.S. president. What do you want POTUS focused on?

Dear Madam / Mr. President:

Over fifty years ago, in March 1964, a document known as the "Triple Revolution Report" landed on the desk of your predecessor, Lyndon Johnson. That report, written by a prominent group of intellectuals that included two Nobel laureates, argued that the United States was on the brink of dramatic social and economic disruption as rapidly advancing industrial automation technology was poised to throw millions out of work.

Needless to say, that dire prediction did not come to pass. However, there are good reasons to believe that technology has finally advanced to the point where such concerns need to be taken seriously. The fear that machines might displace workers and create unemployment has a long history, and because the alarm has been prematurely sounded so many times in the past, there is a real danger that a "little boy who cried wolf" effect will leave us complacent and unprepared if and when the disruption finally arrives.

Recent advances in artificial intelligence and robotics suggest that it is entirely possible that a significant impact on the job market could begin to unfold during the course of your presidency. The most important thing to understand about all this progress is that computers no longer have to be programmed step-by-step. Machine learning—a technology that involves smart algorithms churning through vast amounts of data—in effect allows computers figure out for themselves how to perform tasks or reach specific goals.

The recent triumph of Google's DeepMind technology at learning to play the ancient game of "Go" and then triumphing against one of the world's best players was an especially vivid demonstration of the technology, but, in fact, machine learning is already in widespread use across both industries and occupations. Smart algorithms have already displaced

lawyers and paralegals who once reviewed documents as part of the legal discovery process. An increasing number of news articles published my major U.S. media companies are being generated autonomously by systems that analyze data and create content that is often indistinguishable from a story written by a human journalist. Machine learning is also powering the latest generation of robots, and the machines are rapidly becoming more flexible and dexterous.

As technology continues to accelerate, the number and types of jobs that can be automated is certain to expand dramatically. It's not just factory workers that can be replaced by robots and machines: Rapidly improving software automation and specialized artificial intelligence applications will make knowledge worker and professional occupations requiring college educations and advanced skills increasingly vulnerable. This demonstrated capability for information technology to climb the skills ladder and threaten the jobs taken by college graduates is a special cause for concern because it calls into question the only conventional solution we have to offer workers displaced by automation: ever more training and education.

If technology eventually results in wide-spread unemployment, or if it drives down wages for the majority of workers as jobs are deskilled and commoditized, then we could also run into a serious problem with consumer demand. Jobs are the primary mechanism that gets purchasing power into the hands of consumers so that they buy the products and services generated by the economy. If automation has a negative impact on consumer demand and confidence, then we run the risk of economic stagnation or even a downward, deflationary spiral.

While these concerns may seem either far-fetched science fiction or a return to the Ludditism we've experienced in the past, many of us in the technology community believe the risk is real--and that it deserves serious consideration. At a time when our political system is intensely polarized and seems unable to respond to even the most mundane challenges, the prospect of a dramatic and unanticipated economic and social disruption is not sometime we can afford to take lightly.

If the automation of jobs proves to be a relentless trend, then there will eventually be no alternative but to consider unconventional solutions--perhaps including a guaranteed basic income for all Americans. Needless to say, the implementation of such policies would present a staggering political challenge. Given that there is no reliable way to predict when the disruption will occur, or how fast it will unfold, it is imperative that planning begin well in advance. A logical first step would be to initiate some experimental pilot programs designed to test various policy responses. The data generated by these programs would be invaluable in eventually crafting an effective national policy to adapt our economy and society to the implications of disruptive technology.

I urge you to consider including among those who staff your new administration experts who are familiar with recent advances in artificial intelligence and robotics and with the potential economic and social impact of these technologies, and who are prepared to initiate the planning process.

The Hype—and Hope—of Artificial Intelligence, Om Malik, The New Yorker

Earlier this month, on his HBO show “Last Week Tonight,” John Oliver skewered media companies’ desperate search for clicks. Like many of his bits, it became a viral phenomenon, clocking in at nearly six million views on YouTube. At around the ten-minute mark, Oliver took his verbal bat to the knees of Tronc, the new name for Tribune Publishing Company, and its parody-worthy promotional video, in which a robotic spokeswoman describes the journalistic benefits of artificial intelligence, as a string section swells underneath.

Tronc is not the only company to enthusiastically embrace the term “artificial intelligence.” A.I. is hot, and every company worth its stock price is talking about how this magical potion will change everything. Even Macy’s recently announced that it was testing an I.B.M. artificial-intelligence tool in ten of its department stores, in order to bring back customers who are abandoning traditional retail in favor of online shopping.

Much like “the cloud,” “big data,” and “machine learning” before it, the term “artificial intelligence” has been hijacked by marketers and advertising copywriters. A lot of what people are calling “artificial intelligence” is really data analytics—in other words, business as usual. If the hype leaves you asking “What is A.I., really?,” don’t worry, you’re not alone. I asked various experts to define the term and got different answers. The only thing they all seem to agree on is that artificial intelligence is a set of technologies that try to imitate or augment human intelligence. To me, the emphasis is on augmentation, in which intelligent software helps us interact and deal with the increasingly digital world we live in.

Three decades ago, I read newspapers, wrote on an electric typewriter, and watched a handful of television channels. Today, I have streaming video from Netflix, Amazon, HBO, and other places, and I’m sometimes paralyzed by the choices. It is becoming harder for us to stay on top of the onslaught—e-mails, messages, appointments, alerts. Augmented intelligence offers the possibility of winnowing an increasing number of inputs and options in a way that humans can’t manage without a helping hand.

Computers in general, and software in particular, are much more difficult than other kinds of technology for most people to grok, and they overwhelm us with a sense of mystery. There was a time when you would record a letter or a document on a dictaphone and someone would transcribe it for you. A human was making the voice-to-text conversion with the help of a machine. Today, you can speak into your iPhone and it will transcribe your messages itself. If people could have seen our current voice-to-text capabilities fifty years ago, it would have looked as if technology had become sentient. Now it’s just a routine way to augment how we interact with the world. Kevin Kelly, the writer and futurist, whose most recent book is “The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future,” said, “What we can do now would be A.I. fifty years ago. What we can do in fifty years will not be called A.I.”

You don’t have to look up from Facebook to get his point. Before we had the Internet, we would either call or write to our friends, one at a time, and keep up with their lives. It was a slow process, and took a lot of effort and time to learn about each other. As a result, we had fewer interactions—there was a cost attached to making long-distance phone calls and a time commitment attached to writing letters. With the advent of the Internet, e-mail emerged as a way to facilitate and speed up those interactions. Facebook did one better—it turned your address book into a hub, allowing you to simultaneously stay in touch with hundreds, even thousands, of friends. The algorithm allows us to maintain more relationships with much less effort at almost no cost.

Michelle Zhou spent over a decade and a half at I.B.M. Research and I.B.M. Watson Group before leaving to become a co-founder of Juji, a sentiment-analysis startup. An expert in a field where artificial intelligence and human-computer interaction intersect, Zhou breaks down A.I. into three stages. The first is recognition intelligence, in which algorithms running on ever more powerful computers can recognize patterns and glean topics from blocks of text, or perhaps even derive the meaning of a whole document from a few sentences. The second stage is cognitive intelligence, in which machines can go beyond pattern recognition and start making inferences from data. The third stage will be reached only when we can create virtual human beings, who can think, act, and behave as humans do.

We are a long way from creating virtual human beings. Despite what you read in the media, no technology is perfect, and the most valuable function of A.I. lies in augmenting human intelligence. To even reach that point, we need to train computers to mimic humans. An April, 2016, story in *Bloomberg Business* provided a good example. It described how companies that provide automated A.I. personal assistants (of the sort that arrange schedules or help with online shopping) had hired human "trainers" to check and evaluate the A.I. assistants' responses before they were sent out. "It's ironic that we define artificial intelligence with respect to its ability to replicate human intelligence," said Sean Gourley, the founder of Primer, a data-analytics company, and an expert on deriving intelligence from large data sets with the help of algorithms.

Whether it is Spotify or Netflix or a new generation of A.I. chat bots, all of these tools rely on humans themselves to provide the data. When we listen to songs, put them on playlists, and share them with others, we are sending vital signals to Spotify that train its algorithms not only to discover what we might like but also to predict hits.

Even the much talked-about "computer vision" has become effective only because humans have uploaded billions of photos and tagged them with metadata to give those photos context. Increasingly powerful computers can scan through these photos and find patterns and meaning. Similarly, Google can use billions of voice samples it has collected over the years to build a smart system that understands accents and nuances, which make its voice-based search function possible.

Using Zhou's three stages as a yardstick, we are only in the "recognition intelligence" phase—today's computers use deep learning to discover patterns faster and better. It's true, however, that some companies are working on technologies that can be used for inferring meanings, which would be the next step. "It does not matter whether we will end up at stage 3," Zhou wrote to me in an e-mail. "I'm still a big fan of man-machine symbiosis, where computers do the best they can (that is being consistent, objective, precise), and humans do our best (creative, imprecise but adaptive)." For a few more decades, at least, humans will continue to train computers to mimic us. And, in the meantime, we're going to have to deal with the hyperbole surrounding A.I.

Deep Dive: What are the Next Frontiers and Further Implications of AI?

DeepMind and Blizzard Open StarCraft II as an AI Research Environment, Oriyal VInyals, Stephen Gaffney, Timo Ewalds; DeepMind

DeepMind's scientific mission is to push the boundaries of AI by developing systems that can learn to solve complex problems. To do this, we design agents and test their ability in a wide range of environments from the purpose-built DeepMind Lab to established games, such as Atari and Go.

Testing our agents in games that are not specifically designed for AI research, and where humans play well, is crucial to benchmark agent performance. That is why we, along with our partner Blizzard Entertainment, are excited to announce the release of SC2LE, a set of tools that we hope will accelerate AI research in the real-time strategy game StarCraft II. The SC2LE release includes:

A Machine Learning API developed by Blizzard that gives researchers and developers hooks into the game. This includes the release of tools for Linux for the first time.

A dataset of anonymised game replays, which will increase from 65k to more than half a million in the coming weeks.

An open source version of DeepMind's toolset, PySC2, to allow researchers to easily use Blizzard's feature-layer API with their agents.

A series of simple RL mini-games to allow researchers to test the performance of agents on specific tasks.

A joint paper that outlines the environment, and reports initial baseline results on the mini-games, supervised learning from replays, and the full 1v1 ladder game against the built-in AI.

StarCraft and StarCraft II are among the biggest and most successful games of all time, with players competing in tournaments for more than 20 years. The original game is also already used by AI and ML researchers, who compete annually in the AIIDE bot competition. Part of StarCraft's longevity is down to the rich, multi-layered gameplay, which also makes it an ideal environment for AI research.

For example, while the objective of the game is to beat the opponent, the player must also carry out and balance a number of sub-goals, such as gathering resources or building structures. In addition, a game can take from a few minutes to one hour to complete, meaning actions taken early in the game may not pay-off for a long time. Finally, the map is only partially observed, meaning agents must use a combination of memory and planning to succeed.

The game also has other qualities that appeal to researchers, such as the large pool of avid players that compete online every day. This ensures that there is a large quantity of replay data to learn from - as well as a large quantity of extremely talented opponents for AI agents.

Even StarCraft's action space presents a challenge with a choice of more than 300 basic actions that can be taken. Contrast this with Atari games, which only have about 10 (e.g. up, down, left, right etc). On top of this, actions in StarCraft are hierarchical, can be modified and augmented, with many of them requiring a point on the screen. Even assuming a small screen size of 84x84 there are roughly 100 million possible actions available.

This release means researchers can now tackle some of these challenges using Blizzard's own tools to build their own tasks and models.

Our PySC2 environment wrapper helps by offering a flexible and easy-to-use interface for RL agents to play the game. In this initial release, we break the game down into "feature layers", where elements of the game such as unit type, health and map visibility are isolated from each other, whilst preserving the core visual and spatial elements of the game.

The release also contains a series of 'mini-games' - an established technique for breaking down the game into manageable chunks that can be used to test agents on specific tasks, such as moving the camera, collecting mineral shards or selecting units. We hope that researchers can test their techniques on these as well as propose new mini-games for other researchers to compete and evaluate on.

Our initial investigations show that our agents perform well on these mini-games. But when it comes to the full game, even strong baseline agents, such as A3C, cannot win a single game against even the easiest built-in AI. For instance, the following video shows an early-stage training agent (left) which fails to keep its workers mining, a task that humans find trivial. After training (right), the agents perform more meaningful actions, but if they are to be competitive, we will need further breakthroughs in deep RL and related areas.

One technique that we know allows our agents to learn stronger policies is imitation learning. This kind of training will soon be far easier thanks to Blizzard, which has committed to ongoing releases of hundreds of thousands of anonymized replays gathered from the StarCraft II ladder. These will not only allow researchers to train supervised agents to play the game, but also opens up other interesting areas of research such as sequence prediction and long-term memory.

Our hope is that the release of these new tools will build on the work that the AI community has already done in StarCraft, encouraging more DeepRL research and making it easier for researchers to focus on the frontiers of our field.

We look forward to seeing what the community discovers.

Neuroscience-Inspired Artificial Intelligence, Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, Matthew Botvinick; Neuron

The fields of neuroscience and artificial intelligence (AI) have a long and intertwined history. In more recent times, however, communication and collaboration between the two fields has become less commonplace. In this article, we argue that better understanding biological brains could play a vital role in building intelligent machines. We survey historical interactions between the AI and neuroscience fields and emphasize current advances in AI that have been inspired by the study of neural computation in humans and other animals. We conclude by highlighting shared themes that may be key for advancing future research in both fields.

In recent years, rapid progress has been made in the related fields of neuroscience and artificial intelligence (AI). At the dawn of the computer age, work on AI was inextricably intertwined with neuroscience and psychology, and many of the early pioneers straddled both fields, with collaborations between these disciplines proving highly productive

(Churchland and Sejnowski, 1988, Hebb, 1949, Hinton et al., 1986, Hopfield, 1982, McCulloch and Pitts, 1943, Turing, 1950). However, more recently, the interaction has become much less commonplace, as both subjects have grown enormously in complexity and disciplinary boundaries have solidified. In this review, we argue for the critical and ongoing importance of neuroscience in generating ideas that will accelerate and guide AI research (see Hassabis commentary in Brooks et al., 2012).

We begin with the premise that building human-level general AI (or "Turing-powerful" intelligent systems; Turing, 1936) is a daunting task, because the search space of possible solutions is vast and likely only very sparsely populated. We argue that this therefore underscores the utility of scrutinizing the inner workings of the human brain— the only existing proof that such an intelligence is even possible. Studying animal cognition and its neural implementation also has a vital role to play, as it can provide a window into various important aspects of higher-level general intelligence.

The benefits to developing AI of closely examining biological intelligence are two-fold. First, neuroscience provides a rich source of inspiration for new types of algorithms and architectures, independent of and complementary to the mathematical and logic-based methods and ideas that have largely dominated traditional approaches to AI. For example, were a new facet of biological computation found to be critical to supporting a cognitive function, then we would consider it an excellent candidate for incorporation into artificial systems. Second, neuroscience can provide validation of AI techniques that already exist. If a known algorithm is subsequently found to be implemented in the brain, then that is strong support for its plausibility as an integral component of an overall general intelligence system. Such clues can be critical to a long-term research program when determining where to allocate resources most productively. For example, if an algorithm is not quite attaining the level of performance required or expected, but we observe it is core to the functioning of the brain, then we can surmise that redoubled engineering efforts geared to making it work in artificial systems are likely to pay off.

Of course from a practical standpoint of building an AI system, we need not slavishly enforce adherence to biological plausibility. From an engineering perspective, what works is ultimately all that matters. For our purposes then, biological plausibility is a guide, not a strict requirement. What we are interested in is a systems neuroscience-level understanding of the brain, namely the algorithms, architectures, functions, and representations it utilizes. This roughly corresponds to the top two levels of the three levels of analysis that Marr famously stated are required to understand any complex biological system (Marr and Poggio, 1976): the goals of the system (the computational level) and the process and computations that realize this goal (the algorithmic level). The precise mechanisms by which this is physically realized in a biological substrate are less relevant here (the implementation level). Note this is where our approach to neuroscience-inspired AI differs from other initiatives, such as the Blue Brain Project (Markram, 2006) or the field of neuromorphic computing systems (Esser et al., 2016), which attempt to closely mimic or directly reverse engineer the specifics of neural circuits (albeit with different goals in mind). By focusing on the computational and algorithmic levels, we gain transferrable insights into general mechanisms of brain function, while leaving room to accommodate the distinctive opportunities and challenges that arise when building intelligent machines in silico.

The following sections unpack these points by considering the past, present, and future of the AI-neuroscience interface. Before beginning, we offer a clarification. Throughout this article, we employ the terms "neuroscience" and "AI." We use these terms in the widest possible sense. When we say neuroscience, we mean to include all fields that are involved

with the study of the brain, the behaviors that it generates, and the mechanisms by which it does so, including cognitive neuroscience, systems neuroscience and psychology. When we say AI, we mean work in machine learning, statistics, and

AI research that aims to build intelligent machines (Legg and Hutter, 2007).

We begin by considering the origins of two fields that are pivotal for current AI research, deep learning and reinforcement learning, both of which took root in ideas from neuroscience. We then turn to the current state of play in AI research, noting many cases where inspiration has been drawn (sometimes without explicit acknowledgment) from concepts and findings in neuroscience. In this section, we particularly emphasize instances where we have combined deep learning with other approaches from across machine learning, such as reinforcement learning (Mnih et al., 2015), Monte Carlo tree search (Silver et al., 2016), or techniques involving an external content-addressable memory (Graves et al., 2016). Next, we consider the potential for neuroscience to support future AI research, looking at both the most likely research challenges and some emerging neuroscience-inspired AI techniques. While our main focus will be on the potential for neuroscience to benefit AI, our final section will briefly consider ways in which AI may be helpful to neuroscience and the broader potential for synergistic interactions between these two fields.

The Past:

Deep Learning

As detailed in a number of recent reviews, AI has been revolutionized over the past few years by dramatic advances in neural network, or "deep learning," methods (LeCun et al., 2015, Schmidhuber, 2014). As the moniker "neural network" might suggest, the origins of these AI methods lie directly in neuroscience. In the 1940s, investigations of neural computation began with the construction of artificial neural networks that could compute logical functions (McCulloch and Pitts, 1943). Not long after, others proposed mechanisms by which networks of neurons might learn incrementally via supervisory feedback (Rosenblatt, 1958) or efficiently encode environmental statistics in an unsupervised fashion (Hebb, 1949). These mechanisms opened up the field of artificial neural network research, and they continue to provide the foundation for contemporary research on deep learning (Schmidhuber, 2014).

Not long after this pioneering work, the development of the backpropagation algorithm allowed learning to occur in networks composed of multiple layers (Rumelhart et al., 1985, Werbos, 1974). Notably, the implications of this method for understanding intelligence, including AI, were first appreciated by a group of neuroscientists and cognitive scientists, working under the banner of parallel distributed processing (PDP) (Rumelhart et al., 1986). At the time, most AI research was focused on building logical processing systems based on serial computation, an approach inspired in part by the notion that human intelligence involves manipulation of symbolic representations (Haugeland, 1985). However, there was a growing sense in some quarters that purely symbolic approaches might be too brittle and inflexible to solve complex real-world problems of the kind that humans routinely handle. Instead, a growing foundation of knowledge about the brain seemed to point in a very different direction, highlighting the role of stochastic and highly parallelized information processing. Building on this, the PDP movement proposed that human cognition and behavior emerge from dynamic, distributed interactions within networks of simple neuron-like processing units, interactions tuned by learning procedures that adjust system parameters in order to minimize error or maximize reward.

Although the PDP approach was at first applied to relatively small-scale problems, it showed striking success in accounting for a wide range of human behaviors (Hinton et al., 1986). Along the way, PDP research introduced a diverse collection of ideas that have had a sustained influence on AI research. For example, current machine translation research exploits the notion that words and sentences can be represented in a distributed fashion (i.e., as vectors) (LeCun et al., 2015), a principle that was already ingrained in early PDP-inspired models of sentence processing (St. John and McClelland, 1990). Building on the PDP movement's appeal to biological computation, current state-of-the-art convolutional neural networks (CNNs) incorporate several canonical hallmarks of neural computation, including nonlinear transduction, divisive normalization, and maximum-based pooling of inputs (Yamins and DiCarlo, 2016). These operations were directly inspired by single-cell recordings from the mammalian visual cortex that revealed how visual input is filtered and pooled in simple and complex cells in area V1 (Hubel and Wiesel, 1959). Moreover, current network architectures replicate the hierarchical organization of mammalian cortical systems, with both convergent and divergent information flow in successive, nested processing layers (Krizhevsky et al., 2012, LeCun et al., 1989, Riesenhuber and Poggio, 1999, Serre et al., 2007), following ideas first advanced in early neural network models of visual processing (Fukushima, 1980). In both biological and artificial systems, successive non-linear computations transform raw visual input into an increasingly complex set of features, permitting object recognition that is invariant to transformations of pose, illumination, or scale.

As the field of deep learning evolved out of PDP research into a core area within AI, it was bolstered by new ideas, such as the development of deep belief networks (Hinton et al., 2006) and the introduction of large datasets inspired by research on human language (Deng et al., 2009). During this period, it continued to draw key ideas from neuroscience. For example, biological considerations informed the development of successful regularization schemes that support generalization beyond training data. One such scheme, in which only a subset of units participate in the processing of a given training example ("dropout"), was motivated by the stochasticity that is inherent in biological systems populated by neurons that fire with Poisson-like statistics (Hinton et al., 2012). Here and elsewhere, neuroscience has provided initial guidance toward architectural and algorithmic constraints that lead to successful neural network applications for AI.

Reinforcement Learning

Alongside its important role in the development of deep learning, neuroscience was also instrumental in erecting a second pillar of contemporary AI, stimulating the emergence of the field of reinforcement learning (RL). RL methods address the problem of how to maximize future reward by mapping states in the environment to actions and are among the most widely used tools in AI research (Sutton and Barto, 1998). Although it is not widely appreciated among AI researchers, RL methods were originally inspired by research into animal learning. In particular, the development of temporal-difference (TD) methods, a critical component of many RL models, was inextricably intertwined with research into animal behavior in conditioning experiments. TD methods are real-time models that learn from differences between temporally successive predictions, rather than having to wait until the actual reward is delivered. Of particular relevance was an effect called second-order conditioning, where affective significance is conferred on a conditioned stimulus (CS) through association with another CS rather than directly via association with the unconditioned stimulus (Sutton and Barto, 1981). TD learning provides a natural explanation for second-order conditioning and indeed has gone on to explain a much wider range of findings from neuroscience, as we discuss below.

Here, as in the case of deep learning, investigations initially inspired by observations from neuroscience led to further developments that have strongly shaped the direction of AI research. From their neuroscience-informed origins, TD methods and related techniques have gone on to supply the core technology for recent advances in AI, ranging from robotic control (Hafner and Riedmiller, 2011) to expert play in backgammon (Tesauro, 1995) and Go (Silver et al., 2016).

The Present:

Reading the contemporary AI literature, one gains the impression that the earlier engagement with neuroscience has diminished. However, if one scratches the surface, one can uncover many cases in which recent developments have been inspired and guided by neuroscientific considerations. Here, we look at four specific examples.

Attention

The brain does not learn by implementing a single, global optimization principle within a uniform and undifferentiated neural network (Marblestone et al., 2016). Rather, biological brains are modular, with distinct but interacting subsystems underpinning key functions such as memory, language, and cognitive control (Anderson et al., 2004, Shallice, 1988). This insight from neuroscience has been imported, often in an unspoken way, into many areas of current AI.

One illustrative example is recent AI work on attention. Up until quite lately, most CNN models worked directly on entire images or video frames, with equal priority given to all image pixels at the earliest stage of processing. The primate visual system works differently. Rather than processing all input in parallel, visual attention shifts strategically among locations and objects, centering processing resources and representational coordinates on a series of regions in turn (Koch and Ullman, 1985, Moore and Zirnsak, 2017, Posner and Petersen, 1990). Detailed neurocomputational models have shown how this piecemeal approach benefits behavior, by prioritizing and isolating the information that is relevant at any given moment (Olshausen et al., 1993, Salinas and Abbott, 1997). As such, attentional mechanisms have been a source of inspiration for AI architectures that take “glimpses” of the input image at each step, update internal state representations, and then select the next location to sample (Larochelle and Hinton, 2010, Mnih et al., 2014) (Figure 1A). One such network was able to use this selective attentional mechanism to ignore irrelevant objects in a scene, allowing it to perform well in challenging object classification tasks in the presence of clutter (Mnih et al., 2014). Further, the attentional mechanism allowed the computational cost (e.g., number of network parameters) to scale favorably with the size of the input image. Extensions of this approach were subsequently shown to produce impressive performance at difficult multi-object recognition tasks, outperforming conventional CNNs that process the entirety of the image, both in terms of accuracy and computational efficiency (Ba et al., 2015), as well as enhancing image-to-caption generation (Xu et al., 2015).

While attention is typically thought of as an orienting mechanism for perception, its “spotlight” can also be focused internally, toward the contents of memory. This idea, a recent focus in neuroscience studies (Summerfield et al., 2006), has also inspired work in AI. In some architectures, attentional mechanisms have been used to select information to be read out from the internal memory of the network. This has helped provide recent successes in machine translation (Bahdanau et al., 2014) and led to important advances on memory and reasoning tasks (Graves et al., 2016). These architectures offer a novel implementation of content-addressable retrieval, which was itself a concept originally introduced to AI from neuroscience (Hopfield, 1982).

One further area of AI where attention mechanisms have recently proven useful focuses on generative models, systems that learn to synthesize or "imagine" images (or other kinds of data) that mimic the structure of examples presented during training. Deep generative models (i.e., generative models implemented as multi-layered neural networks) have recently shown striking successes in producing synthetic outputs that capture the form and structure of real visual scenes via the incorporation of attention-like mechanisms (Hong et al., 2015, Reed et al., 2016). For example, in one state-of-the-art generative model known as DRAW, attention allows the system to build up an image incrementally, attending to one portion of a "mental canvas" at a time (Gregor et al., 2015).

Episodic Memory

A canonical theme in neuroscience is that that intelligent behavior relies on multiple memory systems (Tulving, 1985). These will include not only reinforcement-based mechanisms, which allow the value of stimuli and actions to be learned incrementally and through repeated experience, but also instance-based mechanisms, which allow experiences to be encoded rapidly (in "one shot") in a content-addressable store (Gallistel and King, 2009). The latter form of memory, known as episodic memory (Tulving, 2002), is most often associated with circuits in the medial temporal lobe, prominently including the hippocampus (Squire et al., 2004).

One recent breakthrough in AI has been the successful integration of RL with deep learning (Mnih et al., 2015, Silver et al., 2016). For example, the deep Q-network (DQN) exhibits expert play on Atari 2600 video games by learning to transform a vector of image pixels into a policy for selecting actions (e.g., joystick movements). One key ingredient in DQN is "experience replay," whereby the network stores a subset of the training data in an instance-based way, and then "replays" it offline, learning anew from successes or failures that occurred in the past. Experience replay is critical to maximizing data efficiency, avoids the destabilizing effects of learning from consecutive correlated experiences, and allows the network to learn a viable value function even in complex, highly structured sequential environments such as video games.

Critically, experience replay was directly inspired by theories that seek to understand how the multiple memory systems in the mammalian brain might interact. According to a prominent view, animal learning is supported by parallel or "complementary" learning systems in the hippocampus and neocortex (Kumaran et al., 2016, McClelland et al., 1995). The hippocampus acts to encode novel information after a single exposure (one-shot learning), but this information is gradually consolidated to the neocortex in sleep or resting periods that are interleaved with periods of activity. This consolidation is accompanied by replay in the hippocampus and neocortex, which is observed as a reinstatement of the structured patterns of neural activity that accompanied the learning event (O'Neill et al., 2010, Skaggs and McNaughton, 1996) (Figure 1B). This theory was originally proposed as a solution to the well-known problem that in conventional neural networks, correlated exposure to sequential task settings leads to mutual interference among policies, resulting in catastrophic forgetting of one task as a new one is learned. The replay buffer in DQN might thus be thought of as a very primitive hippocampus, permitting complementary learning in silico much as is proposed for biological brains. Later work showed that the benefits of experience replay in DQN are enhanced when replay of highly rewarding events is prioritized (Schaul et al., 2015), just as hippocampal replay seems to favor events that lead to high levels of reinforcement (Singer and Frank, 2009).

Experiences stored in a memory buffer can not only be used to gradually adjust the parameters of a deep network toward an optimal policy, as in DQN, but can also support rapid behavioral change based on an individual experience. Indeed, theoretical neuroscience has argued for the potential benefits of episodic control, whereby rewarded action sequences can be internally re-enacted from a rapidly updateable memory store, implemented in the biological case in the hippocampus (Gershman and Daw, 2017). Further, normative accounts show that episodic control is particularly advantageous over other learning mechanisms when limited experience of the environment has been obtained (Lengyel and Dayan, 2007).

Recent AI research has drawn on these ideas to overcome the slow learning characteristics of deep RL networks, developing architectures that implement episodic control (Blundell et al., 2016). These networks store specific experiences (e.g., actions and reward outcomes associated with particular Atari game screens) and select new actions based on the similarity between the current situation input and the previous events stored in memory, taking the reward associated with those previous events into account (Figure 1B). As predicted from the initial, neuroscience-based work (Lengyel and Dayan, 2007), artificial agents employing episodic control show striking gains in performance over deep RL networks, particularly early on during learning (Blundell et al., 2016). Further, they are able to achieve success on tasks that depend heavily on one-shot learning, where typical deep RL architectures fail. Moreover, episodic-like memory systems more generally have shown considerable promise in allowing new concepts to be learned rapidly based on only a few examples (Vinyals et al., 2016). In the future, it will be interesting to harness the benefits of rapid episodic-like memory and more traditional incremental learning in architectures that incorporate both of these components within an interacting framework that mirrors the complementary learning systems in mammalian brain. We discuss these future perspectives below in more detail later, in “Imagination and planning.”

Working Memory

Human intelligence is characterized by a remarkable ability to maintain and manipulate information within an active store, known as working memory, which is thought to be instantiated within the prefrontal cortex and interconnected areas (Goldman-Rakic, 1990). Classic cognitive theories suggest that this functionality depends on interactions between a central controller (“executive”) and separate, domain-specific memory buffers (e.g., visuo-spatial sketchpad) (Baddeley, 2012). AI research has drawn inspiration from these models, by building architectures that explicitly maintain information over time. Historically, such efforts began with the introduction of recurrent neural network architectures displaying attractor dynamics and rich sequential behavior, work directly inspired by neuroscience (Elman, 1990, Hopfield and Tank, 1986, Jordan, 1997). This work enabled later, more detailed modeling of human working memory (Botvinick and Plaut, 2006, Durstewitz et al., 2000), but it also laid the foundation for further technical innovations that have proved pivotal in recent AI research. In particular, one can see close parallels between the learning dynamics in these early, neuroscience-inspired networks and those in long-short-term memory (LSTM) networks, which subsequently achieved state of the art performance across a variety of domains. LTSMs allow information to be gated into a fixed activity state and maintained until an appropriate output is required (Hochreiter and Schmidhuber, 1997). Variants of this type of network have shown some striking behaviors in challenging domains, such as learning to respond to queries about the latent state of variables after training on computer code (Zaremba and Sutskever, 2014).

In ordinary LSTM networks, the functions of sequence control and memory storage are closely intertwined. This contrasts with classic models of human working memory, which, as mentioned above, separate these two. This neuroscience-based schema has recently inspired more complex AI architectures where control and storage are supported by distinct modules (Graves et al., 2014, Graves et al., 2016, Weston et al., 2014). For example, the differential neural computer (DNC) involves a neural network controller that attends to and reads/writes from an external memory matrix (Graves et al., 2016). This externalization allows the network controller to learn from scratch (i.e., via end-to-end optimization) to perform a wide range of complex memory and reasoning tasks that currently elude LSTMs, such as finding the shortest path through a graph-like structure, such as a subway map, or manipulating blocks in a variant of the Tower of Hanoi task (Figure 1C). These types of problems were previously argued to depend exclusively on symbol processing and variable binding and therefore beyond the purview of neural networks (Fodor and Pylyshyn, 1988, Marcus, 1998). Of note, although both LSTMs and the DNC are described here in the context of working memory, they have the potential to maintain information over many thousands of training cycles and so may thus be suited to longer-term forms of memory, such as retaining and understanding the contents of a book.

Continual Learning

Intelligent agents must be able to learn and remember many different tasks that are encountered over multiple timescales. Both biological and artificial agents must thus have a capacity for continual learning, that is, an ability to master new tasks without forgetting how to perform prior tasks (Thrun and Mitchell, 1995). While animals appear relatively adept at continual learning, neural networks suffer from the problem of catastrophic forgetting (French, 1999, McClelland et al., 1995). This occurs as the network parameters shift toward the optimal state for performing the second of two successive tasks, overwriting the configuration that allowed them to perform the first. Given the importance of continual learning, this liability of neural networks remains a significant challenge for the development of AI.

In neuroscience, advanced neuroimaging techniques (e.g., two-photon imaging) now allow dynamic in vivo visualization of the structure and function of dendritic spines during learning, at the spatial scale of single synapses (Nishiyama and Yasuda, 2015). This approach can be used to study neocortical plasticity during continual learning (Cichon and Gan, 2015, Hayashi-Takagi et al., 2015, Yang et al., 2009). There is emerging evidence for specialized mechanisms that protect knowledge about previous tasks from interference during learning on a new task. These include decreased synaptic lability (i.e., lower rates of plasticity) in a proportion of strengthened synapses, mediated by enlargements to dendritic spines that persist despite learning of other tasks (Cichon and Gan, 2015, Yang et al., 2009) (Figure 1D). These changes are associated with retention of task performance over several months, and indeed, if they are “erased” with synaptic optogenetics, this leads to forgetting of the task (Hayashi-Takagi et al., 2015). These empirical insights are consistent with theoretical models that suggest that memories can be protected from interference through synapses that transition between a cascade of states with different levels of plasticity (Fusi et al., 2005) (Figure 1D).

Together, these findings from neuroscience have inspired the development of AI algorithms that address the challenge of continual learning in deep networks by implementing of a form of “elastic” weight consolidation (EWC) (Kirkpatrick et al., 2017), which acts by slowing down learning in a subset of network weights identified as important to previous tasks, thereby anchoring these parameters to previously found solutions (Figure 1D). This allows multiple tasks to be

learned without an increase in network capacity, with weights shared efficiently between tasks with related structure. In this way, the EWC algorithm allows deep RL networks to support continual learning at large scale.

The Future:

In AI, the pace of recent research has been remarkable. Artificial systems now match human performance in challenging object recognition tasks (Krizhevsky et al., 2012) and outperform expert humans in dynamic, adversarial environments such as Atari video games (Mnih et al., 2015), the ancient board game of Go (Silver et al., 2016), and imperfect information games such as heads-up poker (Moravčík et al., 2017). Machines can autonomously generate synthetic natural images and simulations of human speech that are almost indistinguishable from their real-world counterparts (Lake et al., 2015, van den Oord et al., 2016), translate between multiple languages (Wu et al., 2016), and create "neural art" in the style of well-known painters (Gatys et al., 2015).

However, much work is still needed to bridge the gap between machine and human-level intelligence. In working toward closing this gap, we believe ideas from neuroscience will become increasingly indispensable. In neuroscience, the advent of new tools for brain imaging and genetic bioengineering have begun to offer a detailed characterization of the computations occurring in neural circuits, promising a revolution in our understanding of mammalian brain function (Deisseroth and Schnitzer, 2013). The relevance of neuroscience, both as a roadmap for the AI research agenda and as a source of computational tools is particularly salient in the following key areas.

Intuitive Understanding of the Physical World

Recent perspectives emphasize key ingredients of human intelligence that are already well developed in human infants but lacking in most AI systems (Gilmore et al., 2007, Gopnik and Schulz, 2004, Lake et al., 2016). Among these capabilities are knowledge of core concepts relating to the physical world, such as space, number, and objectness, which allow people to construct compositional mental models that can guide inference and prediction (Battaglia et al., 2013, Spelke and Kinzler, 2007).

AI research has begun to explore methods for addressing this challenge. For example, novel neural network architectures have been developed that interpret and reason about scenes in a humanlike way, by decomposing them into individual objects and their relations (Battaglia et al., 2016, Chang et al., 2016, Eslami et al., 2016) (Figures 2A and 2B). In some cases, this has resulted in human-level performance on challenging reasoning tasks (Santoro et al., 2017). In other work, deep RL has been used to capture the processes by which children gain commonsense understanding of the world through interactive experiments (Denil et al., 2016). Relatedly, deep generative models have been developed that are able to construct rich object models from raw sensory inputs (Higgins et al., 2016). These leverage constraints first identified in neuroscience, such as redundancy reduction (Barlow, 1959), which encourage the emergence of disentangled representations of independent factors such as shape and position (Figure 2C). Importantly, the latent representations learned by such generative models exhibit compositional properties, supporting flexible transfer to novel tasks (Eslami et al., 2016, Higgins et al., 2016, Rezende et al., 2016a). In the caption associated with Figure 2, we provide more detailed information about these networks.

Efficient Learning

Human cognition is distinguished by its ability to rapidly learn about new concepts from only a handful of examples, leveraging prior knowledge to enable flexible inductive inferences. In order to highlight this human ability as a challenge for AI, Lake and colleagues recently posed a "characters challenge" (Lake et al., 2016). Here, an observer must distinguish novel instances of an unfamiliar handwritten character from other, similar items after viewing only a single exemplar. Humans can perform this task well, but it is difficult for classical AI systems.

Encouragingly, recent AI algorithms have begun to make progress on tasks like the characters challenge, through both structured probabilistic models (Lake et al., 2015) and deep generative models based on the abovementioned DRAW model (Rezende et al., 2016b). Both classes of system can make inferences about a new concept despite a poverty of data and generate new samples from a single example concept (Figure 2D). Further, recent AI research has developed networks that "learn to learn," acquiring knowledge on new tasks by leveraging prior experience with related problems, to support one-shot concept learning (Santoro et al., 2016, Vinyals et al., 2016) and accelerating learning in RL tasks (Wang et al., 2016). Once again, this builds on concepts from neuroscience: learning to learn was first explored in studies of animal learning (Harlow, 1949), and has subsequently been studied in developmental psychology (Adolph, 2005, Kemp et al., 2010, Smith, 1995).

Transfer Learning

Humans also excel at generalizing or transferring generalized knowledge gained in one context to novel, previously unseen domains (Barnett and Ceci, 2002, Holyoak and Thagard, 1997). For example, a human who can drive a car, use a laptop computer, or chair a committee meeting is usually able act effectively when confronted with an unfamiliar vehicle, operating system, or social situation. Progress is being made in developing AI architectures capable of exhibiting strong generalization or transfer, for example by enabling zero-shot inferences about novel shapes outside the training distribution based on compositional representations (Higgins et al., 2016; Figure 2C). Others have shown that a new class of architecture, known as a progressive network, can leverage knowledge gained in one video game to learn rapidly in another, promising the sort of "far transfer" that is characteristic of human skill acquisition (Rusu et al., 2016a). Progressive networks have also been successfully employed to transfer knowledge for a simulated robotic environment to a real robot arm, massively reducing the training time required on the real world (Rusu et al., 2016b). Intriguingly, the proposed architecture bears some resemblance to a successful computational model of sequential task learning in humans (Collins and Koechlin, 2012, Donoso et al., 2014). In the neuroscience literature, one hallmark of transfer learning has been the ability to reason relationally, and AI researchers have also begun to make progress in building deep networks that address problems of this nature, for example by solving visual analogies (Reed et al., 2015). More generally however, how humans or other animals achieve this sort of high-level transfer learning is unknown, and remains a relatively unexplored topic in neuroscience. New advances on this front could provide critical insights to spur AI research toward the goal of lifelong learning in agents, and we encourage neuroscientists to engage more deeply with this question.

At the level of neural coding, this kind of transfer of abstract structured knowledge may rely on the formation of conceptual representations that are invariant to the objects, individuals, or scene elements that populate a sensory domain but code instead for abstract, relational information among patterns of inputs (Doumas et al., 2008). However, we currently lack direct evidence for the existence of such codes in the mammalian brain. Nevertheless, one recent report made the very interesting claim that neural codes thought to be important in the representation of allocentric

(map-like) spaces might be critical for abstract reasoning in more general domains (Constantinescu et al., 2016). In the mammalian entorhinal cortex, cells encode the geometry of allocentric space with a periodic "grid" code, with receptive fields that tile the local space in a hexagonal pattern (Rowland et al., 2016). Grid codes may be an excellent candidate for organizing conceptual knowledge, because they allow state spaces to be decomposed efficiently, in a way that could support discovery of subgoals and hierarchical planning (Stachenfeld et al., 2014). Using functional neuroimaging, the researchers provide evidence for the existence of such codes while humans performed an abstract categorization task, supporting the view that periodic encoding is a generalized hallmark of human knowledge organization (Constantinescu et al., 2016). However, much further work is required to substantiate this interesting claim.

Imagination and Planning

Despite their strong performance on goal-directed tasks, deep RL systems such as DQN operate mostly in a reactive way, learning the mapping from perceptual inputs to actions that maximize future value. This "model-free" RL is computationally inexpensive but suffers from two major drawbacks: it is relatively data inefficient, requiring large amounts of experience to derive accurate estimates, and it is inflexible, being insensitive to changes in the value of outcomes (Daw et al., 2005). By contrast, humans can more flexibly select actions based on forecasts of long-term future outcomes through simulation-based planning, which uses predictions generated from an internal model of the environment learned through experience (Daw et al., 2005, Dolan and Dayan, 2013, Tolman, 1948). Moreover, planning is not a uniquely human capacity. For example, when caching food, scrub jays consider the future conditions under which it is likely to be recovered (Raby et al., 2007), and rats use a "cognitive map" when navigating, allowing inductive inferences during wayfinding and facilitating one-shot learning behaviors in maze-like environments (Daw et al., 2005, Tolman, 1948). Of course, this point has not been lost on AI researchers; indeed, early planning algorithms such as Dyna (Sutton, 1991) were inspired by theories that emphasized the importance of "mental models" in generating hypothetical experiences useful for human learning (Craik, 1943). By now, a large volume of literature exists on AI planning techniques, including model-based RL methods, which seek to implement this forecast-based method of action selection. Furthermore, simulation-based planning, particularly Monte Carlo tree search (MCTS) methods, which use forward search to update a value function and/or policy (Browne et al., 2012), played a key role in recent work in which deep RL attained expert-level performance in the game of Go (Silver et al., 2016).

AI research on planning, however, has yet to capture some of the key characteristics that give human planning abilities their power. In particular, we suggest that a general solution to this problem will require understanding how rich internal models, which in practice will have to be approximate but sufficiently accurate to support planning, can be learned through experience, without strong priors being handcrafted into the network by the experimenter. We also argue that AI research will benefit from a close reading of the related literature on how humans imagine possible scenarios, envision the future, and carry out simulation-based planning, functions that depend on a common neural substrate in the hippocampus (Doll et al., 2015, Hassabis and Maguire, 2007, Hassabis and Maguire, 2009, Schacter et al., 2012). Although imagination has an intrinsically subjective, unobservable quality, we have reason to believe that it has a conserved role in simulation-based planning across species (Hassabis and Maguire, 2009, Schacter et al., 2012). For example, when paused at a choice point, ripples of neural activity in the rat hippocampus resemble those observed during subsequent navigation of the available trajectories ("preplay"), as if the animal were "imagining" each possible alternative (Johnson and Redish, 2007, Ólafsdóttir et al., 2015, Pfeiffer and Foster, 2013). Further, recent work has

suggested a similar process during non-spatial planning in humans (Doll et al., 2015, Kurth-Nelson et al., 2016). We have discussed above the ways in which the introduction of mechanisms that replay and learn offline from past experiences can improve the performance of deep RL agents such as DQN (as discussed above in Episodic Memory).

Some encouraging initial progress toward simulation-based planning has been made using deep generative models (Eslami et al., 2016, Rezende et al., 2016a, Rezende et al., 2016b) (Figure 2). In particular, recent work has introduced new architectures that have the capacity to generate temporally consistent sequences of generated samples that reflect the geometric layout of newly experienced realistic environments (Gemici et al., 2017, Oh et al., 2015) (Figure 2E), providing a parallel to the function of the hippocampus in binding together multiple components to create an imagined experience that is spatially and temporally coherent (Hassabis and Maguire, 2007). Deep generative models thus show the potential to capture the rich dynamics of complex realistic environments, but using these models for simulation-based planning in agents remains a challenge for future work.

Insights from neuroscience may provide guidance that facilitates the integration of simulation with control. An emerging picture from neuroscience research suggests that the hippocampus supports planning by instantiating an internal model of the environment, with goal-contingent valuation of simulated outcomes occurring in areas downstream of the hippocampus such the orbitofrontal cortex or striatum (Redish, 2016). Notably, however, the mechanisms that guide the rolling forward of an internal model of the environment in the hippocampus remain uncertain and merit future scrutiny. One possibility is that this process is initiated by the prefrontal cortex through interactions with the hippocampus. Indeed, this notion has distinct parallels with proposals from AI research that a separate controller interacts with an internal model of the environment in a bidirectional fashion, querying the model based on task-relevant goals and receiving predicted simulated states as input (Schmidhuber, 2014). Further, recent efforts to develop agents have employed architectures that instantiate a separation between controller and environmental model to effect simulation-based planning in problems involving the interaction between physical objects (Hamrick et al., 2017).

In enhancing agent capabilities in simulation-based planning, it will also be important to consider other salient properties of this process in humans (Hassabis and Maguire, 2007, Hassabis and Maguire, 2009). Research into human imagination emphasizes its constructive nature, with humans able to construct fictitious mental scenarios by recombining familiar elements in novel ways, necessitating compositional/disentangled representations of the form present in certain generative models (Eslami et al., 2016, Higgins et al., 2016, Rezende et al., 2016a). This fits well with the notion that planning in humans involves efficient representations that support generalization and transfer, so that plans forged in one setting (e.g., going through a door to reach a room) can be leveraged in novel environments that share structure. Further, planning and mental simulation in humans are “jumpy,” bridging multiple temporal scales at a time; for example, humans seem to plan hierarchically, by considering in parallel terminal solutions, interim choice points, and piecemeal steps toward the goal (Balaguer et al., 2016, Solway et al., 2014, Huys et al., 2012). We think that ultimately these flexible, combinatorial aspects of planning will form a critical underpinning of what is perhaps the hardest challenge for AI research: to build an agent that can plan hierarchically, is truly creative, and can generate solutions to challenges that currently elude even the human mind.

Virtual Brain Analytics

One rather different way in which neuroscience may serve AI is by furnishing new analytic tools for understanding computation in AI systems. Due to their complexity, the products of AI research often remain “black boxes”; we understand only poorly the nature of the computations that occur, or representations that are formed, during learning of complex tasks. However, by applying tools from neuroscience to AI systems, synthetic equivalents of single-cell recording, neuroimaging, and lesion techniques, we can gain insights into the key drivers of successful learning in AI research and increase the interpretability of these systems. We call this “virtual brain analytics.”

Recent work has made some progress along these lines. For example, visualizing brain states through dimensionality reduction is commonplace in neuroscience, and has recently been applied to neural networks (Zahavy et al., 2016). Receptive field mapping, another standard tool in neuroscience, allows AI researchers to determine the response properties of units in a neural network. One interesting application of this approach in AI is known as activity maximization, in which a network learns to generate synthetic images by maximizing the activity of certain classes of unit (Nguyen et al., 2016, Simonyan et al., 2013). Elsewhere, neuroscience-inspired analyses of linearized networks have uncovered important principles that may be of general benefit in optimizing learning these networks, and understanding the benefits of network depth and representational structure (McClelland and Rogers, 2003, Saxe et al., 2013).

While this initial progress is encouraging, more work is needed. It remains difficult to characterize the functioning of complex architectures such as networks with external memory (Graves et al., 2016). Nevertheless, AI researchers are in the unique position of having ground truth knowledge of all components of the system, together with the potential to causally manipulate individual elements, an enviable scenario from the perspective of experimental neuroscientists. As such, we encourage AI researchers to use approaches from neuroscience to explore properties of network architectures and agents through analysis, visualization, causal manipulation, not forgetting the need for carefully designed hypothesis-driven experiments (Jonas and Kording, 2017, Krakauer et al., 2017). We think that virtual brain analytics is likely to be an increasingly integral part of the pipeline of algorithmic development as the complexity of architectures increases.

From AI to Neuroscience

Thus far, our review has focused primarily on the role of neuroscience in accelerating AI research rather than vice versa. Historically, however, the flow of information between neuroscience and AI has been reciprocal. Machine learning techniques have transformed the analysis of neuroimaging datasets—for example, in the multivariate analysis of fMRI and magnetoencephalographic (MEG) data (Cichy et al., 2014, Çukur et al., 2013, Kriegeskorte and Kievit, 2013)—with promise for expediting connectomic analysis (Glasser et al., 2016), among other techniques. Going further, we believe that building intelligent algorithms has the potential to offer new ideas about the underpinnings of intelligence in the brains of humans and other animals. In particular, psychologists and neuroscientists often have only quite vague notions of the mechanisms that underlie the concepts they study. AI research can help, by formalizing these concepts in a quantitative language and offering insights into their necessity and sufficiency (or otherwise) for intelligent behavior.

A key illustration of this potential is provided by RL. After ideas from animal psychology helped to give birth to reinforcement learning research, key concepts from the latter fed back to inform neuroscience. In particular, the profile of neural signals observed in midbrain dopaminergic neurons in conditioning paradigms was found to bear a striking resemblance to TD-generated prediction errors, providing neural evidence that the brain implements a form of TD

learning (O'Doherty et al., 2003, Schultz et al., 1997). This overall narrative arc provides an excellent illustration of how the exchange of ideas between AI and neuroscience can create a "virtuous circle" advancing the objectives of both fields.

In another domain, work focused on enhancing the performance of CNNs has also yielded new insights into the nature of neural representations in high-level visual areas (Khaligh-Razavi and Kriegeskorte, 2014, Yamins and DiCarlo, 2016). For example, one group systematically compared the ability of more than 30 network architectures from AI to explain the structure of neural representations observed in the ventral visual stream of humans and monkeys, finding favorable evidence for deep supervised networks (Khaligh-Razavi and Kriegeskorte, 2014). Further, these deep convolutional network architectures offer a computational account of recent neurophysiological data demonstrating that the coding of category-orthogonal properties of objects (e.g., position, size) actually increases as one progresses higher up the ventral visual stream (Hong et al., 2016). While these findings are far from definitive as yet, it shows how state-of-the-art neural networks from AI can be used as plausible simulacra of biological brains, potentially providing detailed explanations of the computations occurring therein (Khaligh-Razavi and Kriegeskorte, 2014, Yamins and DiCarlo, 2016). Relatedly, properties of the LSTM architecture have provided key insights that motivated the development of working memory models that afford gating-based maintenance of task-relevant information in the prefrontal cortex (Lloyd et al., 2012, O'Reilly and Frank, 2006).

We also highlight two recent strands of AI research that may motivate new research in neuroscience. First, neural networks with external memory typically allow the controller to iteratively query or "hop through" the contents of memory. This mechanism is critical for reasoning over multiple supporting input statements that relate to a particular query (Sukhbaatar et al., 2015). Previous proposals in neuroscience have argued for a similar mechanism in human cognition, but any potential neural substrates, potentially in the hippocampus, remain to be described (Kumaran and McClelland, 2012). Second, recent work highlights the potential benefits of "meta-reinforcement learning," where RL is used to optimize the weights of a recurrent network such that the latter is able to implement a second, emergent RL algorithm that is able to learn faster than the original (Duan et al., 2016, Wang et al., 2016). Intriguingly, these ideas connect with a growing neuroscience literature indicating a role for the prefrontal cortex in RL, alongside more established dopamine-based mechanisms (Schultz et al., 1997). Specifically, they indicate how a relatively slow-learning dopaminergic RL algorithm may support the emergence of a freestanding RL algorithm instantiated with the recurrent activity dynamics of the prefrontal cortex (Tsutsui et al., 2016).

Insights from AI research are also providing novel perspectives on how the brain might implement an algorithmic parallel to backpropagation, the key mechanism that allows weights within multiple layers of a hierarchical network to be optimized toward an objective function (Hinton et al., 1986, Werbos, 1974). Backpropagation offers a powerful solution to the problem of credit assignment within deep networks, allowing efficient representations to be learned from high dimensional data (LeCun et al., 2015). However, until recently, several aspects of the backpropagation algorithm were viewed to be biologically implausible (e.g., see Bengio et al., 2015). One important factor is that backpropagation has typically been thought to require perfectly symmetric feedback and feedforward connectivity, a profile that is not observed in mammalian brains. Recent work, however, has demonstrated that this constraint can in fact be relaxed (Liao et al., 2015, Lillicrap et al., 2016). Random backward connections, even when held fixed throughout network training, are sufficient to allow the backpropagation algorithm to function effectively through a process

whereby adjustment of the forward weights allows backward projections to transmit useful teaching signals (Lillicrap et al., 2016).

A second core objection to the biological plausibility of backpropagation is that weight updates in multi-layered networks require access to information that is non-local (i.e., error signals generated by units many layers downstream) (for review, see Bengio et al., 2015). In contrast, plasticity in biological synapses depends primarily on local information (i.e., pre- and post-synaptic neuronal activity) (Bi and Poo, 1998). AI research has begun to address this fundamental issue. In particular, recent work has shown that hierarchical auto-encoder networks and energy-based networks (e.g., continuous Hopfield networks) (Scellier and Bengio, 2016, Whittington and Bogacz, 2017)—models that have strong connections to theoretical neuroscience ideas about predictive coding (Bastos et al., 2012)—are capable of approximating the backpropagation algorithm, based on weight updates that involve purely local information. Indeed, concrete connections have been drawn between learning in such networks and spike-timing dependent plasticity (Scellier and Bengio, 2016), a Hebbian mechanism instantiated widely across the brain (Bi and Poo, 1998). A different class of local learning rule has been shown to allow hierarchical supervised networks to generate high-level invariances characteristic of biological systems, including mirror-symmetric tuning to physically symmetric stimuli, such as faces (Leibo et al., 2017). Taken together, recent AI research offers the promise of discovering mechanisms by which the brain may implement algorithms with the functionality of backpropagation. Moreover, these developments illustrate the potential for synergistic interactions between AI and neuroscience: research aimed to develop biologically plausible forms of backpropagation have also been motivated by the search for alternative learning algorithms. Given the increasingly deep networks (e.g., >20 layer) used in AI research, factors such as the compounding of successive non-linearities pose challenges for optimization using backpropagation (Bengio et al., 2015).

Conclusions

In this perspective, we have reviewed some of the many ways in which neuroscience has made fundamental contributions to advancing AI research, and argued for its increasingly important relevance. In strategizing for the future exchange between the two fields, it is important to appreciate that the past contributions of neuroscience to AI have rarely involved a simple transfer of full-fledged solutions that could be directly re-implemented in machines. Rather, neuroscience has typically been useful in a subtler way, stimulating algorithmic-level questions about facets of animal learning and intelligence of interest to AI researchers and providing initial leads toward relevant mechanisms. As such, our view is that leveraging insights gained from neuroscience research will expedite progress in AI research, and this will be most effective if AI researchers actively initiate collaborations with neuroscientists to highlight key questions that could be addressed by empirical work.

The successful transfer of insights gained from neuroscience to the development of AI algorithms is critically dependent on the interaction between researchers working in both these fields, with insights often developing through a continual handing back and forth of ideas between fields. In the future, we hope that greater collaboration between researchers in neuroscience and AI, and the identification of a common language between the two fields (Marblestone et al., 2016), will permit a virtuous circle whereby research is accelerated through shared theoretical insights and common empirical advances. We believe that the quest to develop AI will ultimately also lead to a better understanding of our own minds and thought processes. Distilling intelligence into an algorithmic construct and comparing it to the human brain might

yield insights into some of the deepest and the most enduring mysteries of the mind, such as the nature of creativity, dreams, and perhaps one day, even consciousness.