Stretch goals: By 2040, users control and trust their AI systems to use persuasion when ethically appropriate, and to alert and protect them from intentional manipulation by others (humans or machines). Milestones along this path include—

Milestones

5 years: AI systems follow ethical guidelines when knowingly attempting to persuade users.

10 years: AI systems understand and notify a user when others (humans or machines) are acting persuasively against the preferences of the user.

15 years: AI systems detect and control malicious persuasion or intentional manipulation by others (humans or machines).

3.2.4 CONCLUSIONS

There is a long way to go before AI systems can interact with people in truly natural ways and can support interaction between people online. Building AI systems that are transparent and that can explain their rationales to end users and developers alike will help enable the development of systems that are unbiased and can be trusted, for it is through explanation that biases can be discovered. In 20 years, if research is properly supported, we will see AI systems that can communicate naturally and collaboratively with a diverse range of users regardless of language, dialect, or ability. We will see AI systems that can communicate using multiple modalities understanding how gesture, speech, images, and sensors complement each other. We will see AI systems that can support the human creative process, suggesting hypotheses and enabling better collaboration. We will see unbiased, transparent AI systems performing high-stakes tasks that humans can trust.

3.3 A Research Roadmap for Self-Aware Learning

3.3.1 INTRODUCTION AND OVERVIEW

The field of machine learning (ML) seeks to provide learning and adaptation capabilities for AI systems. While classical computing based on programming relies on human ingenuity to anticipate the wide range of conditions in which a system may find itself, AI systems must learn autonomously from a variety of sources including labeled training data, tutorial demonstration, instruction manuals and scientific articles, and through interactions with the physical and social world. They must also be able to adapt in order to compensate for changes in the environment and context, changes in the goals of the users, and changes in software, sensors, and computer and robotic hardware. All AI systems are limited by their sensors, their training experiences, and the vocabulary and representations in which their knowledge of the world is expressed; nonetheless, they need to behave robustly in the face of these limitations. This requires a deep understanding of their own uncertainty and a mandate to confront difficult decisions with caution. This is especially important in the face of adversarial attacks that seek to find and exploit the weaknesses and blind spots of all computer systems. Avoiding unintended biases that may be inherent in the way a system has been trained is another important consideration.

While machine learning methods transform the way we build and maintain AI systems, they do not eliminate the need for software engineering. Programming is still required, but at a higher level, where a human expert selects the training materials, designs the representation (i.e., vocabulary and structure) in which the learned knowledge will be captured, and specifies the measures of success. It is often the case that very large quantities of labeled data are needed for training and testing purposes, introducing significant labor and expense. Another critical task of AI designers is to check the correctness of what has been learned. This can be a real challenge, as it requires the AI system to be able to visualize and explain the learned knowledge. Ideally, an AI system would be aware of its own capabilities—it should be able to analyze what it has learned, characterize its boundaries and limitations, and proactively seek opportunities to improve its performance through further learning.

This section of the Roadmap focuses on machine learning and is organized into four areas: learning expressive representations, creating trustworthy systems, building durable systems, and integrating AI and robotic systems.

1. Learning Expressive Representations. Most machine learning today works by discovering statistical correlations between attributes of the input (e.g., a group of pixels in an image) and the value of some target variable (e.g., the kind of object in the image). While this produces excellent results on some tasks, experience reveals that those approaches are very brittle. Slight changes in image size, lighting, and so on rapidly degrade the accuracy of the system. Similarly, while machine learning has improved the quality of machine translation (e.g., from English to Spanish), the results often show that the computer has only a very shallow understanding of the meaning of the translated sentences.

A major challenge for the next 20 years is to develop machine learning methods that can extract, capture, and use knowledge that goes beyond these kinds of surface correlations. To accomplish this, researchers must discover the right representations for capturing more sophisticated knowledge, as well as methods for learning it. For example, scientists and engineers think about the world using mechanistic models that satisfy general principles such as conservation of mass and energy. These models are typically causal, which means that they can predict the effects of actions taken in the world and, importantly, that they can explain what would have happened if different actions had been taken in the past. Two important research directions in AI are how to effectively learn causal models and how to integrate existing mechanistic models into machine learning algorithms, yielding more robust systems while reducing the need for brute-force training and large quantities of labeled data.

Many AI systems represent knowledge in the form of symbolic statements, similar to equations in algebra. These symbolic representations support flexible reasoning, and they have been applied with great success in areas as diverse as manufacturing, transportation, and space flight. Deep neural networks offer an alternative approach in which symbols are replaced by numbers. For example, the word "duck" might be encoded as a list of 300 numbers. Since similar words have similar encodings, deep neural networks have proven very useful in natural language processing tasks such as translation. However, the meaning of the individual numbers may not bear any resemblance to how humans reason about the noun "duck"—that ducks are members of the class of birds and the smaller class of waterfowl, etc. Rather, the internal variables in a neural net are abstract quantities extracted by the network as it learns from the examples that it is given. There is important traction to be gained from symbolic knowledge, though, and a critical research challenge is to develop ways to combine numeric and symbolic representations to obtain the benefits of both; that is, systems that are both highly accurate and produce decisions that are more easily interpretable by humans.

2. Trustworthy Learning. Today's AI systems perform acceptably in low-stakes domains, but that level of performance can be unacceptable in high-stakes environments such as medical diagnosis, robotics, loan approvals, and criminal justice. To be trusted with such decisions, our AI systems must be aware of their own limitations and be able to communicate those limitations to their human programmers and users so that we know when to trust these systems. Each AI system should have a *competence model* that describes the conditions under which it produces accurate and correct behavior. Such a competence model should take into account shortcomings in the training data, mismatches between the training context and the performance context, potential failures in the representation of the learned knowledge, and possible errors in the learning algorithm itself. AI systems should also be able to explain their reasoning and the basis for their predictions and actions. Machine learning algorithms can find regularities that are not known to their users; explaining those can help scientists form hypotheses and design experiments to advance scientific knowledge. Explanations are also crucial for the software engineers who must debug AI systems. They are also extremely important in situations such as criminal justice and loan approvals where multiple stakeholders have a right to contest the conclusions of the system. Some machine learning algorithms produce highly accurate and interpretable models that can be easily inspected and understood. In many other cases, though, machine learning algorithms produce solutions that humans find unintelligible. Going forward, it will be important to assure that future methods produce human-interpretable results. It will be equally important to develop techniques that can be applied to the vast array of legacy machine learning methods that have already been deployed, in order to assess whether they are fair and trustworthy.

3. Durable ML Systems. The knowledge acquired through machine learning is only valid as long as the regularities discovered in the training data hold true in the real world. However, the world is continually changing, and we need AI systems that can work for significant stretches of time without manual re-engineering. This will require machine learning methods that can detect and track trends and other changes in the data and adapt to those changes when possible (and appropriate). Our AI systems also need to remember the past, because there are often cyclical patterns in data over time. A medical diagnosis system that learns to diagnose flu one winter should retrieve that learned knowledge the next winter, when the flu season begins again, and determine how to sensibly incorporate that knowledge into the current environment. Existing machine learning systems lack this form of long-term memory and the ability to effectively merge new and old knowledge.

People are able to detect novel situations (objects, events, etc.) and learn new knowledge from just a few examples. We need so-called "one-shot" or "few-shot" learning algorithms that can do the same; this is likely to require the kinds of expressive representations discussed above. People can also learn skills in one setting (e.g., driving on the left in Britain) and transfer them to substantially different situations (e.g., driving on the right in the US). Such "transfer learning" is not perfect, but it is much faster than having to learn to drive all over again from scratch.

Of course, it is not always possible to avoid the need for retraining. Hence, we need methods that can detect when the current learned model is no longer applicable and flag the system for wholesale re-engineering and careful re-testing.

4. Integrating AI and Robotic Systems. Robotics has made tremendous advances in the past two decades. Autonomous driving in both urban and off-road environments has entered a stage of rapid commercialization, attracting immense investments of capital and personnel from both private venture capital and large public technology and automobile companies. Robotic manipulation in structured or semi-structured environments is fairly well understood and has been widely adopted in industry situations that allow for carefully co-engineered workspaces that protect humans in proximity, such as assembly lines for manufacturing.

Still, there are many challenges facing robotics in the next two decades. Current robotic systems lack a holistic understanding of their own behavior, especially when AI technologies are introduced. Moreover, today's robots are not able to work in unstructured home/hospital environments in assisting and caregiving roles, where they must interact safely and effectively with people and where they must deal with non-rigid objects such as cloth and cables. Interaction with humans—both when humans are training the robots and when the robots are assisting or caring for the humans—requires much deeper knowledge about both the physical and social worlds. Acquiring and using this knowledge will require learning from data coming from multiple modalities such as speech, prosody, eye gaze, and body language.

In recent years, there has been impressive progress in both software and hardware for robotics. Sensors such as cameras, radars, and lidars have become smaller, faster, cheaper, and more accurate under a wider envelope of operating environments. Frameworks such as the open-source Robot Operating System (ROS) have grown in scope and quality and enjoy wide adoption and support in both industry and academia. As a major step toward more flexible, adaptable robot systems, there is both a strong need and a powerful opportunity to integrate advances in AI and machine learning to create intelligent software middleware layers that integrate sensing, reasoning, and acting in ways that are tuned to the current needs of the robot and the environment. More work is also needed to establish standards for testing, security, deployment, and monitoring of intelligent robotic systems.

3.3.2 SOCIETAL DRIVERS FOR EXPRESSIVE, ROBUST, AND DURABLE LEARNING

This section presents motivating vignettes for five of the societal drivers identified in the overall AI Roadmap. These highlight the research breakthroughs required in each of the four areas listed above that will be necessary to build AI systems that could successfully realize these vignettes.

ENHANCE HEALTH AND QUALITY OF LIFE

Vignette 14

In the spring of 2040, Sue founds a startup pharmaceutical company with the goal of creating a drug for controlling childhood asthma. If she had tried to develop a drug back in 2020, she would have needed to first identify a target receptor to inhibit. Then she would have needed to synthesize and test many different candidate compounds to find a good inhibitor. Following that, she would need to spend millions of dollars for phase I and phase II trials—trials that most drug compounds fail. But thanks to 20 years of investment in AI, medicine, biology, and chemistry, the process of drug design has been revolutionized. A consortium of research hospitals has standardized their health records and deployed a robust data-collection infrastructure that includes advanced wearable sensors as well as a carefully deployed sensor network over the city. This has produced an extensive set of data that capture fine-grained information about health and daily living, including disease and quality of life outcomes. Concurrently, machine learning analysis of the tens of thousands of experiments performed by biologists has produced high-fidelity simulators for basic metabolism and many other biological processes. By linking these simulations and medical records, AI algorithms can now identify causal hypotheses for many disease processes, including asthma. Sue's company uses these resources to identify a small number of candidate target receptors. The company focuses its effort on creating and validating an in vitro system for measuring binding affinity to those receptors. It then licenses large combinatorial libraries of small molecules (the result of federal investments in organic chemistry), and screens them against these receptors. Machine learning strategies developed in cooperation with chemists permits rapid optimization of the shape and structure of the candidate drug molecules. These are then processed by the metabolism simulators to predict required dosages and identify potential side effects. The handful of candidate drug molecules that pass these tests then undergo clinical trials. Health data from the research hospital consortium—maintained with algorithmic assurances of privacy and anonymity—make it easy to recruit participants for these trials. The candidate drugs survive all phases of the trials because of the strong data and simulations that have guided the design process. The availability of these open data and simulation resources dramatically lowers the cost of bringing drugs to market, greatly reduces side effects, and boosts overall health and quality of life.

Research Challenges: Realizing this vision of improved health care will require a number of breakthroughs in machine learning:

1. Learning Expressive Representations: To analyze the scientific literature and extract causal hypotheses, machine learning for natural language processing needs to go beyond surface regularities of words to a deep understanding of the organ systems and biological and chemical processes described in each paper, the scientific claims made by the paper, and the evidence supporting those claims. It must integrate this information into a growing knowledge base that carefully quantifies the uncertainty arising both from the experiments described in the papers and from the AI system's interpretation of those papers.

To relate the data collected from patients to the underlying biology, machine learning methods must form and evaluate causal models. These will lie at the heart of the simulators of metabolism and other biological processes.

Finally, to optimize the shape and structure of candidate drug molecules, machine learning models must be integrated with physical models of molecule conformation and dynamics.

2. Trustworthy Learning: To carry out clinical trials, Sue's company, the hospital consortium, and the patients (as well as their physicians) must all trust the results of the machine learning analysis. Proper quantification of all sources of uncertainty must be achieved, and all stakeholders (including the developers of the AI system) must be able to obtain useful explanations for why certain drug compounds were chosen and others rejected, why certain doses were chosen, and why side effects are predicted to be minimal. Privacy must be protected in appropriate ways throughout this process.

3. Durable ML Systems: The laboratory procedures in the published literature are under continuous evolution and improvement, as are the sensors and medical tests recording information about the patients. The AI systems analyzing the literature and patient health data must manage these changes as well as handling the extreme heterogeneity of the patient populations in different hospitals. As new biological discoveries are made, previous simulations must be re-executed to check whether existing drugs may be causing unobserved side effects or whether existing drugs may have new positive applications.

4. Integrating AI and Robotic Systems: The use of wearable sensors and the deployment of sensor networks over the city introduces many challenges. Sensors may not be fully reliable and may fail, so the data collected may not be consistent. As data is collected, sensor systems may learn a model of which sensors fail and how. The systems must discover which frequency of collection is most appropriate for the type of data being collected. For example, on rainy days the pollution may be reduced and data can be collected at longer time intervals, while on hot days the network should activate all sensors and collect data more often.

REINVENT BUSINESS INNOVATION AND COMPETITIVENESS

Vignette 15

Juan is an artist who wants to turn his passion for personalized augmented reality board games into a business. To design such games, Juan uses an intelligent design framework that plays the proposed game with him, perceives his behavior and enjoyment level, and suggests ways to improve the game's rules and layout. Once he is happy with his design, Juan uploads it to a prototype manufacturing site where the local manufacturer quickly teaches a robot how to build and package the game. The robot employs multimodal human interaction to learn from its trainer. Even though the game components are unique to each user, including a wide range of potential materials, the robot can now deploy robust manipulation skills to assemble the novel game components into customized packaging. The next morning, game testers receive initial prototypes and give it rave reviews, and demand for the game explodes. The trained robot shares what it learned through building and packaging Juan's game to hundreds of robots at other manufacturing sites. Because each robot is slightly different, the learned models have to be adapted to work with their particular hardware and software setup. The new robots immediately begin making games to meet the demand, and Juan is able to rapidly expand his sales.

Research Challenges: This vision requires research advances in machine learning along the four areas mentioned:

1. Learning Expressive Representations: Juan's game design framework understands his desire to build a fun game. It is able to learn from multiple modalities, including video cameras, audio microphones, and wearable physiological sensors, as well as from spoken statements that Juan makes. This allows it to understand Juan's intent and emotional response at a deeper level, which in turn allows it to propose improvements to the game.

2. Durable ML Systems. The game design framework is able to transfer what it has learned from interacting with Juan to learning from other customers. Similarly, the manufacturing robots can transfer their knowledge to other companies, where the robots are slightly different and are manufacturing similar, but not identical, game pieces.

3. Integrating AI and Robotic Systems. The prototype manufacturer is able to teach the robot how to build Juan's game using direct interaction, including verbal communication and visual demonstration. The robot assistant is adept at handling a wide variety of physical materials, both rigid and non-rigid, to produce games that are highly customized.

ACCELERATE SCIENTIFIC DISCOVERY AND TECHNOLOGY INNOVATION

Vignette 16

Aishwarya is a climate scientist trying to make predictions of future climate at the local and regional scale. It is essential that such predictions correctly quantify uncertainty. She chooses a climate model that is based on mathematical models of atmospheric physics, solar radiation, and land surface-atmosphere interactions. Unfortunately, running the model at the required fine level of detail is not possible due to the computational cost and the lack of sufficient observation data. Fortunately, recent advances in ML research have produced new physics-aware ML approaches that learn from data while incorporating knowledge about the underlying physics. These approaches run efficiently and produce models at a much finer level of detail than the original climate models. This makes it easy to run multiple models efficiently, which in turn allows Aishwarya to provide clear uncertainty bounds for the resulting predictions.

The results of these models are then used by Jia, who works at FEMA. Using machine learning methods, she combines climate predictions under different policy scenarios (no change versus reduced carbon emissions, etc.) to identify regions that are most vulnerable to extreme weather events such as hurricanes, floods, droughts, heat waves, and forest fires. With the aid of these causal models, she can plan appropriate responses. For example, her physics-aware ML model produces inundation maps in response to extreme meteorological events (hurricane, heavy rain) to identify areas of flooding, which is fed into an AI system for smart cities to perform evacuation planning, emergency relief operations, and planning for long-term interventions (e.g., building a sea wall to ward off storm surge).

Thanks to the 20 years of research investment in AI, physics-aware machine learning techniques are available that can process multimodal, multi-scale data and also handle heterogeneity in space and time, as well as quantify uncertainty in the results. The combination of physics-based climate and hydrological models with machine learned components allows Jia to produce more accurate predictions than would be possible with pure physics-based or pure machine learned models alone. This hybrid approach also generalizes better to novel scenarios, identifying new threats that could result in injury or death. In 2035, these models are applied to revise flood maps, saving many lives in the floods caused by hurricane Thor in North Carolina.

Research Challenges: This vignette illustrates the productive integration of mathematical models based on first principles with massive quantities of real-world data. Each approach complements and extends the other and helps address the inherent limitations of each. The methods apply hard-earned physics knowledge but also detect when it no longer applies and must be adapted and refined to cover new situations.

1. Learning Expressive Representations. Mathematical models based on physical laws, such as conservation of matter and energy, are causal and generalizable to unseen scenarios. In contrast, the statistical patterns extracted by traditional machine learning algorithms from the available observations often violate these physical principles (especially when extrapolating beyond the training data). On the other hand, the mathematical models are not perfect. They are necessarily approximations of reality, which introduces bias. In addition, they often contain a large number of parameters whose values must be estimated with the help of data. If data are sparse, the performance of these general mathematical models can be further degraded. While machine learning models have the potential to ameliorate such biases, the challenge is to ensure that they respect the required physical principles and are generalizable to unseen scenarios.

2. Durable ML Systems. Building Aishwarya's model requires combining sparse data measured at different locations around the planet, measured at different spatial and temporal scales using different instruments and protocols, and stored in different formats. Much of the knowledge brought to bear is adapted from different circumstances, from past natural disasters that occurred in other locations but with similar features. One-shot and few-shot learning are applied to anticipate and address previously unseen situations that could have tragic consequences.

3. Trustworthy Learning. Aishwarya seeks to properly account for all sources of uncertainty, including those that result from transferring knowledge from some parts of the planet to others and from one model to another. In this context, machine learning methods are needed that can assess uncertainty in transfer learning. Jai and FEMA will be making life-and-death decisions based in part on the uncertainties in Aishwarya's predictions and in part on uncertainties in the hydrological models.

SOCIAL JUSTICE AND POLICY

Vignette 17

During a training procedure in a virtual reality facility, Officer Smith and the more experienced Officer Rodriguez are dispatched to a house to handle a situation in which a husband and wife are having a very loud dispute with threats of violence that caused the neighbors to become concerned. Each one tells Officer Smith a story that conflicts with the other. They are very manipulative, using emotions, misrepresentations, and guilt to manipulate each other and the officers. Smith's challenge is to say the right things to de-escalate the situation and resolve the dispute. She also needs to say things in the right way: the virtual characters are perceiving her body position and body language, as well as her eye gaze and facial expressions. If she is either too assertive or not assertive enough (which will be indicated by both what she says and how she says it), things will go badly. In this case, Smith becomes angry and moves too abruptly; the husband pulls a gun and Officer Rodriguez shoots in reaction. During the after-action review, Officer Smith walks through the scenario, and an AI coach presents things she did badly and encourages her to think about how she could have done better. Officer Smith is also able to "interview" the virtual husband and wife to find out what they were thinking and how they perceived her statements and actions.

Research Challenges: The use of AI in advanced training involving life-or-death decisions and split-second reactions requires accurate modeling all of the visible and audible cues that humans depend on as well as hidden motivations, coupled with the ability to engage in a collaborative dialog on what happened after-the-fact.

1. Learning Expressive Representations. To create the virtual characters, machine learning and computer vision must be trained on data from actors in order to build models of the motivation and personality of each character enough to capture how those are revealed through their speech, emotions, and body language. This requires deep understanding of the meaning of their spoken language as well as prosody, facial expression, and physical interactions. The learned simulation models must be able to synthesize responses to all of the new behaviors that a trainee police officer might exhibit.

2. Trustworthy Learning. The virtual characters must be able to explain (in the after-action review) what they were thinking and why, which requires explainable AI as well as the capability to synthesize natural speech in response to questions from the trainee officer.

TRANSFORM NATIONAL DEFENSE AND SECURITY

Vignette 18

Susan works in the emergency operations center of a smart city. There is a major hurricane bearing down, and her task is to muster hundreds of thousands of sensors, cameras, and actuators to plan and execute the safe evacuation of 10 million people. The system that she uses for this procedure has been pre-trained to estimate street capacity based on normal traffic and to predict where residents are typically located at each hour of the day based on behavioral data. From this, it can plan how to set traffic signals to facilitate rapid evacuation of the city. However, additional data must be integrated into the system in real time—weather predictions, flooding, road construction, vehicle accidents, downed utility poles—to which the plan must be adapted. In adapting to changing situations, the system can also make real-time use of residents' social media postings as they provide feedback on their own personal situations in the storm.

This sophisticated planning system has been designed to be resilient to physical damage and to surprises it cannot correct but must work around. Among other things, it can detect when the situation is beyond the conditions it can successfully handle, at which point it falls back on different systems or even hands control over to manual processes. For example, as smart sensors are damaged by the storm, missing data can be interpolated from nearby sensors, or inferred from regional models; as internet infrastructure fails in some neighborhoods, human staffers operating radios can be called in.

With the goal of instilling civil unrest, a nation state at conflict with the US seizes the opportunity to launch a coordinated campaign of disruption to interfere with the evacuation. They release malware to disrupt automated car systems and a bot-powered misinformation campaign on social media. Fortunately, safety systems in the individual cars detect inconsistencies between the user's commands and the car's physical behavior, and the city's social media analysis software identifies the bot activity and deploys communication countermeasures that have been found to be effective across many different human populations. Susan and her teammates then apply a collaborative planning system to take back control of some parts of the system, and employ the failover methods originally designed for storm resilience to route around the maliciously compromised portions of the system.

Research Challenges. In this scenario, the AI system combines smart detection, software countermeasures, and overall resilience to mitigate and recover from the attacks. Achieving all of this requires addressing important research challenges:

1. Durable ML Systems. Traffic behavior and road capacities change over time as the vehicle mix changes: e.g., the percentage of highly automated (or even autonomous) connected vehicles. Electric vehicles that run out of battery power cannot be easily recharged, which can affect these dynamics. One challenge will be to develop machine learning methods that can deal with such change in the distribution of vehicle capabilities.

A second need for managing change arises in the context of social media traffic. The behavior of people on social media is constantly changing. Any machine learning system that attempts to model the state of the city from social media posts must be prepared to deal with such changes.

A third durability requirement arises from the need to test such systems. One could accomplish this using reconstructions of historical weather evacuations, for instance. These occurred in situations very different from the current one, however, with different traffic signals and sensing capabilities, different driving behavior, and so on, and hence must be adapted to the current scenario.

2. Trustworthy Learning. The ubiquity of smartphones and the advent of connected vehicles provides an opportunity to coordinate the behavior of different vehicles. A central planning system could give instructions to each driver (or each vehicle), but, of course, the drivers must trust this information. Another possibility is to create a kind of market in which people could indicate their preferred destination and the planning system would coordinate traffic so that people could get to their destinations quickly. This could even be integrated with social media analysis—e.g., so that the system could ensure that family groups were all sent to the same destination in the case of an evacuation. In all cases, the system should be able to explain why it is sending people along the chosen route.

3.3.3 TECHNICAL CHALLENGES FOR SELF-AWARE LEARNING

Learning Expressive Representations

A central tenet of artificial intelligence is to build models of the world (world knowledge) and use them to both track the state of the world and formulate plans to achieve goals. While routine intelligent behaviors can be achieved via reactive (stimulus-response) mechanisms, the ability to deal with complex and novel situations is believed to require model-based reasoning. Until recently, machine learning systems were designed to map directly from an input representation (e.g., a sentence) to an output representation (e.g., a parse tree that captures its syntactic structure). The difficulty of learning is related to the size of the gap between the input and the output, and consequently this gap was kept small in order to ensure that the learning was feasible.

This section outlines three different research threads that will be required to span these gaps: automatic learning of 1) intermediate representations that act as stepping stones, 2) models that capture cause and effect, and 3) models that capture underlying mechanisms.

Learning Better Intermediate Representations

With the advent of deep learning—networks that use many *layers* of computational elements to map from input to output—it has been possible in some cases to learn across much larger gaps. In computer vision, our algorithms are able to map from raw image pixels to high-level image summaries (e.g., a sentence describing the image or a division of the image into its objects and background). There is some evidence that machine learning algorithms achieve this by deducing useful *intermediate representations*: ways of capturing or describing information at a level that is "between" those of the input and output. Many machine learning algorithms, for instance, leverage the notion of an *embedding space*. In word embedding, for example, each word is represented by a list of numbers. Many different natural language processing tasks can then be attacked by first mapping the words into this embedding space and then learning appropriate transformations in that space. Machine-translation systems work in this fashion, mapping the words into this embedding space prior to translating them. Similarly, systems for visual question answering proceed by combining the embedded

representation of the words and the information in the image. This staged decomposition of the task has a number of advantages besides effectiveness, including robustness and extensibility; For example, systems trained on the ImageNet database of 1000 object categories can easily be "fine tuned" to work on new problems, such as medical or biological recognition. The fine tuning typically involves retraining only the final layer or two of the network (which may have 150 layers in total).

Reusable object recognition and word embeddings show that *some* forms of reusable knowledge can be learned by computer algorithms. But it is unclear how to move from these two cases to general approaches for the many problems in which reusable intermediate representations can play important roles. For example, it is widely believed that successful natural language understanding requires some event schema. Similarly in computer vision, it would be useful to build a three-dimensional model of all of the objects in an image, plus the light sources and the position and settings of the camera at the time the image was taken. In robot locomotion, there is a natural hierarchy in which the overall navigational goal ("go out the door") is the most abstract, the main trajectory—the path around the desk and through the door—is next, and at the lowest level are the individual actuator commands that direct the robot's wheels accordingly.

The main technical challenge in this endeavor is to develop such representations, either by automatic discovery from experience or by some mix of manual design and algorithmic learning. These must include transformations that bridge from the numerical representations (e.g., word embeddings, trajectory coordinates) to the symbolic representations (e.g., parse trees, movement plans) that support flexible abstract reasoning. A critical consideration in such mappings is for the AI system to know which detailed aspects of the world can be ignored within the current context. It is rare that a person needs to know the exact shape and location of every object in a room. When we are searching for our car keys, we focus on the most relevant size range for objects. In the opposite direction, the AI systems must learn how to translate abstract meanings and plans into numerical commands for generating actions (spoken language, robot motion, etc.)

Stretch goals: By 2040, we should have achieved a fundamental understanding of the relationships between knowledge representation, learning, and reasoning. This will enable us to create AI systems that understand natural language and engage in conversations about both the physical and the social worlds. Such systems will be able to learn from natural language instruction and in turn explain and teach via natural language. The underlying representations will allow robots to reason effectively about the physical world through direct manipulation, pointing, computer vision, and natural language. Milestones along this path include—

5 years: Robotics is likely to be the first field where we have good ideas about intermediate representations needed for both visual perception and physical action, making it a good starting point. Using simulated physical worlds, develop sufficient intermediate representations to enable robotic systems to learn complex, hierarchical activities through demonstration and practice. Initial work should focus on manipulation and assembly, because these will also be immediately useful.

10 years: Develop generalized representation-learning methodologies and use them to build effective representations of social and mental worlds in application areas where humans interact with computers (e.g., conversational agents with a broad competence). Achieve good performance in human-computer interaction by learning from examples, demonstrations, and natural language instruction.

15 years: Using these representations, create an AI system that can read a textbook, work the exercises, and prepare new instructional materials for customized education.

Learning Causal Models

Developments over the past decade in causal modeling hint at the power of these techniques in machine learning systems. Existing algorithms rely entirely on correlative patterns between the input and output variables in the training data. These correlations can be very brittle, so that even very small changes in the problem can invalidate the learned knowledge. In contrast, most scientific theories are causal, and they allow us to learn causal regularities on Earth and extrapolate far away. For example, the causal theories of physics allow us to predict with high confidence what happens inside black holes even though it is impossible

for us to make observations there. The potential applications are significant. Advances in understanding causal inference could, for example, greatly advance research in biology and medicine, where these relationships are the key to designing medical interventions to fight disease.

The technical challenge here is to discover the causal relationships and assess their strength. The randomized trial (or A/B test) is the best understood method for achieving this, but it requires good data. In some settings, causal relationships can be confirmed with high confidence even in the face of noise. In some situations, one can even assess the causal impact of one variable on *multiple* result variables. Research is needed to develop a comprehensive understanding of the set of situations under which causal relationships can be inferred. This will be particularly challenging in situations where feedback loops are present, producing circular or bidirectional causal relationships. This is commonplace in real-world problems.

Stretch goals: By 2040, we will have a comprehensive theory of learning and inference with causal models and with models that mix causal and correlational inference. Every machine learning package and every application will include causal modeling to the extent permitted by the data and the problem. Milestones along this path include—

5 years: An open-source machine learning package will support general learning of causal models for supervised learning with semantically useful features. Applications in at least two domains (e.g., biology and medicine) will employ learned causal models.

10 years: Robotic and cyber-physical systems will learn causal models based on active exploration and experimentation in the domain. This will greatly accelerate learning in these domains.

20 years: Development of machine learning systems that can discover, explain, and apply new causal models for significant physical or social phenomena that rival what human experts would consider publishable in a scientific journal.

Leveraging Mechanistic Models to Build Robust Systems

The scientific community is increasingly considering machine learning as an alternative to hand-crafted mathematical and mechanistic models (e.g., Newton's laws of motion). These traditional scientific models are broadly powerful, but they have some major limitations, including the effort required to create them. Machine learning models, on the other hand, require little human effort, but they are "black boxes." They cannot explain their results in the language of the scientists: physical relationships, mechanistic insights, etc. There are other challenges as well. Unless they are provided with adequate information about the physical mechanisms of complex real-world processes, in the form of comprehensive training data, machine learning approaches can produce incorrect answers, false discoveries, and serious inconsistencies with known science. This reflects shortcomings in the training data, which may not be rich enough to capture the important scientific relationships of the problem at hand. It also reflects the tendency of current machine learning algorithms to focus on correlational rather than causal models (see above).

Leveraging scientific knowledge, both directly (e.g., by incorporating fundamental physical laws) and indirectly (e.g., as embodied in mechanistic models), is one way to address these issues. Incorporation of relevant physical constraints into a machine learning model, such as conservation of mass and energy, will ensure that it is both physically realistic and generalizable to situations beyond those covered in the training data. Indeed, incorporation of this kind of knowledge can reduce the amount of data required for successful training. Simulations of science-based models can even be used to generate synthetic data to "bootstrap" the training process so that only a small amount of observation data is needed to finalize the model.

Producing science-aware machine learning models will require addressing a number of challenges. Effective representation of physical processes will require development of novel abstractions and architectures that can simultaneously account for evolution at multiple spatial and temporal scales. Effective training of such models will need to consider not just accuracy (I.e., how well the output matches the specific observations) but also overall scientific correctness. This will become even more challenging when the underlying system contains disparate interacting processes that need to be represented by a collection of physical models developed by distinct scientific communities (e.g., climate science, ecology, hydrology, population dynamics, etc.).

Stretch goals: By 2040, we will have AI tools and methodologies that merge physical/process-based models with machine learning in a manner that effectively scaffolds the integration of multiple, disparate models that cover multiple aspects of complex systems (e.g., climate, weather, hydrology, ecology, economics, agriculture, and natural disaster mitigation and recovery). Such unified models could identify areas that are expected to be hit by hurricanes of increased intensity (due to climate change); project the path of a specific hurricane; identify areas to be impacted by flooding; create projections about damage to people, infrastructure, and biodiversity; and plan for emergency help or adaptation decisions (building a more resilient power grid, higher sea walls, etc.). Milestones along this path include—

5 years: Identify several use cases where the science is well understood but the physical models have limited performance. Build machine learning models that can leverage scientific knowledge to significantly exceed the performance of the associated state-of-the-art physical model.

10 years: Expand these efforts to a larger and more diverse set of domains with the goal of detecting commonalities in causal modeling tasks in those domains and developing a methodology that is applicable to a wide range of problems (e.g., climate, hydrological, ecological, econometric, population dynamics, crop yield, and social models).

15 years: Extend these tools and methodologies to handle cases where many disparate models are collectively used to solve a complex problem. Address problems of end-to-end debugging and explanation. Combine statistical and symbolic AI to help scientists come to an improved understanding of causes and effects of physical processes.

Trustworthy Learning

For most of its history, artificial intelligence research has focused on building systems to solve specific problems, whether it be recognizing objects in images or winning at the game of chess. But as AI is deployed in commerce, healthcare, law enforcement, and the military, it is critical that we know when we should and should not trust our AI systems. In this section, we consider the reasons why we might not trust these systems and suggest research directions to address these "threats to trust."

Every engineered system is based on a set of assumptions; if those assumptions are violated, that may be a reason to distrust the system. In machine learning, the assumptions typically include: 1) the training data capture a representative picture of the problem at hand, 2) the data have been accurately measured and correctly labeled, 3) the problem itself has been properly formulated, and 4) the solution space contains an accurate model that can be identified by the machine learning algorithm. Violations of these assumptions can arise due to bias and error introduced during the measurement, collection, and labeling of data, or when the underlying phenomena change: e.g., when a new disease spreads through the patient population. A different issue arises when the machine learning system is trained on one population (e.g., a university hospital) but then applied to a different population (e.g., a Veterans Administration hospital). A particularly extreme case of this kind of *data shift* is when the situation at hand involves categories that were not in the training data: e.g., a medical AI system that needs to recognize a disease that it has never encountered. This is known as the "open category" problem. Indeed, the classic case where the training data are fully representative of the test data is rarely encountered in practice.

A closely related threat to trust is any violation of the second two assumptions: that the problem is properly formulated and that the choice of model representation is correct. For many standard learning settings, the assumption is that each data point and query is independent of the others, which is not true if there are correlations or trends in the data. This is similar to data shift, but it is more serious, because it can require completely changing the problem formulation: e.g., considering past history, rather than just the current state, or considering other agents (humans or robots), rather than assuming the system is acting alone. To trust an AI system, we need to know what modeling assumptions it is making so that we (especially the system engineers) can check those assumptions.

A third threat to trust is adversarial attack, for which machine learning presents new challenges—e.g., modification of the training data (known as "data set poisoning"), which can cause the learning algorithm to make errors. An error in a face recognition

system, for example, might allow an adversary to gain physical access to a restricted facility. If the attacker can modify the machine learning software itself, then they can also cause the system to make mistakes. An attacker may also modify examples to which the system is later applied, seeking to gain advantage. Related to adversarial models are incentive models, where the learning systems play a role in the algorithmic economy, and where participants are profit maximizing. Consider, for example, recommender systems, where some actors may seek to promote or demote particular products or methods of dynamic pricing where it is of interest to achieve collusive outcomes or strategic advantages over other learning systems.

A fourth threat to trust is the high complexity of machine learning models, such as large ensemble classifiers (e.g., boosted ensembles or random forests) and deep neural networks. These models are too complex to be directly understood either by programmers or end users.

There are four major ways to address these challenges. If the AI system can maintain complete control of the system and its data, it can avoid most of the cyberattacks and some of the sources of bias in data collection. This does not, however, address data shift or incorrect problem formulation. Moreover, it is not clear how it is possible to maintain control of data in many settings: data inherently involves interacting with the world outside of the system.

The second strategy is the *robustness* approach. In traditional engineering, the designer of a bridge or of an airplane wing can include a margin of safety so that if the assumed load on the bridge or the wing is violated by a modest amount, the system will still work. An important research challenge in AI is to discover ways to achieve a similar effect in our algorithms. A related research direction seeks to make trained models inherently less susceptible to adversarial modifications to future examples, by promoting simpler dependencies on inputs, for example. In the context of incentive considerations, an important research direction is to model costs and utility functions explicitly, and apply game theory and mechanism design to provide robustness to strategic behavior.

The third strategy is the *detect and repair* approach. In this approach, the AI system monitors the input data to verify that the underlying assumptions are not violated and takes appropriate steps if violations occur. For example, if it detects a novel category (of disease or of animal), it can apply an algorithm for one-shot learning; if it detects a change of context, it can employ transfer learning techniques to identify the ways in which the new context is similar to the old one and transfer knowledge accordingly. An important research direction is to systematize the many ways in which the data may shift or the machine learning problem may be incorrectly formulated and develop methods for diagnosing and repairing each of those errors. The field of model diagnostics in statistics provides a good starting point for this.

The fourth strategy is to *quantify, interpret, and explain*. Here, the AI system seeks to replace complex models with models that are more easily interpreted by the user. It also seeks to quantify its uncertainty over its learned knowledge and its individual predictions so that the user (or a downstream computational component) can decide whether to trust those predictions.

We now consider several topics related to creating trustworthy machine learning systems.

Data Provenance: Information about how the data were selected, measured, and annotated is known as data provenance. Strong data provenance principles are a determining factor in the accessibility of data and govern our ability apply machine learning to that data, but training sets with good provenance information are rare. Strong data provenance promotes meaningful labeling and metadata—data about the data—across domains. We lose the ability both to leverage this data and to track changes in an accessible and concise way.

Solutions here will require understanding how data propagates across domains and over networks, how consensus is developed around annotation and structure, and where the responsibility for defining those principles lies. Improved understanding of these principles would enable us to build and evaluate robust and flexible models in changing situations. For instance, understanding the effects of climate change on building materials and indoor environments requires integrating domain knowledge from climate

scientists, materials scientists, and construction engineers. Beyond changes in statistical distribution of the data, there will be shifts in terminology associated with each of these industries over time as new materials are created and building methods are refined or improved.

Stretch goals: By 2040, many fields will have created shared, durable data repositories with strong provenance information. These will be continually maintained and tested so that the data can be trusted. Virtually all commercial machine learning and data mining systems employ provenance information to improve resilience and support updates and extensions to provenance standards. Milestones along this path include—

5 years: Provenance standards are developed in consultation with stakeholders in at least two different disciplines (e.g., pharmaceuticals and electronic health records).

10 years: Provenance standards are adopted by most scientific and engineering fields. At least two fields have shared, durable data repositories conforming to these standards.

15 years: Many fields have developed and adopted provenance standards, and these are supported by all of the leading machine learning and data mining toolkits. Most fields have shared, durable data repositories, and those repositories have well-exercised methods for making updates and extensions to the standards and testing how well machine learning and data mining systems handle updates and extensions.

Explanation and Interpretablity

An important technical challenge is to create interpretable, explainable machine learning methods. A model is interpretable if a person can inspect it and easily predict how it will respond to an input query or to changes in such a query. This relates also to *agency*, which is the idea that the user should retain enough understanding in order to retain control over decisions that are implied by a model. An explanation is some form of presentation (e.g., an argument, a sequence of inference steps, a visualization) that similarly allows a person to predict how the system will respond to changes in the query. In addition to being able to make query-specific predictions of model behavior, we often want an overall explanation. For example, we might want to know that a face recognition system *always* ignores the person's hair or that a pedestrian detection system is *never* fooled by shadows. Answering questions like this in the context of complex models such as deep neural networks is an open research topic. A primary reason for the success of deep neural networks is their ability to discover *intermediate representations* that bridge the gap between the inputs and the outputs. These intermediate representations, discussed in the previous section, rarely take on a form that a human can understand. An active research direction seeks to assist humans in understanding the functioning of neural networks, for example, through visualization.

Research on explanations must consider the users who are the targets of the explanations (the software engineer, the test engineer, the end user) and the possible queries in which those users might be interested (Why? Why not? What if? etc.). A major challenge here is that sometimes the explanation relies on regularities that the machine learning system has discovered about the world. For example, the system might explain why it predicts that company X2, which is located in Orlando, Florida, is not in the coal mining business by stating the general rule that there are no mining companies in Florida (a regularity that it has discovered). Such regularities may be obvious to a subject matter expert but not to end users.

Stretch goals: By 2040, every AI system should be able to explain its beliefs about the world and its reasoning about specific cases. The system's users (software engineers, maintenance engineers, and end users) will then be able to trust the system appropriately and understand its limitations. Milestones along this path include—

5 years: Algorithms that can learn high-performing, interpretable, and explainable models are developed for data with meaningful features.

10 years: Interpretable and/or explainable machine learning models become a requirement for all high-risk applications.

15 years: Algorithms are developed to explain their actions effectively. These allow AI systems to work much more successfully in open worlds.

Quantification of Uncertainty

In many applications, it is important for an AI system to have an internal measure of its confidence. A system that can accurately assess probabilities of various outcomes and events can assist users in taking risks and identify situations in which an automated system should cede control to a user. But what if the system is not making accurate assessments of these probabilities? An accurate understanding of the sources of uncertainty can help differentiate situations that are fundamentally uncertain from those in which additional data collection would be useful.

Quantifying uncertainty is difficult. Even in the classical setting where the examples are independent and identically distributed, many existing machine learning models can be extremely confident while making incorrect predictions. Some methods do exist for quantifying uncertainty in these cases, but there are many gaps in our knowledge. For example, if the data were collected in a biased way, or if some of the attributes are missing in some of the data, how can we quantify the resulting uncertainty? When we move beyond the classical setting to consider data shift, incorrect problem formulation, and incorrect modeling assumptions, there are essentially no existing methods for uncertainty quantification. The problems of open category detection and transfer learning are particularly vexing. Although we do have some learning methods (such as meta-learning and hierarchical modeling) for one-shot learning and transfer learning—both of which are described in the previous section—the problem of uncertainty quantification in those situations is completely unexplored.

In many cases, it is important to distinguish between situations where uncertainty stems from insufficient data or an incorrect model—situations that could be remedied by additional data collection or construction of a better model—and situations in which some aspect of the world is inherently random. A related challenge has to do with accurate comparison of uncertainties across different kinds of models: When different models disagree, how should decisions be made about gathering more data or rebuilding the model?

Stretch goals: By 2040, every AI system should be able to build and maintain a model of its own competence and uncertainty. This should include uncertainties about the problem formulation and model choice, as well as uncertainties resulting from imperfect and non-stationary data, and uncertainties arising from the behavior of other decision makers. Every AI system should exhibit some robustness to those uncertainties, and be able to detect when that robustness will break down. Milestones along this path include—

5 years: Good theoretical understanding of the different sources of data and model error is achieved so that we understand what kinds of problems can be detected and repaired. Uncertainty quantification for basic cases of data shift are developed.

10 years: Algorithms are developed to detect most forms of flawed data. These algorithms are able to adjust the learning process to compensate for these flaws, to the extent theoretically possible. Uncertainty quantification is developed for all forms of data shift and some types of model formulation errors.

15 years: Algorithms are developed to detect and adapt to many forms of model formulation error. These allow AI systems to work much more successfully in open worlds.

Machine Learning and Markets

Just as the interactions among collections of neurons lead to aggregate behavior that is intelligent at the level of whole organisms, the interactions among collections of individual decision makers leads to the phenomena of markets, which carry out tasks such as supplying all of the goods needed for a city (food, medicine, clothing, raw materials, etc.) on a daily basis, and have done so for thousands of years. Such markets work at a variety of scales and in a great variety of conditions. They can be adaptive, durable,

robust and trustable—exactly the requirements outlined in this section. Moreover, the field of microeconomics provides not only conceptual tools for understanding markets—including their limitations—but also provides a theoretical framework to guide their design, e.g., through mechanism design theory. Is there a role for economic thinking in the design of trustworthy machine learning systems, and in addressing many of the major challenges that we have identified in this report?

First, consider a multi-agent view on learning systems, a *market-of-learners* view, where open platforms promote competition between learning systems (and composition of learning systems) with profits flowing to learning systems with the best functionality. Can this market-of-learners view lead to durability, with good performance over long stretches of time, offering robustness as surrounding context changes? Markets promote complements, incentivizing new products and technologies that help when used together with existing products and technologies. What is the analogy for learning systems, in the ability to promote new data, new algorithm design, and new forms of interpretability?

Second, learning systems will necessarily act in markets, leading to a *market-based AI systems* view. In domains such as commerce, transportation, finance, medicine, and education, the appropriate level of design and analysis is not individual decisions, but large numbers of coupled decisions, and the processes by which individuals are brought into contact with other individuals and with real-world resources that must be shared. We need trustworthy machine learning systems for these economic contexts. The field of multi-agent systems studies both the design of individual agents that interact with other agents (both human and artificial) and the design of the rules or mechanisms by which multi-agent systems are mediated.

The market perspective acknowledges that there may be scarcity of resources: one cannot simply offer to each individual what they want most. If a machine learning system recommends the same route to the airport to everyone, it will create congestion, and the nominally fastest route will no longer be the fastest. Learning systems need to take into account human preferences and respect these preferences. A user may be willing to go on a slower route today because she's not in such a rush, or, on the other hand, she may be willing to pay more today precisely because she's in a rush. One AI system cannot know all of these hidden facts about participants; rather, we need the participants, or AI systems representing these participants, to understand their options, how those options help them to meet their goals, and how to understand and diagnose their role in the overall system. This is what markets do. The recommendations of market-based AI systems will, like recommendation systems, be based on data and responsive to human behavior and preferences. Such blends of microeconomic principles with statistical learning will be essential for AI systems to exhibit intelligence and adaptivity, particularly in real-world domains where there is scarcity. An economic view may also inform considerations of fairness, building from theories of taste-based discrimination and statistical discrimination in an economic context.

Of course, markets have imperfections, and new kinds of imperfections will arise as we develop new kinds of markets in the context of massive data flows and machine learning algorithms. But these imperfections are properly viewed as research challenges, not as show-stoppers. Moreover, an active research direction is to use learning systems for the automatic design of the rules of markets. Moreover, the theories of normative design from economics may prove more relevant for market-based AI systems than for systems with people, as AI system come to better respect idealized rationality than people. AI must take on this research agenda of bringing together the collective intelligence of markets with the single-agent intelligence that has been the province of much of classical AI and classical machine learning.

Stretch goals: By 2040, we understand how to build robust AI systems that can interact in market-based contexts, making decisions that are responsive to preferences, and with market-based AI systems that achieve normative goals, and more effectively than the markets that preceded them. Market-based AI systems are adaptive to global shifts in behavior, robust against perturbations or malevolent interventions, and transparent in helping individuals to understand their role in a system. We can build robust learning systems that adjust the rules of markets and mechanisms, in order to promote objectives (efficiency, fairness, etc.). We also have vibrant, open platforms that support *economy-of-learning-systems* that promote innovation, competition, and composition of capabilities. Milestones along this path include—

5 years: Market-based AI systems are able to provide explanations to users about outcomes (e.g., purchases, prices), and learning systems can model other learning systems, and reason about behavior, including the way in which behavior depends on behavior of others. Learning systems can learn the preferences of individuals and build user confidence through transparency.

10 years: We have open platforms that support economy-of-learning-systems and succeed in driving innovation in data and algorithms, together with easy composition of capabilities. Learning systems can be employed to adapt the rules of interaction and mechanisms, in response to global changes in behavior (preferences, economic conditions, etc.), which will result in robust, market-based systems. Learning systems understand the cause-and-effect between actions and outcomes in multi-agent settings.

15 years: AI systems can work successfully in open, multi-agent worlds and market-based worlds, and they are effective in representing the interests of individuals and firms. Markets become more efficient, with better decisions about scarce resources, and humans spend less time making small consumption decisions. Learning systems are robust against the uncertainties that arise from the decision making of others. Learning systems can be used to solve inverse problems, going from normative goals, including those related to fairness, to mechanisms and rules of interaction.

Durable Machine Learning Systems

Despite tremendous progress, there remain intrinsic challenges in building machine learning systems that are robust in real-world environments. Tasks that naturally have minimal training data and situations that are observed only rarely are not well handled by current methods. Furthermore, intelligent systems today often live in a transactional setting, reacting to a single input, but without long-term evolution and accommodation of new scenarios and new distributions of data. Often learned modules are treated as endpoints, yet they should be reused and composed to exhibit greater intelligence. As a result, today's machine learning systems generally have a very short life span. They are frequently retrained from scratch, often on a daily basis. They are not durable.

These challenges cross-cut many disciplines in learning and artificial intelligence, such as computer vision, robotics, natural language processing, and speech. The ultimate goal for AI has always been general artificial intelligence: pioneers in the field envisioned general-purpose learning and reasoning agents; the recent focus on specialized systems can be seen as a step along the way but is not viewed as the pinnacle for AI research.

Consider systems that are designed to work over significant stretches of time, involving data gathered at many different places and involving many individuals and stakeholders. Systems like this arise in domains such as commerce, transportation, medicine, and security. For example, the knowledge employed in the medical treatment for a single patient can be conceptualized as a vast network involving data flows and the results of experiments, studies, and past treatments conducted at diverse locations on large numbers of individuals, against the background of changing ecologies, lifestyles and climates. Bringing all this information together requires ascertaining how relevant a given datum is to a given decision, even if the context has changed (e.g., different measurement devices, populations, etc.). It requires recognizing when something significant has changed (e.g., a mutation in a flu virus, availability of a new drug). It requires coping with the inherent uncertainty in diagnostic processes, and it requires managing biases that may creep into the way experiments are run and data are selected. It requires being robust to errors and to adversarial attack. Finally, while we wish for a certain degree of autonomy from the system—it should learn from data and not require detailed human programming or management—we must also recognize that for the foreseeable future human judgment and oversight will be required throughout the lifetime of any complex, real-world system.

In this section, we identify some of the specific research challenges associated with building durable machine learning systems, progressing toward a world of more and more general artificial intelligence.

Dealing with Data Shift

The knowledge acquired through machine learning is only valid as long as the regularities discovered in the training data hold true in the real world. However, the world is continually changing, and we need AI systems that can work for significant stretches of time without manual re-engineering. This is known as the problem of data shift.

It is generally acknowledged that current state-of-the-art machine learning, particularly in supervised settings, is brittle and lacks a dynamic understanding of the way data evolves over time and in response to changing environments or interactions. We need improved methods for detecting data shift, diagnosing the nature of the shift, and repairing the learned model to respond to the shift.

Detecting data shift need not be a purely statistical problem. Strong provenance information can directly tell the AI system how the data have changed and provide valuable information for responding to the shift. For example, suppose a sensor has been upgraded and the new sensor performs internal normalization. Provenance information can suggest that it is reasonable to learn how to transform the new normalized values into the old value scale so that the learned model does not need to be changed.

To build more flexible systems over time and in the event of an abrupt shift, we must be able to rapidly respond when a shift is detected and develop a resolution. Each system must ask whether it has the ability to re-train in a timely fashion and if it has easily accessible and available data to build a new model. This becomes even more important when handling real-time systems where the scale of data may make re-training prohibitively expensive. In this case, we must ensure rapid model switching and evaluation. Diagnostics, in this case, would act proactively to evaluate existing models and train parallel models that can account for whether data is anomalous and develop replacement models with varying distributions that examine ways to effectively assess new/online or unstructured data.

Machine Learning and Memory

The ability to create, store, and retrieve memories is of fundamental importance to the development of intelligent behavior. Just as humans depend on memory, advanced machine learning systems must be able to leverage the same kinds of capabilities. Humans also make use of external, collective memories to supplement their own, for example, through the use of libraries or, more recently, the World Wide Web. Machine learning systems must do the same through the use of external memory structures and knowledge bases.

Deep learning has become the de facto tool for analysis of many types of sequential data, such as text data for machine translation and audio signals for speech recognition. Beyond these marquee applications, deep learning is also being deployed in a broad range of areas including analyzing product demand, weather forecasting, and analyzing biological processes. Other critical areas involve systems for perception and planning in dynamic environments. Deep learning methods provide significant promise in providing flexible representational power of complex sequential data sources. However, in these areas, there are three fundamental challenges: 1) forming long-range predictions, 2) adapting to potentially rapidly changing and non-stationary environments, and 3) persisting memory of relevant past events.

Deep learning models have the potential to capture long-term dependencies; these dependencies can be critical for the task at hand, such as suggesting routes in a traffic system. However, a question is how to flexibly adapt to new settings. For example, a traffic system has to rapidly adapt to road closures and accidents. Unfortunately, most existing deep learning training procedures require batch (offline) training, and so they are unable to adapt to rapidly changing environments while persisting relevant memory. The common approach of dividing the training data into chunks (or batches) and learning a model initialized from the previous chunk (warm start) leads to catastrophic forgetting, where the model parameters adapt too quickly and overfit to the new chunk of training data. For example, when the road opens again, there is no memory from previous chunks when the road

was open, so the system needs to learn about open (normally functioning) roads from scratch. Current methods for alleviating catastrophic forgetting only address one-shot classification tasks and do not address handling the temporal dependencies in sequence models. Note that there is a push-and-pull between rapid adaptation and persisting relevant information from the past, especially when one returns to previously seen settings. A related challenge is one of domain adaptation. For example, suppose the road closure was due to construction leading to a short new road segment, but an overall similar road network.

One potential avenue is to leverage new modeling techniques instead of modifying training methods. Attention mechanisms, for example, have shown promise in identifying contexts relevant for making predictions. In traffic forecasting, for example, the predicted congestion at an intersection depends on flow into and out of that intersection as well as congestion at similar types of intersections; attention can learn such relationships. As another example, memory mechanisms allow deep learning methods to store important past contexts, which can help alleviate the problem of forgetting past learning. For example, the traffic system could leverage memory to store the state of traffic dynamics prior to a road closure and recall them when the road reopens. Although attention and memory can highlight past states relevant to predicting long-term outcomes, there are many opportunities for improving long-term predictions in sequential models by learning seasonal patterns and long-term trends.

Turning to another domain where memory is critical to the next stage of advances in machine learning, the building of conversational agents has been a long-standing goal in natural language processing. Most current agents are little more than chatbots, however. Progress toward more natural, complex, and personalized utterances will hinge on agents having a large amount of knowledge about the world, how it works, and its implications for the ways in which the agent can assist the user. It will also require the agent to be able to build a personalized model of the user's goals and emotional state. Such a model will need long-term memory and lifelong learning, since it must evolve with each interaction and remember and surface information from days, weeks, or even years ago. Another related technical challenge is collecting and defining datasets and evaluation methodologies. Unlike domains such as image classification where the input can be treated as an isolated object, dialog is an ongoing interaction between the system and a user, making static datasets and discrete label sets useless. How to create good testbeds (and collect good datasets) and how to evaluate such interactive systems is an open research question.

Machine learning systems must also be designed to exploit the wealth of knowledge that has been accumulated through the ages. As we have previously noted, it should be possible to leverage the domain knowledge developed in scientific disciplines (e.g., basic principles such as conservation of mass and energy, mechanistic/process-based models) in machine learning frameworks to deal with data-sparse situations and to make machine learning algorithms generalizable to a greater range of data distributions. In the absence of adequate information about the physical mechanisms of real-world processes, machine learning approaches are prone to false discoveries and can also exhibit serious inconsistencies with known physics (the wealth of knowledge accumulated through the ages). This is because scientific problems often involve complex spaces of hypotheses with non-stationary relationships among the variables that are difficult to capture solely from raw data. Leveraging physics will be key to constraining hypothesis spaces in machine learning for small sample regimes and to produce models that can generalize to unseen data distributions.

In the natural language domain, existing recurrent neural network methods can be applied to compose word vectors into sentence and document vectors. However, this loses the structure of the text and does not capture the logical semantics of the language. By combining symbolic representations of linguistic meaning such as semantic networks, knowledge graphs, and logical forms, with continuous word embeddings, a hybrid method could integrate the strengths of both approaches. New representational formalisms and learning algorithms are needed to allow an AI system to construct such hybrid representations, reason with them efficiently, and learn to properly compose them. Such techniques are another way to exploit memory and accumulated knowledge in machine learning systems and could lead to improved document understanding, question answering, and machine translation.

Overall, new advances are necessary to form well-calibrated long-term predictions and persist relevant information from the past while performing continual, life-long learning.

Transfer Learning

A notion related to memory is the ability to transfer or adapt hard-earned knowledge from one setting to a different setting. Many of the most effective machine learning methods today require very large amounts of training data, which entails significant investments of time and expense in their development. However, some tasks seem so similar in their inputs and intended outputs that it ought to be possible to leverage the investments in building one system to reduce the development time for another. Imagine a classifier trained to recognize dogs in an image then being adapted to recognize cats.

Transfer learning and meta-learning (or learning to learn) have received much attention as approaches to transfer what has been learned in a source domain (e.g., with large amounts of data) to a destination domain where there is little available data. Computer vision and natural language processing are two popular application areas for transfer learning. We might ask, for example, whether classifiers trained on standard image datasets can be applied in medical imaging scenarios where much less data is available. Transfer learning is particularly valuable in the case of deep learning systems, since they require significant resources to build. Its success is based on the notion that at some layer of the network, the features learned for one problem can be effectively adapted for another problem without having to rebuild the network from scratch.

Transfer learning can be divided into *forward* transfer, which learns from one task and applies to another task, and *multi-task* transfer, which learns from a number of tasks and applies to a new task. Diversity in the training data is the key to success, but defining and capturing an effective notion of diversity remains an open research challenge. Transfer learning is often discussed in the context of few-shot, one-shot, and zero-shot learning, addressed in the next section. Meta-learning, on the other hand, is the concept of learning how to solve a new problem from previously accomplished tasks, not just from labeled training data as with traditional machine learning. Ideally, a meta-learner can acquire useful features for solving a new problem more efficiently and learn to avoid exploring alternatives that are clearly not productive. While offering intriguing possibilities, meta-learning approaches to date are sensitive to the initial task distributions they are presented with; designing effective task distributions is an open research problem.

While transfer learning has demonstrated some promising potential, more research is needed to continue to quantify its inherent uncertainty. The path forward will require innovation: clearly we should strive to be fully confident, but if there is transfer to be had from a large data source, we should be more confident than if we only had a smaller dataset to learn from.

Learning from Very Small Amounts of Data

As AI and deep learning models are applied to more domains, we are confronted with more and more learning scenarios that have a very limited number of labeled training examples. Furthermore, there can be continual changes in data distributions (e.g., dynamic systems) resulting in insufficient examples to retrain the model. In other scenarios, it is infeasible to obtain examples in some low-resource applications (e.g., rare diseases in health care domains). Therefore, there is a pressing need to develop learning methods that can learn effectively with fewer labels.

One promising solution is few-shot (and zero-shot) learning. These approaches typically work by first learning about various sub-parts of objects and their attributes. Then when a new object is observed for the first time, its sub-parts and attributes are compared to known sub-parts and attributes, and a recognizer for the new class is synthesized. For example, consider a computer vision system for recognizing birds. By studying multiple bird species, it can discover that important properties include the shape and length of the beak, color of the back, color of the breast, and so on. It can learn that an American robin has a black back, an orange breast, and a short light-orange beak. The first time it is shown a wood thrush, it can infer that it is similar to a robin but its back and beak are reddish-brown and its breast is white with black spots.

While many few-shot classification algorithms have achieved successes in some learning scenarios, there is still major room for improvement. First, the effectiveness of existing few-shot learning models varies. Second, it is usually assumed that the training samples (with limited examples) and the testing samples have the same distributions; while in practice there could be data shift.

Another important path to addressing low-resource learning is incorporation of domain knowledge or physical models with data-driven models (as discussed above). In many dynamical systems, such as energy systems, traffic flows, turbulent flow, aerospace and climate models, pure data-driven approaches may lead to incorrect and uninformed decisions due to limited data compared with vast state space. Models based on physical principles and domain knowledge have been developed with some initial success, although mathematical models based on these physical principles may have flaws and omit important factors. An important goal is to achieve highly accurate models by integrating physical principles with data. Research in this area will develop successful methodologies and easy-to-use tools for solving a wide range of problems in science, engineering, medicine, and manufacturing.

Stretch goals: Building durable machine learning systems requires developing sound principles for coping with the challenges discussed earlier in this section. By 2040, AI systems must deal with data shift, realizing that the real world is constantly changing and recognizing when the data they were trained with is no longer representative. They must find ways to exploit memory—knowledge that they or others have previously acquired. They must be able to adapt from problems they know how to solve to problems that are new through transfer learning. And they must be able to learn effectively when confronted by new and novel situations they have not encountered before. Milestones along this path include—

5 years: AI systems that are able to identify and correct for data shift in focused domains (e.g., a specific kind of medical diagnostic). Continued melding of data-based learning and first-principles for tasks with physical interpretations. Development of more effective paradigms for transfer learning, including increased ability to characterize task sets and their mappings into new domains. Better, more effective schemes for few-shot and zero-shot learning.

10 years: AI systems that begin to integrate durability along two or more of the dimensions described above, at the same time.

20 years: Truly durable AI systems that can simultaneously recognize and correct for data shift, exploit past knowledge and first principles in a manner reminiscent of human experts, transfer hard-earned knowledge from known problems to another related problem, and identify and perform competently when confronted by completely new tasks.

INTEGRATING AI AND ROBOTIC SYSTEMS

Learning From, and For, Physical Interaction

Robots must interact with the physical world. More so, robots can be seen as a tool for automating our physical world, similar to how digital computing has automated our organization of and access to information. Through autonomous robotics, society will be able to envision and operationalize new ways to make our physical world programmable. The challenge we face, however, is that the interface of autonomous systems and the real world is filled with uncertainty and physical constraints. The convergence of robotics and AI is well suited to address these challenges and to yield next-generation autonomous systems that are robust in the context of changing real-world situations.

Robots, in order to move, must rely on physical contacts between surfaces and wheels or feet, or interactions between fluids (water, air) and wings and propellers. Locomotion often involves repeated, cyclical motions in which energy may be stored and released. Often the energy is stored mechanically, for example in springs. Currently, the mechanical and actuation components of a robot are designed first by one team; another team then develops the controllers and algorithms. Neither of these teams is able to fully model and understand the uncertainty imposed on the entire system, which limits the adaptability and robustness of the design. An exciting challenge for future work in locomotion is to replace this process with a simultaneous, collaborative design of both the hardware and the software.

Robot dexterity for grasping and manipulating objects remains a grand challenge for robotics research and development. These are crucial capabilities for robots to be useful in flexible manufacturing, logistics, and health and home-care settings, among many others. Manipulators that are currently deployed in industrial settings are designed to perform highly repetitive and exactly specified tasks. In contrast, manipulators operating in unstructured environments, such as homes, hospitals, or small-scale

manufacturing sites, must handle a variety of objects, including fluids and non-rigid objects (e.g., cables and cloth). Operating in such environments requires the ability to recognize novel objects, assess their physical properties and affordances, and combine real-time feedback from visual, force, and touch sensors to constantly update those assessments and use them to generate robust controls.

While the robotics research community has developed techniques that enable manipulators to pick up rigid objects or cloth in research lab settings, fundamental advances in perception, control, and learning are necessary if these skills are to transfer to complex tasks in unconstrained environments. Consider, for example, grasping and sliding a cable through a small hole or mixing flour into cake batter, while being robust to details such as changing lighting conditions, cable stiffness, unpredictable tools and containers, and batter viscosity. Representations that capture the salient knowledge for these kinds of tasks are too complex to be manually designed and need to be learned from both real-world and simulated experiences. While recent advances have enabled progress in isolated manipulation settings, future manipulators must be able to learn far more capable representations and adapt them on the spot, without needing extensive offline learning or manual data labeling. This will require compression of high-dimensional sensor data into representations that are useful for a robot's task and allow for effective communication and collaboration with humans.

Stretch goals: By 2040, every robotic system should be able to safely and dexterously manipulate objects in common everyday human environments, including physically assisting people. Every system should also be continually updating models of how to perform various manipulation tasks and behaviors, helping itself and other robots to understand the form and function of objects and the behaviors of people in order to perform ever-safer and more effective physical manipulation, collaboration, and assistance. Every robot will be able to use AI systems to reason over affordances to realize goals and perform effectively in complex human environments. Every robot will have access, through the cloud, to a large repository of data, knowledge, and plans to facilitate real-time understanding of the form and function of the vast range of objects and of human behavior in the physical world. Newly created robots will be able to adapt and use these cloud repositories immediately, or through simulation-based training. Milestones along this path include—

5 years: Learning of representations that combine multi-sensor data to support robust operation in cluttered scenes. Learning from explicit instruction and demonstration by non-experts in day-to-day environments, one-on-one and group settings.

10 years: Learning of representations that support a broad set of manipulation tasks, including handling of novel rigid and deformable objects and liquids. Learning from observations of real-world environments and interactions. Safe physical manipulation of objects around people, such as in eldercare settings and homes.

20 years: Learning of representations that support dexterous manipulation of arbitrary objects in unstructured environments. Systems deployed in real-world human environment engaging in lifelong learning and continuous adaptation to the changing dynamics of the environment, including people. Safe physically assistive manipulation of people, such as in care for the elderly.

Learning From Humans

As the applications of AI and robotic systems expand, robots will increasingly be introduced to human environments, including homes, workplaces, institutions, and public venues. In these settings, they will interact and collaborate with people, which will require understanding the requests, actions, and intents of those people, and require acting in ways that do not violate both explicit and explicit human norms and expectations, including ethical standards. Such capabilities will demand designs that incorporate fundamental capabilities for interaction, collaboration, and human awareness and that can adapt to different environments and users. These capabilities are extremely complex, and information needed to achieve the necessary adaptations is often unavailable at design time, necessitating the use of learning methods.

Humans can facilitate robot learning in a number of ways, including providing explicit instruction and demonstrations through multimodal input, structuring the environment in ways that facilitate the robot but do not inconvenience people, and giving

feedback to the robot as it explores a new environment to adapt its capabilities. These approaches require rich perceptual information from humans in relation to their environment, such as understanding verbal commands, including spatial references, detecting entities that are referenced through pointing or gazing, continuously tracking objects that are manipulated by the human, and inferring intent or missing information that is not directly and explicitly communicated. Making sense of this information and acting appropriately in response requires advances in sensing, recognition, prediction, and reasoning. Robots will need to not only construct models of human behavior and plan their own actions accordingly, but they must also communicate their own goals, intent, and intended actions to users in natural and informative ways.

Robots deployed in mission-critical settings (e.g., search and rescue, law enforcement, space exploration), in physical and social assistance (e.g., therapy, rehabilitation, home healthcare), and as expert tools (e.g., robot-assisted surgery) especially need to have robust and adaptive models of human input, states, and goals.

Stretch goals: By 2040, robotic systems will obtain new physical and social capabilities through increasingly more sophisticated ways of learning and applying learned capabilities appropriately, depending on social context and user preferences, to improve human trust and rapport. Milestones along this path include—

5 years: Learning from explicit instruction and demonstration by non-experts in day-to-day environments, one-on-one and group settings.

10 years: Learning from observations of implicit human behavior and environmental changes, adaptively engaging in active learning to improve model robustness.

20 years: Systems deployed in real-world human environment engaging in lifelong learning and continuous adaptation to the changing dynamics of the environment. Systems also execute learned capabilities in appropriate ways to improve human trust.

Infrastructure and Middleware for Capable Learning Robots

Effective robot autonomy can be thought of as following a sense-plan-act pipeline involving many AI components that must work in synchrony: a computer vision system that senses the state of the world from the robot sensors, an automated planner that chooses actions to achieve a goal state, and a motion controller that carries out those planned actions. Machine learning plays important roles throughout this pipeline, drawing inferences from data to provide functionality where hand-coded solutions are infeasible.

The Robot Operating System (ROS) provides a common framework for developing AI software for robots, allowing heterogeneous communities of researchers, companies, and open-source developers to build on each other's work. Since the original robotics Roadmap in 2009, this infrastructure has served as a foundation for the rapid advancement of robotics research and development. However, these initial robot middleware efforts were geared specifically for robotics researchers. More general open-source middleware platforms, with broader functionality, are needed for prototyping new AI systems with direct uses in real-world settings involving sensing and action in the real world. A significant challenge here will be to fully understand the sense-plan-act pipeline and its interplay with robot components (software, hardware, communication) from disparate, and possibly malicious, sources. Solutions will require generalizable standards and testing suites for widely varying robot systems.

Stretch goals: By 2040, highly capable AI systems will be employed broadly across robotics domains, to the point that such use becomes commonplace. Just as today we take for granted that any mobile robot can map, localize, and navigate in a new environment, we will come to expect that all future robots will have the new capabilities that will be provided by forthcoming advances in AI. Reaching this level will require that AI applications can be used portably, safely, and securely in a manner that earns the trust of users. Milestones along this path include—

5 years: Widespread deployment of next-generation middleware for AI and robotics and accompanying testing standards for real-world conditions.

10 years: Industry adoption of robust security and safety standards for AI applications built on common middleware.

20 years: Rich ecosystem of user-centered and privacy-preserving AI applications that are portable across different robot platforms.

4. Major Findings

Several major findings from community discussions drive the recommendations in this Roadmap:

I. Enabled by strong algorithmic foundations and propelled by the data and computational resources that have become available over the past decade, AI is poised to have profound positive impacts on society and the economy. AI has become a mature science, leveraging large datasets and powerful computing resources to produce substantial progress in many areas: exploration and training of statistical models, for instance, and powerful image and video-processing techniques. Many other areas of AI might be amenable to the same dramatic leaps forward, but are starving for appropriate data. And while collection, processing, and annotation of data are key aspects of an experimental science, architectures and frameworks are also instrumental in AI solutions. Enabled by the right theoretical and applied foundations and fueled by massive datasets and growing computational power, future AI-driven successes could affect many aspects of society, including healthcare, education, business, science, government, and security (as shown in Figure 3): removing humans from harm's way in dangerous yet vital occupations; aiding in the response to public health crises and natural disasters; expanding educational opportunities for increasingly larger segments of our society; helping local, state, and federal government agencies deliver broader range of valuable services to their citizens; or providing personalized lifelong health- and wellness care accessible to each individual's changing needs.

II. To realize the potential benefits of AI advances will require audacious AI research, along with new strategies, research models, and types of organizations for catalyzing and supporting it. Over its first few decades, AI research was characterized by steady progress in understanding and recreating intelligent behaviors, with deployments of AI-enhanced systems in narrow application areas. Since the mid-1980s, fundamental AI research has been supported largely by short-term grant-funded projects in small single-investigator labs, limiting the types of empirical work and advances possible. In recent years, the experimental possibilities enabled by data and computationally resource-rich and generously staffed industry labs have yielded significant advances, enabling wide deployment of AI-enabled systems in many societally relevant and important venues. Cross-fertilization with other fields, ranging from social science to computer architecture, is also critical to modern AI, given the demands, breadth, and implications of its applications. To adequately address the next generation of AI challenges will require sustained effort by large, interdisciplinary teams supported by appropriate resources: massive datasets, common architectures and frameworks, shared hardware and software infrastructure, support staff, and sustained, long-term funding. This Roadmap offers new models for the resources, the critical mass, and the long-term stability that will be needed in order to enable a new era of audacious AI research that is significantly more integrative and experimental while also recognizing the need for caution regarding the impact of AI in society.

III. The needs and roles of academia and industry, and their interactions, have critically important implications for the future of AI. Building on the foundations of past AI research, most of which was conducted in academia, the private sector has compiled and leveraged massive resources—datasets, knowledge graphs, special-purpose computers, and large cadres of AI engineers—to propel powerful innovations. These assets, which provide major competitive advantages, are generally proprietary. Furthermore, the constraints, incentives, and timelines in these two communities are very different: Industry is largely driven by practical, near-term solutions, while academia is where many of the fundamental long-term questions are studied. Solutions to