

DOI:10.1145/3271625

What just happened in artificial intelligence and how it is being misunderstood.

BY ADNAN DARWICHE

Human-Level Intelligence or Animal-Like Abilities?

“The vision systems of the eagle and the snake outperform everything that we can make in the laboratory, but snakes and eagles cannot build an eyeglass or a telescope or a microscope.”
— Judea Pearl^a

THE RECENT SUCCESSES of neural networks in applications like speech recognition, vision, and autonomous navigation has led to great excitement by members of the artificial intelligence (AI) community, as well as by the general public. Over a relatively short time, by the science clock, we managed to automate some tasks that have defied us for decades, using one of the more classical techniques due to AI research.

^a Lecture by Judea Pearl, *The Mathematics of Causal Inference, with Reflections on Machine Learning and the Logic of Science*; <https://www.youtube.com/watch?v=zHjdd-W6o4>

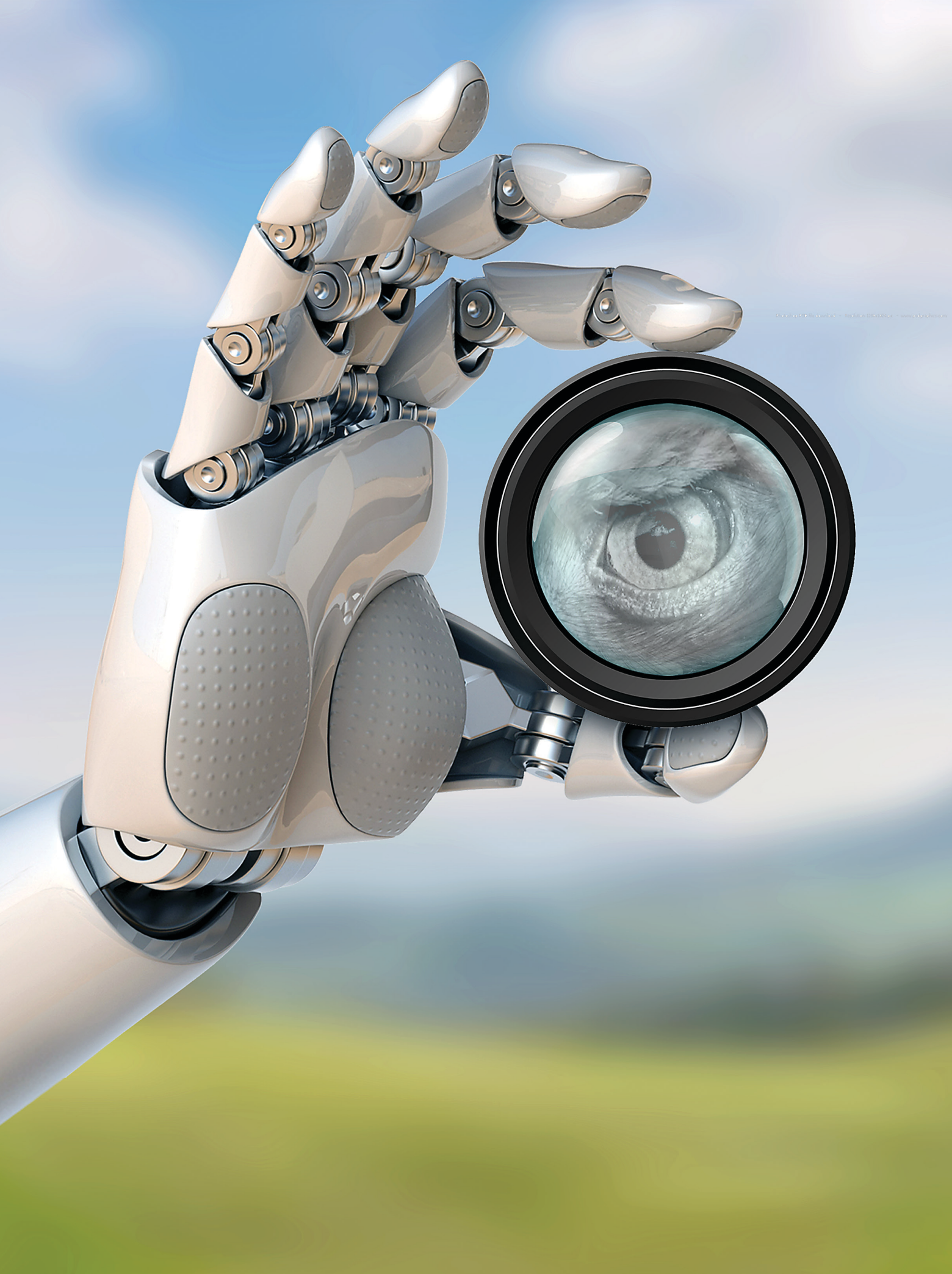
The triumph of these achievements has led some to describe the automation of these tasks as having reached human-level intelligence. This perception, originally hinted at in academic circles, has gained momentum more broadly and is leading to some implications. For example, some coverage of AI in public arenas, particularly comments made by several notable figures, has led to mixing this excitement with fear of what AI might bring us all in the future (doomsday scenarios).^b Moreover, a trend is emerging in which machine learning research is being streamlined into neural network research, under its newly acquired label “deep learning.” This perception has also caused some to question the wisdom of continuing to invest in other machine learning approaches or even other mainstream areas of AI (such as knowledge representation, symbolic reasoning, and planning).

This turn of events in the history of AI has created a dilemma for researchers in the broader AI community. On the one hand, one cannot but be impressed with, and enjoy, what we have been able to accomplish with neural networks. On the other hand, mainstream scientific intuition stands in the way of accepting that a method

^b Stephen Hawking said: “The development of full artificial intelligence could spell the end of the human race;” and Elon Musk said AI is: “... potentially more dangerous than nukes.”

» key insights

- The recent successes of deep learning have revealed something very interesting about the structure of our world, yet this seems to be the least pursued and talked about topic today.
- In AI, the key question today is not whether we should use model-based or function-based approaches but how to integrate and fuse them so we can realize their collective benefits.
- We need a new generation of AI researchers who are well versed in and appreciate classical AI, machine learning, and computer science more broadly while also being informed about AI history.




that does not require explicit modeling or sophisticated reasoning is sufficient for reproducing human-level intelligence. This dilemma is further amplified by the observation that recent developments did not culminate in a clearly characterized and profound scientific discovery (such as a new theory of the mind) that would normally mandate massive updates to the AI curricula. Scholars from outside AI and computer science often sense this dilemma, as they complain they are not receiving an intellectually satisfying answer to the question: “What just happened in AI?”

The answer lies in a careful assessment of what we managed to achieve with deep learning and in identifying and appreciating the key scientific outcomes of recent developments in this area of research. This has unfortunately been lacking to a great extent. My aim here is to trigger such a discussion, encouraged by the positive and curious feedback I have been receiving on the thoughts expressed in this article.


Background

To lay the ground for the discussion, I first mark two distinct approaches for tackling problems that have been of interest to AI. I call the first one “model-based” and the second “function-based.” Consider the object-recognition and -localization task in Figure 1. To solve it, the model-based approach requires one to represent knowledge about dogs and hats, among other things, and involves reasoning with such knowledge. The main tools of the approach today are logic and probability (mathematical modeling more generally) and can be thought of as the “represent-and-reason”^c approach originally envisioned by the founders of AI. It is also the approach normally expected, at some level, by informed members of the scientific community. The function-based approach, on the other hand, formulates this task as a function-fitting problem, with function inputs coming directly from the image pixels and outputs corresponding to the high-level recognitions we seek. The function must have a form that can be evaluated efficiently so no

c This term might be likened to what has been called “good old-fashioned AI.”



In my own quest to fully appreciate the progress enabled by deep learning, I came to the conclusion that recent developments tell us more about the problems tackled and the structure of our world than about neural networks per se.



reasoning is required to compute the function outputs from its inputs. The main tool of this approach is the neural network. Many college students have exercised a version of it in a physics or chemistry lab, where they fit simple functions to data collected from various experiments, as in Figure 2. The main difference here is we are now employing functions with multiple inputs and outputs; the structure of these functions can be quite complex; and the problems being tackled are ones we tend to associate with perception or cognition, as opposed to, say, estimating the relationship between volume and pressure in a sealed container.^d

The main observation in AI recently is that the function-based approach can be quite effective at certain AI tasks, more so than the model-based approach or at least earlier attempts at using this approach. This has surprised not only mainstream AI researchers, who mainly practice the model-based approach, but also machine learning researchers who practice various approaches, of which the function-based approach is but one.^e This has had many implications, some positive and some giving grounds for concern.

On the positive side is the increasing number of tasks and applications now within reach, using a tool that can be very familiar to someone with only a broad engineering background, particularly one accustomed to estimating functions and using them to make predictions. What is of concern, however, is the current imbalance between exploiting, enjoying, and cheering this tool on the one hand and *thinking* about it on the other. This thinking is not only important for realizing the full potential of the tool but also for scientifically characterizing its potential

d This is also called the “curve-fitting” approach. While the term “curve” highlights the efficient evaluation of a function and captures the spirit of the function-based approach, it underplays the complex and rich structure of functions encoded by today’s (deep) neural networks, which can have millions if not billions of parameters.

e Machine learning includes the function-based approach but has a wide enough span that it overlaps with the model-based approach; for example, one can learn the parameters and structure of a model but may still need non-trivial reasoning to obtain answers from the learned model.

reach. The lack of such characterization is a culprit of current misconceptions about AI progress and where it may lead us in the future.

What Just Happened in AI?

In my own quest to fully appreciate the progress enabled by deep learning, I came to the conclusion that recent developments tell us more about the problems tackled and the structure of our world than about neural networks per se. These networks are parameterized functions that are expressive enough to capture any relationship between inputs and outputs and have a form that can be evaluated efficiently. This has been known for decades and described at length in textbooks. What caused the current turn of events?

To shed some light on this question, let me state again what we have discovered recently. That is, some seemingly complex abilities that are typically associated with perception or cognition can be captured and reproduced to a reasonable extent by simply fitting functions to data, without having to explicitly model the environment or symbolically reason about it. While this is a remarkable finding, it highlights problems and thresholds more than it highlights technology, a point I explain next.

Every behavior, intelligent or not, can be captured by a function that maps inputs (environmental sensing) to outputs (thoughts or actions). However, the size of this function can be quite large for certain tasks, assuming the function can be evaluated efficiently. In fact, the function may have an unbounded size in general, as it may have to map from life histories. The two key questions then are the following: For tasks of interest, are the corresponding functions simple enough to admit a compact representation that allows mapping inputs to outputs efficiently, as in neural networks (without the need for reasoning)? And, if the answer is yes, are we currently able to estimate these functions from input-output pairs (labeled data)?

What has happened in AI recently are three developments that bear directly on these questions: The first is our improved ability to fit functions to data, which has been enabled by the availability of massive amounts

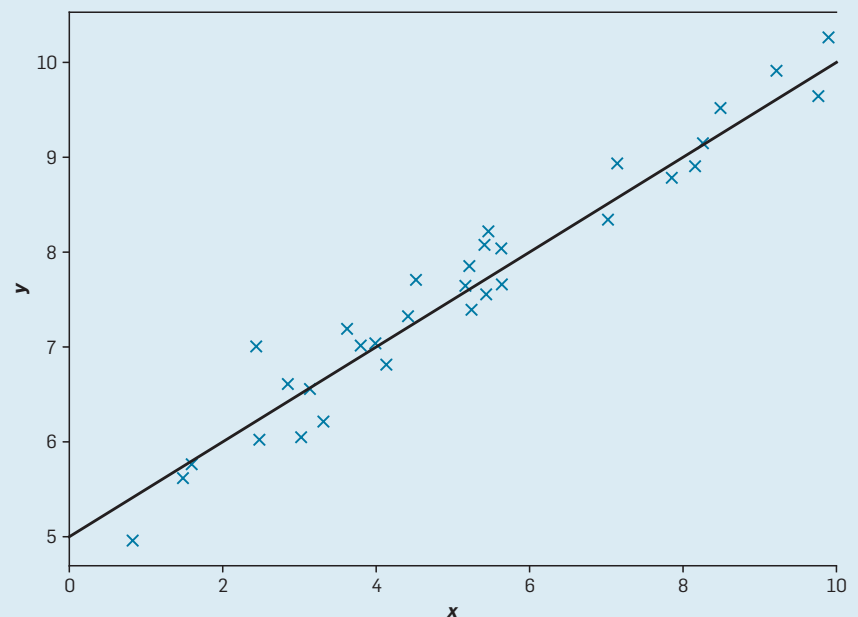
of labeled data; the increased computational power we now have at our hands; and the increasingly sophisticated statistical and optimization techniques for fitting functions (including new activation functions and new/deeper network structures). The second is that we have identified

a class of practical applications that correspond to functions that, we now know, are simple enough to allow compact representations that can be evaluated efficiently (again, without the need for reasoning), and whose estimation is within reach of current thresholds for gathering data, com-

Figure 1. Object recognition and localization in an image (ImageNet).



Figure 2. Fitting a simple function to data.



putational speed, and estimation techniques. This includes recognizing and localizing objects in some classes of images and certain tasks that pertain to natural language and speech. The third development, which goes largely unnoticed, is that we gradually changed our objectives and measures for success in ways that reduced the technical challenges considerably, at least as entertained by early AI researchers, while maintaining our ability to capitalize on the obtained results commercially, a point I discuss further later in the section on objectives and success.

Interestingly, none of these developments amounts to a major technical breakthrough in AI per se (such as the establishment of probability as a foundation of commonsense reasoning in the late 1980s and the introduction of neural networks more than 50 years ago).^f Yet the combination of these factors created a milestone in AI history, as it had a profound impact on real-world applications and the successful deployment of various AI techniques that have been in the works for a very long time, particularly neural networks.^g

'I Beg to Differ'

I shared these remarks in various contexts during the course of preparing this article. The audiences ranged from AI and computer science to law and public-policy researchers with an interest in AI. What I found striking is the great interest in this discussion and the com-

fort, if not general agreement, with the remarks I made. I did get a few "I beg to differ" responses though, all centering on recent advancements relating to optimizing functions, which are key to the successful training of neural networks (such as results on stochastic gradient descent, dropouts, and new activation functions). The objections stemmed from not having named them as breakthroughs (in AI). My answer: They all fall under the enabler I outlined earlier: "increasingly sophisticated statistical and optimization techniques for fitting functions." Follow up question: Does it matter that they are statistical and optimization techniques, as opposed to classical AI techniques? Answer: It does not matter as far as acknowledging and appreciating scientific inquiry and progress, but it does matter as far as explaining what just happened and, more important, forecasting what may happen next.

Consider an educated individual sitting next to you, the AI researcher, on a plane; I get that a lot. They figure out you do AI research and ask: What are the developments that enabled the current progress in AI? You recount the function-based story and lay out the three enablers. They will likely be impressed and also intellectually satisfied. However, if the answer is, "We just discovered a new theory of the mind," you will likely not be surprised if they also end up worrying about a Skynet coming soon to mess up our lives. Public perceptions about AI progress and its future are very important. The current misperceptions and associated fears are being nurtured by the absence of scientific, precise, and bold perspectives on what just happened, leaving much to the imagination.

This is not to suggest that only a new theory of the mind or an advance of such scale would justify some of the legitimate concerns surrounding AI. In fact, even limited AI technologies can lead to autonomous systems that may pose all kinds of risks. However, these concerns are not new to our industrialized society; recall safety concerns when the autopilot was introduced into the aerospace industry and job-loss concerns when ATMs were introduced into the banking industry. The headline here should therefore be "automation" more than "AI," as the latter is just a technology that happened to improve and

speed up automation.^h To address these concerns, the focus should be shifted toward policy and regulatory considerations for dealing with the new level of automation our society is embarking on, instead of fearing AI.

On Objectives and Success

Let me now address the third reason for the current turn of events, which relates to the change in objectives and how we measure success as a broad AI community. This reason is quite substantial yet goes largely unnoticed, especially by younger researchers. I am referring here to the gradual but sustained shift over AI history from trying to develop technologies that were meant to be intelligent and part of integrated AI systems to developing technologies that perform well and are integrated with consumer products; this distinction can be likened to what has been called "Strong AI" vs. "Weak AI."

This shift was paralleled by a sharpening of performance metrics and by progress against these metrics, particularly by deep learning, leading to an increased deployment of AI systems. However, these metrics and corresponding progress did not necessarily align with improving intelligence, or furthering our understanding of intelligence as sought by early AI researchers.ⁱ One must thus be careful not to draw certain conclusions based on current progress, which would be justified only if one were to make progress against earlier objectives. This caution particularly refers to current perceptions that we may have made considerable progress toward achieving "full AI."

Consider machine translation, which received significant attention in the early days of AI. The represent-and-reason approach aimed to comprehend text before translating it and is considered to have failed on this task, with function-based approaches being the state of the art today. In the early days of AI, success was measured by how far a system's accuracy was

^f Research on neural networks has gone through many turns since their early traces in the 1940s. Nils Nilsson of Stanford University told me he does not think the pessimistic predictions of the 1969 book *Perceptrons: An Introduction to Computational Geometry* by Marvin Minsky and Seymour Papert was the real reason for the decline in neural network research back then, as is widely believed. Instead, it was the inability to train multiple layers of weights that Nilsson also wrestled with at SRI during that time "but couldn't get anywhere," as he explained to me.

^g A perspective relayed to me by an anonymous reviewer is that science advances because instruments improve and that recent developments in neural networks could be viewed as improvements to our machine learning instruments. The analogy given here was to genomics and the development of high-throughput sequencing, which was not the result of a scientific breakthrough but rather of intense engineering efforts, yet such efforts have indeed revealed a vast amount about the human genome.

^h See also the first report of the *One Hundred Year Study on Artificial Intelligence (AI100)* for a complementary perspective; <https://ai100.stanford.edu/>


ⁱ An anonymous reviewer said that throughout AI there are metrics for evaluating task performance but not for evaluating the fit among an agent, its goals, and its environment. Such global metrics may be needed to assess and improve the intelligence of AI systems.

from 100% compared to humans, and successful translation was predicated on the ability to comprehend text. Government intelligence was a main driving application; a failure to translate correctly can potentially lead to a political crisis. Today, the main application of machine translation is to webpages and social-media content, leading to a new mode of operation and a different measure of success. In the new context, there is no explicit need for a translation system to comprehend text, only to perform well based on the adopted metrics. From a consumer's viewpoint, success is effectively measured in terms of how far a system's accuracy is from 0%. If I am looking at a page written in French, a language I do not speak, I am happy with any translation that gives me a sense of what the page is saying. In fact, the machine-translation community rightfully calls this "gist translation." It can work impressively well on prototypical sentences that appear often in the data (such as in social media) but can fail badly on novel text (such as poetry). It is still very valuable yet corresponds to a task that is significantly different from what was tackled by early AI researchers. We did indeed make significant progress recently with function-based translation, thanks to deep learning. But this progress has not been directed toward the classical challenge of comprehending text, which aimed to acquire knowledge from text to enable reasoning about its content,^j instead of just translating it.^k


Similar observations can be made about speech-recognition systems.

^j There are other views as to what "comprehension" might mean, as in, say, what might be revealed about language from the internal encodings of learned translation functions.

^k With regard to the observation that the represent-and-reason approach is considered to have failed on machine translation, Stuart Russell of the University of California, Berkeley, pointed out to me that this is probably a correct description of an incorrect diagnosis, as not enough effort was directed toward pursuing an adequate represent-and-reason approach, particularly one that is trainable, since language has too many quirks to be captured by hand. This observation is part of a broader perspective I subscribe to calling for revisiting represent-and-reason approaches while augmenting them with advances in machine learning. This task would, however, require a new generation of researchers well versed in both approaches; see the section in this article on the power of success for hints as to what might stand in the way of having this breed of researchers.



Some seemingly complex abilities that are typically associated with perception or cognition can be captured and reproduced to a reasonable extent by simply fitting functions to data.



Perhaps one of the broadest applications of these systems today is in user interfaces (such as automated technical support and the commanding of software systems, as in phone and navigation systems in vehicles). These systems fail often; try to say something that is not very prototypical or not to hide your accent if you have one. But when these systems fail, they send the user back to a human operator or force the user to command the software through classical means; some users even adjust their speech to get the systems to work. Again, while the performance of these systems has improved, according to the adopted metrics, they are today embedded in new contexts and governed by new modes of operation that can tolerate lack of robustness or intelligence. Moreover, as in text, improving their performance against current metrics is not necessarily directed toward, nor requires addressing, the challenge of comprehending speech.^l

Moving to vision applications, it has been noted that some object-recognition systems, based on neural networks, surpass human performance in recognizing certain objects in images. But reports also indicate how making simple changes to images may sometimes hinder the ability of neural networks to recognize objects correctly. Some transformations or deformations to objects in images, which preserve the human ability to recognize them, can also hinder the ability of networks to recognize them. While this does not measure up to the expectations of early AI researchers or even contemporary vision researchers, as far as robustness and intelligence is concerned, we still manage to benefit from these technologies in a number of applications. This includes recognizing faces during autofocus in smart cameras (people do not normally deform their faces but if they do, bad luck, an unfocused image); looking up images that contain cats in online search (it is ok if you end up getting a dog instead); and localizing surrounding vehicles in an image taken by

^l An anonymous reviewer suggested that transcription is perhaps the main application of speech systems today, with substantial progress made toward the preferred metric of "word error rate." The same observation applies to this class of applications.

the camera of a self-driving car (the vulnerability of these systems to mistakes remains controversial in both its scope and how to deal with it at the policy and regulatory levels).

The significance of these observations stems from their bearing on our ability to forecast the future and decisions as to what research to invest in. In particular, does the success in addressing these selected tasks, which are driven by circumscribed commercial applications, justify the worry about doomsday scenarios? Does it justify claims that AI-based systems can now comprehend language or speech or do vision at the levels that humans do? Does it justify this current imbalance of attitudes toward various machine learning and AI approaches? If you work for a company that has an interest in such an application, then the answer is perhaps, and justifiably, yes. But, if you are concerned with scientific inquiry and understanding intelligence more broadly, then the answer is hopefully no.

In summary, what has just happened in AI is nothing close to a breakthrough that justifies worrying about doomsday scenarios. What just happened is the successful employment of AI technology in some widespread applications, aided greatly by developments in related fields, and by new modes of operation that can tolerate lack of robustness or intelligence. Put another way—and in response to headlines I see today, like “AI Has Arrived” and “I Didn’t See AI Coming”—AI has not yet arrived according to the early objective of capturing intelligent behavior. What really has arrived are numerous applications that can benefit from improved AI techniques that still fall short of AI ambitions but are good enough to be capitalized on commercially. This by itself is positive, until we confuse it with something else.

Let me close this section by stressing two points: The first is to reemphasize an earlier observation that while current AI technology is still quite limited, the impact it may have on automation, and hence society, may be substantial (such as in jobs and safety). This in turn calls for profound treatments at the technological, policy,



We succeeded in these applications by having circumvented certain technical challenges instead of having solved them directly.



and regulatory levels.^m The second is that while function-based systems have been an enabling and positive development, we do need to be acutely aware of the reasons behind their success to better understand the implications. A key finding here is that some tasks in perception and cognition can be emulated to a reasonable extent without having to understand or formalize these tasks as originally believed and sought, as in some text, speech, and vision applications. That is, we succeeded in these applications by having circumvented certain technical challenges instead of having solved them directly.ⁿ This observation is not meant to discount current success but to highlight its nature and lay the grounds for this question: How far can we go with this direction? I revisit this issue later in the article.

Human-Level or Animal-Level?

Let me now get to the thoughts that triggered the title of this article in the first place. I believe human-level intelligence is not required for the tasks currently conquered by neural networks, as such tasks barely rise to the level of abilities possessed by many animals. Judea Pearl cited eagles and snakes as having vision systems that surpass what we can build today. Cats have navigation abilities that are far superior to any of those in existing automaton-navigation systems, including self-driving cars. Dogs can recognize and react to hu-

^m Eric Horvitz of Microsoft Research brought up the idea of subjecting certain AI systems to trials as is done to approve drugs. The proper labeling of certain AI systems should also be considered, also as is done with drugs. For example, it has been suggested that the term “self-driving car” is perhaps responsible for the misuse of this AI-based technology by some drivers who expect more from the technology than is currently warranted.

ⁿ For example, one can now use learned functions to recognize cats in images without having to describe or model what a cat is, as originally thought and sought, by simply fitting a function based on labeled data of the form: (image, cat), (image, not cat). While this approach works better than modeling a cat (for now), it does not entail success in “learning” what a cat is, to the point where one can recognize, say, deformed images of cats or infer aspects of cats that are not relayed in the training dataset.

man speech, and African grey parrots can generate sounds that mimic human speech to impressive levels. Yet none of these animals has the cognitive abilities and intelligence typically attributed to humans.

One of the reactions I received to such remarks was: “I don’t know of any animal that can play Go!” This was in reference to the AlphaGo system, which set a milestone in 2016 by beating the world champion in the game. Indeed, we do not know of animals that can play a game as complex as Go. But first recall the difference between performance and intelligence: A calculator outperforms humans at arithmetic without possessing human or even animal cognitive abilities. Moreover, contrary to what seems to be widely believed, AlphaGo is not a neural network since its architecture is based on a collection of AI techniques that have been in the works for at least 50 years.^o This includes the minimax technique for two-player games, stochastic search, learning from self-play, use of evaluation functions to cut off minimax search trees, and reinforcement learning, in addition to two neural networks. While a Go player can be viewed as a function that maps a board configuration (input) to an action (output), the AlphaGo player was not built by learning a single function from input-output pairs; only some of its components were built that way.^p The issue here is not only about assigning credit but about whether a competitive Go function can be small enough to be represented and estimated under current data-gathering, storage, and computational thresholds. It would be quite interesting if this was the case, but we do not yet know the answer. I should also note that AlphaGo is a great example of what one can achieve today by integrating model-based and function-based approaches.

Pushing Thresholds

One cannot of course preclude the possibility of constructing a competitive Go function or similarly complex

functions, even though we may not be there today, given current thresholds. But it begs the question: If it is a matter of thresholds, and given current successes, why not focus all our attention on moving thresholds further? While there is merit to this proposal, which seems to have been adopted by key industries, it does face challenges that stem from both academic and policy considerations. I address academic considerations next while leaving policy considerations to a later section.

From an academic viewpoint, the history of AI tells us to be quite cautious, as we have seen similar phenomena before. Those of us who have been around long enough can recall the era of expert systems in the 1980s. At that time, we discovered ways to build functions using rules that were devised through “knowledge engineering” sessions, as they were then called. The functions created through this process, called “expert systems” and “knowledge-based systems,” were claimed to achieve performance that surpassed human experts in some cases, particularly in medical diagnosis.^q The term “knowledge is power” was used and symbolized a jubilant state of affairs, resembling what “deep learning” has come to symbolize today.^r The period following this era came to be known as the “AI Winter,” as we could finally delimit the class of applications that yielded to such systems, and that class fell well short of AI ambitions.

While the current derivative for progress on neural networks has been impressive, it has not been sustained long enough to allow sufficient visibil-

ity into this consequential question: How effective will function-based approaches be when applied to new and broader applications than those already targeted, particularly those that mandate more stringent measures of success? The question has two parts: The first concerns the class of cognitive tasks whose corresponding functions are simple enough to allow compact representations that can be evaluated efficiently (as in neural networks) and whose estimation is within reach of current thresholds—or thresholds we expect to attain in, say, 10 to 20 years. The second alludes to the fact that these functions are only approximations of cognitive tasks; that is, they do not always get it right. How suitable or acceptable will such approximations be when targeting cognitive tasks that mandate measures of success that are tighter than those required by the currently targeted applications?

The Power of Success

Before I comment on policy considerations, let me highlight a relevant phenomenon that recurs in the history of science, with AI no exception. I call it the “bullied-by-success” phenomenon, in reference to the subduing of a research community into mainly pursuing what is currently successful, at the expense of pursuing enough what may be more successful or needed in the future.

Going back to AI history, some of the perspectives promoted during the expert-systems era can be safely characterized today as having been scientifically absurd. Yet, due to the perceived success of expert systems then, these perspectives had a dominating effect on the course of scientific dialogue and direction, leading to a bullied-by-success community.^s I saw a similar phenomenon during the transition from logic-based approaches to probability-based approaches for commonsense reasoning in the late 1980s. Popular arguments then, like “People don’t reason probabilistically,”

^o Oren Etzioni of the Allen Institute for Artificial Intelligence laid out this argument during a talk at UCLA in March 2016 called *Myths and Facts about the Future of AI*.

^p AlphaZero, the successor to AlphaGo, used one neural network instead of two and data generated through self-play, setting another milestone.

^q One academic outcome of the expert system era was the introduction of a dedicated master’s degree at Stanford University called the “Master’s in AI” that was separate from the master’s in computer science and had significantly looser course requirements. It was a two-year program, with the second year dedicated to building an expert system. I was a member of the very last class that graduated from the program before it was terminated and recall that one of its justifications was that classical computer science techniques can be harmful to the “heuristic” thinking needed to effectively build expert systems.

^r The phrase “knowledge is power” is apparently due to English philosopher Sir Francis Bacon (1561–1626).

^s A colleague could not but joke that the broad machine learning community is being bullied today by the success of its deep learning sub-community, just as the broader AI community has been bullied by the success of its machine learning sub-community.

which I believe carries merit, were completely silenced when probabilistic approaches started solving commonsense reasoning problems that had defied logical approaches for more than a decade. The bullied-by-success community then made even more far-reaching choices in this case, as symbolic logic almost disappeared from the AI curricula. Departments that were viewed as world centers for representing and reasoning with symbolic logic barely offered any logic courses as a result. Now we are paying the price. As one example: Not realizing that probabilistic reasoning attributes numbers to Boolean propositions in the first place, and that logic was at the heart of probabilistic reasoning except in its simplest form, we have now come to the conclusion that we need to attribute probabilities to more complex Boolean propositions and even to first-order sentences. The resulting frameworks are referred to as “first-order probabilistic models” or “relational probabilistic models,” and there is a great need for skill in symbolic logic to advance these formalisms. The only problem is that this skill has almost vanished from within the AI community.

The blame for this phenomenon cannot be assigned to any particular party. It is natural for the successful to be overjoyed and sometimes also inflate that success. It is expected that industry will exploit such success in ways that may redefine the employment market and influence the academic interests of graduate students. It is also understandable that the rest of the academic community may play along for the sake of its survival: win a grant, get a paper in, attract a student. While each of these behaviors seems rational locally, their combination can be harmful to scientific inquiry and hence irrational globally. Beyond raising awareness about this recurring phenomenon, decision makers at the governmental and academic levels bear a particular responsibility for mitigating its negative effects. Senior members of the academic community also bear the responsibility of putting current developments in historical perspective, to empower junior researchers in pursuing their

genuine academic interests instead of just yielding to current fashions.^t

Policy Considerations

Let me now address some policy concerns with regard to focusing all our attention on functions instead of also on models. A major concern here relates to interpretability and explainability. If a medical-diagnosis system recommends surgery, we would need to know why. If a self-driving car kills someone, we would also need to know why. If a voice command unintentionally shuts down a power-generation system, it would need to be explained as well. Answering “Why?” questions is central to assigning blame and responsibility and lies at the heart of legal systems. It is also now recognized that opacity, or lack of explainability, is “one of the biggest obstacles to widespread adoption of artificial intelligence.”^u

Models are more interpretable than functions.^v Moreover, models offer a wider class of explanations than functions, including explanations of novel situations and explanations that can form a basis for “understanding” and “control.” This is due to models having access to in-

formation that goes beyond what can be extracted from data. To elaborate on these points, I first need to explain why a function may not qualify as a model, a question I received during a discussion on the subject.

Consider an engineered system that allows us to blow air into a balloon that then raises a lever that is positioned on top of the balloon. The input to this system is the amount of air we blow (X), while the output is the position of the lever (Y). We can learn a function that captures the behavior of the system by collecting X - Y pairs and then estimating the function $Y = f(X)$. While this function may be all we need for certain applications, it would not qualify as a model, as it does not capture the system mechanism. Modeling that mechanism is essential for certain explanations (Why is the change in the lever position not a linear function of the amount of air blown?) and for causal reasoning more generally (What if the balloon is pinched?). One may try to address these issues by adding more inputs to the function but may also blow up the function size, among other difficulties; more on this next.

In his *The Book of Why: The New Science of Cause and Effect*, Judea Pearl explained further the differences between a (causal) model and a function, even though he did not use the term “function” explicitly. In Chapter 1, he wrote: “There is only one way a thinking entity (computer or human) can work out what would happen in multiple scenarios, including some that it has never experienced before. It must possess, consult, and manipulate a mental causal model of that reality.” He then gave an example of a navigation system based on either reasoning with a map (model) or consulting a GPS system that gives only a list of left-right turns for arriving at a destination (function). The rest of the discussion focused on what can be done with the model but not the function. Pearl’s argument particularly focused on how a model can handle novel scenarios (such as encountering roadblocks that invalidate the function recommendations) while pointing to the combinatorial impossibility of encoding such contingencies in the function, as it must have a bounded size.

^t I made these remarks over a dinner table that included a young machine learning researcher, whose reaction was: “I feel much better now.” He was apparently subjected to this phenomenon by support-vector-machine (SVM) researchers during his Ph.D. work when SVMs were at their peak and considered “it” at the time. Another young vision researcher, pressed on whether deep learning is able to address the ambitions of vision research, said, “The reality is that you cannot publish a vision paper today in a top conference if it does not contain a deep learning component, which is kind of depressing.”


^u See Castellanos, S. and Norton, S. Inside Darpa’s push to make artificial intelligence explain itself. *The Wall Street Journal* (Aug. 10, 2017); <http://on.wsj.com/2vmZKlM>; DARPA’s program on “explainable artificial intelligence”; <https://www.darpa.mil/program/explainable-artificial-intelligence>; and the E.U. general data protection regulation on “explainability”; <https://www.privacy-regulation.eu/en/r71.htm>

^v I am referring here to learned and large functions of the kind that stand behind some of the current successes (such as neural networks with thousands or millions of parameters). This excludes simple or well-understood learned functions and functions synthesized from models, as they can be interpretable or explainable by design.


There is today growing work on explaining functions, where the vocabulary of explanations is restricted to the function inputs. For example, in medical diagnosis, an explanation may point to important inputs (such as age, weight, and heart attack history) when explaining why the function is recommending surgery. The function may have many more additional inputs, so the role of an explanation is to deem them irrelevant. In vision applications, such explanations may point to a specific part of the image that has led to recognizing an object; again, the role of an explanation is to deem some pixels irrelevant to the recognition. These explanations are practically useful, but due to their limited vocabulary and the limited information they can access, they could face challenges when encountering novel situations. Moreover, they may not be sufficient when one is seeking explanations for the purpose of understanding or control.

Consider a function that predicts the sound of an alarm based on many inputs, including fire. An input-based explanation may point to fire as a culprit of the alarm sound. Such an explanation relies effectively on comparing this scenario to similar scenarios in the data, in which the sound of the alarm was heard soon after fire was detected; these scenarios are summarized by the function parameters. While this may explain why the function reached a certain conclusion, it does not explain why the conclusion (alarm sound) may be true in the physical world.^w Nor does it explain how fire triggers the alarm; is it, say, through smoke or through heat? The importance of these distinctions surfaces when novel situations arise that have not been seen before. For example, if the alarm is triggered by smoke, then inviting a smoker into our living room might trigger an alarm even in the absence of fire. In this case, pointing to fire as an explanation of the sound would be problematic. Humans arrive at such conclusions without ever seeing a smoker, which can also be achieved through reasoning on an appropriate

^w The function imitates data instead of reasoning about a model of the physical world.



Human-level intelligence is not required for the tasks currently conquered by neural networks, as such tasks barely rise to the level of abilities possessed by many animals.



model. However, to do this based on a learned function, the function would need to be trained in the presence of smokers or other smoke-producing agents while defining smoke as an input to the function and assuring that smoke mediates the relationship between fire and alarm, a task that requires external manipulation.

As Pearl told me, model-based explanations are also important because they give us a sense of “understanding” or “being in control” of a phenomenon. For example, knowing that a certain diet prevents heart disease does not satisfy our desire for understanding unless we know why. Knowing that the diet works by lowering the cholesterol level in the blood partially satisfies this desire because it opens up new possibilities of control. For instance, it drives us to explore cholesterol-lowering drugs, which may be more effective than diet. Such control possibilities are implicit in models but cannot be inferred from a learned, black-box function, as it has no access to the necessary information (such as that cholesterol level mediates the relationship between diet and heart disease).

A number of researchers contacted me about the first draft of this section, which was focused entirely on explanations, to turn my attention to additional policy considerations that seem to require models. Like explanations, they all fell under the label “reasoning about AI systems” but this time to ensure that the developed systems would satisfy certain properties. At the top of these properties were safety and fairness, particularly as they relate to AI systems that are driven only by data. These considerations constitute further examples where models may be needed, not only to explain or compensate for the lack of enough data, but to further ensure we are able to build the right AI systems and reason about them rigorously.

A Theory of Cognitive Functions

One reaction I received concerning my model-based vs. function-based perspective was during a workshop dedicated to deep learning at the Simons Institute for the Theory of Computing in March 2017. The workshop

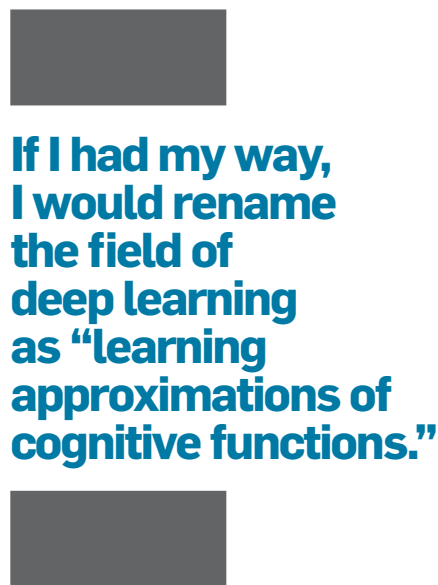
title was “Representation Learning,” a term used with increasing frequency by deep learning researchers. If you have followed presentations on deep learning, you will notice that a critical component of getting these systems to work amounts to finding the correct architecture of the neural network. Moreover, the architectures vary depending on the task, and some of their components are sometimes portrayed as doing something that can be described at an intuitive level. For example, in language, one uses an encoder-decoder architecture in which the encoder transforms a sentence in the source language into an internal encoding, and the decoder then generates a sentence in the target language.

The reaction here was that deep learning is not learning a function (black box) but a representation since the architecture is not arbitrary but driven by the given task.^x I see this differently. Architecting the structure of a neural network is “function engineering” not “representation learning,” particularly since the structure is penalized and rewarded by virtue of its conformity with input-output pairs. The outcome of function engineering amounts to restricting the class of functions that can be learned using parameter estimation techniques. This process is akin to restricting the class of distributions that can be learned after one fixes the topology of a probabilistic graphical model. The practice of representation learning is then an exercise in identifying the classes of functions that are suitable for certain tasks.^y

In this context, I think what is needed most is a theory of cognitive functions. A cognitive function captures a relationship that is typically associated with cognition (such

^x There are other broader interpretations of the term “representation learning.”

^y An anonymous reviewer suggested today’s practice of building deep neural networks can be viewed as the application of a new programming paradigm called “differentiable programming.” In this view, networks are carefully structured by a programmer using various differentiable program modules (such as convolutional layers, pooling layers, LSTM layers, residual blocks, and embedding layers). The compiler then differentiates and structures them for GPU execution. The key is to structure the program so the gradients are guided to do the right thing.



If I had my way, I would rename the field of deep learning as “learning approximations of cognitive functions.”

as mapping audio signals to words and mapping words to some meaning). What is needed is a catalogue of cognitive functions and a study of their representational complexity—the size and nature of architectures needed to represent them—in addition to a study of their learnability and approximability. For Boolean functions, we have a deep theory of this kind. In particular, researchers have cataloged various functions in terms of the space needed to represent them in different forms (such as CNFs, DNFs, and OBDDs). What we need is something similar for real-valued functions that are meant to capture cognitive behaviors. In a sense, we already have some leads into such a theory; for example, researchers seem to know what architectures, or “function classes,” can be more effective for certain object-recognition tasks. This needs to be formalized and put on solid theoretical ground.^z Such a theory would also include results on the learnability of function classes using estimation techniques employed by the deep learning community, particularly “gradient descent.” Interestingly, such results were presented at the Representation Learning workshop I referenced earlier in a talk called “Failures of Deep Learning” in which very simple functions were presented that defeat current estimation techniques. Even more interestingly, some have dismissed the importance of such results in side discussions on the grounds that the identified functions are not of practical significance; read “these are not cognitive functions” or “we have come a long way by learning approximations to functions.” In fact, if I had my way, I would rename the field of deep learning as “learning approximations of cognitive functions.”

The term “cognitive functions” surprised some colleagues who told me that “perception functions” may be more suitable, given that the current successes of deep learning have been

^z The properties of learned functions may carry quite a bit of insight about the structure of our world; for example, linguists are called upon to study this phenomenon and unveil what learned translation functions may be revealing about the structure of language.

mostly in instinct-based perception (such as computer vision and language processing). I agree with this observation, except nothing at this stage prohibits functions from providing reasonable approximations to more high-level cognitive tasks. In fact, Go functions have been constructed using neural networks, even though they are not yet competitive with hybrid systems (such as AlphaGo). Admittedly, it is also possible that we might later realize that functions (of practical size) cannot provide reasonable approximations to a wide enough class of cognitive functions despite progress on pushing computational and data thresholds. The association with perception would then be more established in that case. Time will tell.

Conclusion

This article was motivated by concerns I and others have had on how current progress in AI is being framed and perceived. Without a scholarly discussion of the causes and effects of recent achievements, and without a proper perspective on the obtained results, one stands to hinder further progress by perhaps misguiding the young generation of researchers or misallocating resources at the academic, industrial, and governmental levels. One also stands to misinform a public that has developed a keen interest in AI and its implications. The current negative discussions by the general public on the AI singularity, also called “super intelligence,” is partly due to the lack of accurate framings and characterizations of recent progress. With almost everyone being either overexcited or overwhelmed by the new developments, substantial scholarly discussions and reflections have gone missing.

I had the privilege of starting my research career in AI around the mid-to-late 1980s during one of the major crises in the field, a period marked by inability instead of ability. I was dismayed then, as I sat in classes at Stanford University, witnessing how AI researchers were being significantly challenged by some of the simpler tasks performed routinely by humans. I now realize how such crises can be enabling for scientific discovery, as they fuel academic thinking, empower researchers, and create grounds for


profound scientific contributions.^{aa} On the other hand, I am reminded how times of achievements can potentially slow scientific progress by shifting academic interests, resources, and brain power too significantly toward exploiting what was just discovered, at the expense of understanding the discoveries and preparing for the moment when their practical applications have been delimited or exhausted.

There are many dimensions to such preparation. For the deep learning community, perhaps the most significant is a transition from the “look what else we can do” mode to a “look what else you can do” mode. This is not only an invitation to reach out to and empower the broader AI community; it is also a challenge since such a transition is not only a function of attitude but also an ability to characterize progress in ways that enable people from outside the community to understand and capitalize on it. The broader AI community is also both invited and challenged to identify fundamental ways in which functions can be turned into a boon for building and learning models. Given where we stand today, the question is not whether it is functions or models but how to profoundly integrate and fuse functions with models.^{ab} This aim requires genuine cross-fertilization and the training of a new generation of researchers who are well-versed in and appreciative of various AI methods, and who are better informed about the history of AI.

I conclude with this reflection: I wrote the first draft of this article in November 2016. A number of colleagues provided positive feedback then, with one warning about a negative tone. I put the draft on hold for some months as a result while con-

tinuing to share its contents verbally in various contexts and revising accordingly. The decision to eventually release a first draft in July 2017 was triggered by two events: a discussion of these thoughts at a workshop organized by the UCLA School of Law and other discussions with colleagues outside of AI, including architecture, programming languages, networks, and theory. These discussions revealed a substantial interest in the subject and led me to conclude that the most important objective I should be seeking is “starting a discussion.” I may have erred in certain parts, I may have failed to give due credit, and I may have missed parts of the evolving scene. I just hope the thoughts I share here will start that discussion, and the collective wisdom of the community will correct what I may have gotten wrong.

Acknowledgments

I benefited greatly from the feedback I received from anonymous reviewers and from colleagues who are too many to enumerate but whose input and discussions were critical to shaping the thoughts expressed here. However, I must specifically acknowledge Judea Pearl for inspiring the article and for helping with various arguments; Stuart Russell for providing very thoughtful and constructive feedback; Guy Van den Broeck for keeping me interested in the project every time I almost gave up; and Arthur Choi for being a generous and honest companion to the thinking that went into it. Finally, I wish to thank Nils Nilsson for telling me that he wished he had written the article and for kindly inviting me to share his feedback with others. This is an ultimate reward. 

^{aa} Judea Pearl’s seminal work on probabilistic approaches to commonsense reasoning is one example outcome of the crisis.

^{ab} An anonymous reviewer brought to my attention works on the analyses of human cognition, particularly Daniel Kahneman’s book *Thinking Fast and Slow*. The reviewer said “fast” naturally maps onto function-based and “slow” onto model-based, and there is a strong argument in the literature on cognitive science that people must at least combine them both. The reviewer further pointed out that there are a variety of cognitive architectures that embody specific hypotheses about such hybrids.

Adnan Darwiche (darwiche@cs.ucla.edu) is a professor in and chairman of the Computer Science Department at the University of California, Los Angeles, CA, USA.

Copyright held by author.



Watch the author discuss his work in this exclusive *Communications* video. <https://cacm.acm.org/videos/human-level-intelligence-or-animal-like-abilities>