2. Evaluating revolutions in artificial intelligence from a human perspective

Kenneth D. Forbus, Northwestern University

Artificial intelligence (AI) has made considerable progress over the past 40 years, leading to both important applications now in daily use but also a lot of hype in the popular press. This chapter outlines a framework for understanding progress in AI to help understand how to better measure that progress. It describes three revolutions currently under way: *deep learning*, *knowledge graphs* and *reasoning*. It summarises the progress and limitations in each and contrasts them with human capabilities for learning, knowledge and reasoning. The chapter also summarises a fourth revolution to come, *integrated intelligence*, where the goal is to create systems with agency. It discusses some additional implications for measuring AI progress with respect to learning, knowledge, reasoning and agency.

Introduction

There has been considerable progress in artificial intelligence (AI) over the past four decades. Due to vast increases in the availability of computing power and data, this progress has accelerated over the last 20 years. These increases, combined with steady scientific progress, have led to revolutionary advances. In this context, a revolution occurs when scientific progress and its application through technology lead to a radical improvement in existing capabilities, or new capabilities, that have become widely used and provide significant benefits. This notion of revolution factors out stunts and laboratory experiments. This makes it useful for understanding how to measure AI and to think about how it might impact education and work.

By this criterion, there are three AI revolutions in progress: *deep learning*, *knowledge graphs* and *reasoning*. This chapter discusses them, comparing and contrasting with human capabilities. It then describes a fourth AI revolution just beginning, *integrated intelligence*, which builds on the other three. It concludes with further discussion of measuring progress in AI relative to human capabilities.

Artificial intelligence Revolution 1: Deep learning

The revolution everyone has heard about through the popular press is deep learning (LeCun, Bengio and Hinton, 2015[1]). Neural networks organise computation via large collections of simple computational units metaphorically inspired by neurons in brains. The term “deep” refers to the number of layers in the network. Research on neural networks has received varying amounts of attention since the 1950s. Until recently, the lack of computing power and training data meant that only small networks could be explored. Several decades of research on algorithm improvements, combined with graphical processing units originally developed for video games, finally enabled large networks to be trained using massive amounts of data. This has led to systems that performed far better than previous approaches on tasks such as image captioning, facial recognition, speech recognition and automatic natural language translation. These performance improvements have led to deployed applications in these areas and many others that are in daily use around the world.

Although deep learning models are inspired by biological neurons, they do not necessarily learn in human-like ways. Indeed, given the state of knowledge in neuroscience, claims of biological plausibility must be taken sceptically (Bengio et al., 2015[2]; Marcus, 2018[3]). Research on learning in psychology and AI highlights some important differences between deep learning and human learning. Human learning involves multiple kinds of processes and occurs at multiple time-scales. For example, conceptual change involves articulation of knowledge, working through implications of models over weeks, months and years. It can involve radical discontinuities in conceptual understanding. By contrast, skill learning, which can also take years, is often hard to articulate. Skill learning can be described by a power law, which is smooth. Both conceptual change and skill learning are incremental. By contrast, deep learning requires its training data all at once, but otherwise is more smooth and implicit, like skill learning.

The widespread use of deep learning has exposed three limitations. The first is brittleness, where small changes in inputs lead to dramatically different outputs. The second is data efficiency, i.e. the amount of training data required. The third is explainability, i.e. it is hard to understand why errors occur and how to fix them. Each limitation is discussed in turn.

Brittleness

Deep learning models typically approximate a function, e.g. given an input, it produces an output. Inputs that are close to what the model was trained on often yield reasonable results. Unfortunately, inputs from outside the training set can sometimes yield arbitrarily incorrect results. Changing a few pixels, for example,

can cause a scene to be classified differently (Goodfellow, Shlens and Szegedy, 2015[4]). Adding a few small patches can cause the system to perceive a stop sign as a speed limit sign (Eykholt et al., 2018[5]). Indeed, images that look like random noise to people are sometimes confidently recognised by deep learning systems as everyday objects (Nguyen, Yosinski and Clune, 2015[6]). Human vision has limitations, as evidenced by well-known visual illusions. However, human visual illusions are nothing like these misperceptions.

Brittleness is a limitation of deep learning models in general, not just in vision. Deep learning models for question-answering and reading comprehension are subject to similar problems. For example, including extra irrelevant information that a human would ignore can mislead such systems (Jia and Liang, 2017[7]). Part of the problem is that these systems are being tested using instruments that do not necessarily measure what they are intended to measure. Reading comprehension measures designed by AI researchers, for instance, almost always use multiple-choice questions because they are simple to score. However, this method enables systems to pick up correlations between terms in the texts and the answers. Indeed, the systems will sometimes do almost as well in picking out answers even when it does not know the questions (Kaushik and Lipton, 2018[8]).

Data efficiency

Data efficiency is analogous to fuel efficiency. It is the amount of data needed to achieve a particular level of performance. The more layers in a deep learning system, the more parameters there are, and the more training data is required. How much data? Consider learning to play the game of Go. AlphaGo (Silver et al., 2016[9]) and AlphaZero (Silver et al., 2018[10]), by combining deep learning with a hand-built search routine optimised for such games, were able to learn to play Go well enough to beat the best human players. The amount of data used was massive. Achieving AlphaGo's exposure to games of Go, in human terms, would require a person to live about 6 800 years (Forbus, Liang and Rabkina, 2017[11]).[1] Similarly, GPT-3, a large-scale statistical language model, was trained on many billions of tokens, far more than people experience in a lifetime. Nevertheless, while it can sometimes generate text that looks locally plausible, it is often globally incoherent and often the statements it produces are simply wrong (Marcus and Davis, 2020[12]).[2]

By contrast, consider teaching a human assistant to do a routine office task or assembly task. One does not need hundreds of training examples, let alone millions. A growing AI research area, *interactive task learning* (Gluck and Laird, 2019[13]), seeks to train systems with the same amounts and kinds of experience that a person would receive. This requires data efficiency but also the ability to communicate with the system using natural modalities (e.g. language, sketching, gesture, speech). Interactive task learning is a step towards integrated intelligences, which are discussed more below.

Explainability

Explainability is important so that people can understand the rationale for decisions that a system makes and to troubleshoot when problems arise. Deep learning systems are opaque, since they rely on massive sets of numbers that typically do not have any clear interpretation. It should be noted that human learning is not always explainable. The field of psychophysics, for example, explores how perception works, including what is built-in versus learnt. Similarly, cognitive psychologists have studied learning procedural knowledge. In so doing, they have generated both a wealth of behavioural results and computational models that capture temporal properties and error patterns in such learning. For evaluating skills, performance measures often suffice.

By contrast, in science, engineering, law and medicine, as well as many other fields, practitioners are required to explain their results. This puts a premium on explicit conceptual knowledge, as discussed further below. Even in these professions, some forms of knowledge are less easily articulated, commonly

called *tacit knowledge* (Forbus, 2019[14]). Such knowledge often concerns how situations in the everyday world can be understood in terms of professional knowledge and the construction of a *mental model* (Gentner and Stevens, 1983[15]). Research in qualitative reasoning (Forbus, 2019[14]) suggests this split between formulating and using mental models is common in professional reasoning (Klein, 1999[16]; Engel, 2008[17]).

Deep learning is indeed a valuable technology, despite these limitations. The trajectory of deep learning also highlights an important lesson about new technologies and their potential impacts. Ideas can take considerable time to hone – conceptual advances and algorithmic improvements can take decades. They may only take off when environmental factors are favourable (e.g. massive amounts of data and of computation). Other approaches to learning in AI (including inductive logic programming, analogy and interactive task learning) are earlier in their trajectories. However, they may yet hit that combination of progress and environmental factors to become revolutions on their own.

There is more to AI than learning. These same factors have led to revolutions in two other areas, as described next.

Artificial intelligence Revolution 2: Knowledge graphs

Knowledge is a key ingredient of intelligence. Decades of research on knowledge representation in AI has led to formal methods for capturing many kinds of facts in machine-usable form at industrial scale. Such compendia are called knowledge graphs.[3]

An overview of knowledge graphs

Knowledge graphs provide explicit representations of knowledge, including descriptions of entities in terms of statements about them, relationships among entities and a variety of other conceptual structures, such as arguments, explanations and plans. Knowledge graphs include *ontologies*, i.e. representations that describe types of things in the world (entities) and facts about those entities. Ontologies are typically constructed by hand. Facts are added via a combination of hand-engineering and automatic methods, such as importing information from databases and extracting from texts such as webpages.

Knowledge graphs have been applied in multiple industries. For example, Google's Knowledge Graph uses 70 billion facts describing over a billion entities to improve web search and perform question-answering (Noy et al., 2019[18]). If a user is looking for a restaurant in San Francisco, for instance, the system needs to understand whether they are searching from California or Colombia. Similarly, Amazon uses one knowledge graph of product information in its recommendation system (Dong, 2018[19]) and another of general information for question-answering via Alexa.[4] In its operations, Facebook uses a knowledge graph of people and their relationship with others (Noy et al., 2019[18]). Several scientific communities have created their own ontologies using Semantic Web representations (Allemang, Hendler and Gandon, 2020[20]) to help index and retrieve data, results and publications. Similarly, publishers have created ontologies to help others access their articles and works.[5]

It is interesting to compare today's knowledge graphs with human knowledge. Billions of facts is quite a lot. Certainly, no human being has memorised, say, all of the named locations on Earth plus all of the people mentioned in Wikipedia. Therefore, along some dimensions, knowledge graphs go beyond what people know. On the other hand, people know massive amounts about the physical world, social world and mental worlds of agents. Estimating the amount of knowledge people have is challenging (Forbus, Liang and Rabkina, 2017[11]). What is clear, however, is that humans use at least three types of knowledge that are rarely found in deployed knowledge graphs. These are multimodal grounding of knowledge, episodic knowledge and inferential knowledge.

Multimodal grounding of knowledge

Some human knowledge is grounded in perception and sensorimotor experience. Humans know what a cat looks like in repose versus at play, and how many dishes can be carried safely. This multimodal grounding of knowledge – what the world looks, sounds, feels and tastes like – is almost entirely missing from today's knowledge graphs.[6]

Episodic knowledge

Episodic knowledge is one's personal experience, which serves as a resource in efficient reasoning. For example, human experience suggests that dogs make good pets whereas sharks do not. Humans can decide this quickly, even if it takes a moment to articulate clear reasons for the difference. People also learn from the experiences of others, as transmitted via stories, allowing cultures to accumulate knowledge across space and time.

The family of techniques called *case-based reasoning* (CBR) use forms of analogy to apply knowledge about previous situations to new problems. Some CBR applications have scaled to millions of cases (Jalali and Leake, 2018[21]). However, current large-scale CBR systems use vector representations[7] rather than the relational information used in other areas in CBR and in knowledge graphs. Relational knowledge is crucial to capture the full range of human knowledge. It includes the ability to represent plans, theories, arguments and social relationships.[8] Although people's general knowledge (often called *semantic memory*) is tightly integrated with their episodic knowledge, large-scale relational CBR systems do not currently exist.

Inferential knowledge

Inferential knowledge consists of general knowledge that can be used to derive new consequences. For example, one might know that dogs are commonly kept as pets in some cultures, because dogs are often affectionate towards people and people often keep pets to provide affection.

Inferential knowledge is valuable because it supports reasoning, which enables systems to come up with novel conclusions about new situations and problems. It typically takes the form of rules in a formal language, allowing algorithms to use them automatically and unambiguously. There are a variety of formal languages used in practice, varying considerably in expressive power. Some are strictly logical. Others incorporate probabilities and statistical reasoning. Still others incorporate procedural (i.e. how-to) knowledge. Cyc is the knowledge graph with the most inferential knowledge. It uses a highly expressive logic, enabling it to be used for a variety of applications.[9]

The trajectory of knowledge graphs

The limitations of knowledge graphs summarised above are natural given that industrial, commercial and scientific knowledge graphs are purpose-built. For example, a genomics ontology will not include cultural products, which are a major component of Google and Bing's knowledge graphs, and they in turn do not include detailed representations of genomes. There is a recognition that the field needs to start building *open knowledge networks*.[10] This will allow existing knowledge graphs to be integrated more easily and extended in ways that will benefit everyone.

As knowledge graphs are applied to more tasks, the range of types of knowledge they include will grow. Multimodal grounding of concepts in perception is important for creating robots that can operate in less constrained environments for tasks such as elder care and housekeeping. Personal assistant software is already helping drive development of formal representations of human events and preferences, to better model the particular people they work with.

Understanding context and higher-quality understanding of natural language is leading to more work on representing inferential knowledge. Today's assistants might handle questions like "What is the weather today?" followed by "How about the weekend?" If the initial question was "What is my schedule today?", the follow-up question about the weekend concerns schedules, not the weather. Handling such changes can be done via semantic representations that explicitly represent the computations used in answering questions (Andreas et al., 2020[22]).

Artificial intelligence Revolution 3: Reasoning

Reasoning is a hallmark of human intelligence. There are many kinds of reasoning, but in their essence they involve combining facts to reach new conclusions. This sub-section examines several dimensions of reasoning, common sense versus professional reasoning, some examples of the reasoning revolution, and compares human and AI reasoning capabilities.

Dimensions of reasoning

Depth of reasoning

Depth of reasoning relates to how many steps are needed to perform a task. A question like "How old is the president of Germany?"[11] provides a simple example. Answering the question requires two steps: identifying the president of Germany and then finding out that person's age. The reasoning revolution is due in part to the ability of some specialised reasoners to go far beyond human capabilities in terms of depth.

Types and breadth of knowledge

Since reasoning involves combining information to produce new conclusions, it invariably involves two dimensions of knowledge: *types* and *breadth*. Consider the kinds of question-answering that IBM's original Watson system performed. To beat the best human players at the game of *Jeopardy!*, Watson integrated a massive amount of knowledge: Wikipedia, multiple reference works, volumes of literature and other information (Fan et al., 2012[23]). Watson's knowledge was thus quite broad but only moderately deep. It was capable of evaluating spatial and temporal constraints on answers, for example, and reasoning about types of entities.

Flexibility of reasoning

Humans marshal vast amounts of multiple types of knowledge efficiently and effectively. We are able to come to some conclusions rapidly, even with very little information, and can combine experience with principles to come up with effective plans in novel circumstances.

Common sense reasoning versus professional reasoning

One of the grand challenges for AI is common sense reasoning (Davis and Marcus, 2015[24]). This includes everyday knowledge of the physical world, such as knowing that it is possible to pull with a string but not push with it.[12] Common sense reasoning gets people through the activities of their days – from navigating, cooking and cleaning to interacting with others.

Professional reasoning is grounded in common sense reasoning. New problems come couched in a mix of professional and everyday terms. For example, a doctor must diagnose a patient's symptoms; a judge reviews a legal case to judge; a designer develops a device. In doing so, they translate from their everyday terms and models into their professional terms and models. Thus, a key part of professional reasoning is

model formulation. This involves construction of a mental model for a situation in professional terms that enables a problem to be solved.[13] The knowledge needed for professional reasoning includes explicit principles taught (e.g. the definition of a tort, the equations of thermodynamics). However, it also includes strategies and tactics to apply principles to new situations. This strategic and tactical knowledge is often tacit, communicated via apprenticeship and reflection on experience.

Examples from the reasoning revolution

While many issues in understanding reasoning remain open scientific questions, several areas have shown revolutionary progress. The most spectacular AI reasoning systems today operate after model formulation has occurred. In designing a new distributed computing scheme, for example, Amazon engineers write a description of their system using a temporal logic (Newcombe et al., 2015[25]). A *model checker*, which is a type of reasoning system, scrutinises the possible combinations of events in the system to search for bugs. These bugs can be subtle, sometimes only happening after a sequence of over 30 events. Yet, given the velocity of modern computation, such bugs are likely to manifest several times a week, making their detection important.

Many applications use automatic model formulation within specific narrow domains. For example, Facebook uses model checking to evaluate the code being added to the 100 million lines of code that runs their services (Distefano et al., 2019[26]). If an engineer has checked in a software patch that could be problematic, it is flagged for review automatically.

Similar superhuman feats of reasoning occur daily by *satisfiability solvers* used in logistics and operations planning. These satisfiability solvers operate over systems of millions of constraints far more quickly and generate better solutions than human planners can produce [e.g. (Simonis, 2001[27])].

In engineering design, combinations of *qualitative and quantitative models* have been used to do several types of commercially important reasoning. These systems perform causal reasoning, using human-like conceptual models that enable them to provide natural explanations that engineers understand. For example, AutoSteve (Price, 2000[28]) performs failure modes and effects analyses to evaluate what could go wrong in designs for automobile electrical systems. At Xerox, such models are used to evaluate potential new modules for high-performance print engines at design time for cost effectiveness. They are also used as the basis for model-based control software. This software reconfigures itself dynamically on-site as the complex electro-mechanical components of high-end printers change due to reconfiguration of physical parts or parts wearing out (Fromherz, Bobrow and de Kleer, 2003[29]).

Table 2.1. Dimensions of reasoning in people versus AI systems

Dimension	People	AI systems
Depth of reasoning	Medium	Very High
Types of knowledge	Many	Few
Breadth of knowledge	Wide	Narrow
Flexibility of reasoning	High	Low

Reasoning in people vs. AI systems

As Table 2.1 illustrates, AI systems and humans have different strengths and weaknesses. AI systems are better at high-depth reasoning than humans. Asking whether people could be trained to do as well as model checkers and satisfiability solvers is like asking whether people can be trained to outrun an automobile or outlift a crane. The firing frequency of neurons ranges roughly between 1-500 Hertz, but brains have massive parallelism, with billions of neurons. Modern computers, by contrast, are much more

serial, but operate around 10 million times faster. For tasks with many sequential reasoning steps (i.e. high-depth reasoning), AI systems have an overwhelming advantage.

People still have the edge over AI systems on common sense reasoning, model formulation and many kinds of professional reasoning. These require many types of reasoning, a wide breadth of knowledge and high flexibility. Of these, more aspects of professional reasoning are likely to be the next to be more broadly automated. It is already happening in engineering domains to some degree, due to the mathematical nature of many engineering models.

However, the experiential component to engineering is harder to capture, and that will take extending more AI systems to accumulate episodic memories. Medicine and law seem to rely even more heavily on experience and rules distilled from experience by practitioners. Ultimately, AI systems will be able to accumulate far more experience in these domains than any human being due to our lifespan limitations and the ability of AI to combine experience across many systems.[14] Scaling up analogical reasoning to handle CBR and distillation of rules from such massive collections will be an important technical challenge.

The coming fourth artificial intelligence Revolution: Integrated intelligence

The three revolutions – deep learning, knowledge graphs and reasoning – are far from over. Moreover, they interact synergistically. For example, deep learning has been used to help build knowledge graphs and estimate which subgoals are worth pursuing in reasoning (like choosing good moves in a game). Knowledge graphs are being integrated with deep learning systems to improve their reasoning capabilities. However, there is a key limitation that none of these revolutions addresses. Even when machine learning is used, today's AI systems are still constructed and maintained by humans. This sub-section outlines the *integrated intelligence* revolution to come, the fourth AI revolution.

From pipelines to organisms

People hand-craft the structure of today's AI systems. Even when machine learning is used, people gather and curate the data, inspect the results, and then tune/tweak the data and the learning algorithms. AI systems do not, on the whole, set their own learning goals and decide what data to gather to achieve them. Human scientists and engineers decide on the ontology for systems and what kinds of information are needed in knowledge graphs. If the kinds of tasks change, humans must decide on the need for new knowledge and how to acquire it. Human engineers tune and tweak reasoning systems and problem formulations to wring the most accurate results and effective performance out of systems.

Today's deployed AI systems can be thought of as pipelines – machines whose architecture is purpose-built for a particular range of tasks. For personal assistant AI systems, such as Google Assistant, Amazon's Alexa, Microsoft's Cortana, Samsung's Bixby or SK Telecom's Aria, these pipelines are massive. Human engineers are constantly updating or even replacing models and algorithms as needed to adapt them to changing circumstances. The exact number of human engineers needed for these tasks is held closely by the companies involved. However, depending on the system, it likely ranges from hundreds to thousands of people.

This reliance on human intervention does not scale. For example, one of the visions in the 20-year AI Roadmap for the US (Gil and Selman, 2019[30]) is the idea of personal AI assistants. Unlike today's assistants, these systems will adapt themselves to the work and personal lives of humans to help them, with humans owning all their own data. Such a vision is impossible with the pipeline model – the systems themselves must be able to maintain, adapt and learn on their own. In other words, pipelines are missing *agency*. Building AI systems that, like biological organisms, have agency is the heart of the coming fourth revolution of integrated intelligence.

The road to the fourth revolution

Several AI research threads are building up the science base for this revolution. Cognitive systems research[15] typically involves multiple AI capabilities, as well as exploring how specific capabilities need to be modified or extended to function as part of larger-scale systems. Cognitive architecture research [e.g. (Anderson, 2009[31]; Laird, 2012[32]; Forbus and Hinrichs, 2017[33])] explores computational models of larger-scale phenomena in human cognition. None of these are deployed for daily use at the scale that would constitute a revolution – yet.[16]

Examining the difference between human development and the way AI systems are engineered today provides a lens for identifying the capabilities needed to build organisms. People successfully learn to become members of a culture and profession. This learning is cumulative, with new material building on previously learnt material in ways that none of today's AI technologies can match. People quickly adapt when joining a new group and when dealing with new circumstances. They can be apprentices, which catalyses learning through interactions with others.

Endowing machines with these capabilities involves numerous scientific challenges, most of which are still being formulated. The recent 20-year AI Roadmap for the US (Gil and Selman, 2019[30]) provides a decomposition of many of these issues and suggests milestones to bridge the gaps.[17] It is thus a resource for projecting progress of the field over that period, based on a consensus view developed by over 90 researchers.

A roadmap is not a schedule, of course. Our understanding of the issues and their difficulties are naturally incomplete and progress requires ample resources. Nor does this enable prediction for the start of daily-use beneficial applications that will herald the start of this revolution. Resources for development of deep learning, knowledge graphs and reasoning ramped up as successful applications were fielded, and the same will apply to integrated intelligence.

Additional implications for measuring artificial intelligence progress

Clearly, no single test of AI progress will suffice for measuring its capabilities. This should not be surprising, since psychometricians long ago gave up on defining a single test for measuring human intelligence. AI systems are still well below the capability and sophistication of human minds in most regards. However, the need to measure their progress on human scales demands at least a battery of tests.

These tests will need to be designed by teams of psychometricians and AI researchers. The psychometricians have a better understanding of how to test extremely complex systems (i.e. humans) and to achieve reliable results over time. The AI researchers have a better understanding of the strengths and weaknesses of the technologies.

A battery should incorporate tests for learning, knowledge, reasoning and agency.

Learning

Separate tests will be needed for perceptual, procedural and conceptual learning as each has different expectations and requirements for data efficiency and explainability. Prior work on modelling conceptual change has used tests developed by cognitive scientists. These tests measure the kinds of models a system learns. They also measure whether trajectories of learnt models are similar to those of human learners (Friedman and Forbus, 2010[34]). Assembling a set of such developmental benchmarks could help ensure that intermediate learnt models make sense to the human collaborators of AI systems.

Knowledge

With explicit knowledge graphs, the amount of knowledge in different areas can be measured directly, with effectiveness for reasoning tested by sampling. Some approaches attempt to use language models and other distributed representation systems as knowledge bases. Evaluating such models is far more difficult, given their lack of interpretability.

For inferential knowledge in some domains, such as verification and scheduling, the knowledge is entirely deductive. The consistency of such knowledge can be automatically verified in many cases. As the range of inferential knowledge is expanded, to include more abductive reasoning and common sense reasoning, benchmark domains will be required for testing, since plausible-sounding rules can turn out to be insufficient and/or problematic.

Testing episodic memory is harder. It requires looking at the experiences required for a job and comparing what the system/person has versus what is required (see vocational education and training discussion below).

Reasoning

Reasoning in technical domains, such as engineering, is worth measuring. Performing technical analyses in engineering domains is reasonably well understood now in many cases, but two more difficult frontiers should also be included. The first is reasoning for model formulation, i.e. can a system frame problems for technical analyses when expressed in everyday terms? The second is reasoning for communication, i.e. can the system interact with professionals using the mixture of technical concepts and everyday concepts that people use when working together to solve problems?

Common sense reasoning is especially difficult to measure since it is broad and often tacit. The knowledge may be tacit because it is grounded in experience. For example, one knows that tipped cups spill because it has been seen to happen, not because of a simulation or an explicit rule.

Agency

Testing the ability of AI systems to manage their own operations and learning over extended periods of time – ultimately, decades – will be extraordinarily difficult. Measuring the ability to handle a suite of complex problems, of varying types and progressive in difficulty, may be a reasonable way to approximate the process. As the complexity of AI systems grows, models of trust will likely be developed that look more like collaboration with other organisms (e.g. working dogs, draft horses, human collaborators) than engineering verification and validation. In other words, experience will lead to trust in systems over time. It is not even clear what verification and validation would mean when, for example, training a system via apprenticeship to fit within the practices of an organisation. In its assessments used in vocational education and training developed in Germany, Chapter 9 provides both a useful breakdown of many workplace skills and examples of assessment techniques that might be adaptable for this purpose.

Conclusions

The three AI revolutions of deep learning, knowledge graphs and reasoning are already having considerable impact, and the fourth revolution of integrated intelligence will likely go further. For understanding AI progress in human terms, these four areas (learning, knowledge, reasoning, agency) are coarser than traditional psychometric taxonomies. However, they provide a useful perspective by focusing on sufficiency for tasks, which is directly relevant for considering the impact of AI on the future of work. For this purpose, they are incomplete: language and vision capabilities need to be assessed as well, for example.

Still, a level of analysis in terms of task capabilities seems useful. Psychometric tests were designed to assess people. They have been, and remain, one source of useful measures to understand progress in AI. However, today's AI systems differ in many ways from people, and tomorrow's likely will as well. In some cases, these differences will be limitations to be overcome; in others, they will be sought deliberately to build systems that complement human strengths and weaknesses. Consequently, more task-oriented measures will be needed as well.

Are there other AI revolutions to come? It seems likely. While deep learning has received much of the press, other learning techniques are being scaled up and more broadly applied as well. Therefore, more revolutions in learning may emerge.

What about natural language? Statistical learning and knowledge graphs have been powering progress on natural language, but there are limits to language on its own. For example, approaches that try to build dialogue systems based on large corpora of human interactions often ignore the critical role of context in dialogue. For example, the meaning of "What shall we do next Thursday?" must be determined in the context of the current conversation. This means that, beyond a certain level, progress in natural language is bound up with progress in integrated intelligences, which provide context for understanding.

What about robotics? There, issues of materials and mechanical engineering are key bottlenecks, making it much harder to estimate progress.

References

Allemang, D., J. Hendler and F. Gandon (2020), *Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL*, Third edition, ACM Books, New York. [20]

Anderson, J. (2009), *How can the Human Mind Occur in the Physical Universe?*, Oxford University Press. [31]

Andreas, J. et al. (2020), "Task-oriented dialogue as dataflow synthesis", *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 556-571, http://dx.doi.org/10.1162/tacl_a_00333. [22]

Bengio, Y. et al. (2015), "Towards biologically plausible deep learning", *arXiv*, Vol. 1502.04156v2, https://arxiv.org/abs/1502.04156. [2]

Davis, E. and G. Marcus (2015), "Commonsense reasoning and commonsense knowledge in artificial intelligence", *Communications of the ACM*, Vol. 58/9, pp. 92-103, https://doi.org/10.1145/2701413. [24]

Distefano, D. et al. (2019), "Scaling static analyses at Facebook", *Communications of the ACM*, Vol. 62/8, pp. 62-70, http://dx.doi.org/10.1145/3338112. [26]

Dong, L. (2018), "Challenges and innovations in building a product knowledge graph", presentation, January, Paul Allen School of Science and Engineering, University of Washington, Seattle, https://db.cs.washington.edu/events/database_day/2018/slides/luna_productgraph.pdf. [19]

Engel, P. (2008), "Tacit knowledge and visual expertise in medical diagnostic reasoning: Implications for medical education", *Medical Teacher*, Vol. 307/7, pp. 184-188, http://dx.doi.org/10.1080/01421590802144260. [17]

Eykholt, K. et al. (2018), "Robust physical-world attacks on deep learning visual classification", *arXiv*, Vol. 08945, http://dx.doi.org/arXiv:1707.08945. [5]

Fan, J. et al. (2012), "Automatic knowledge extraction from documents", *IBM Journal of Research and Development*, Vol. 56/3/4, pp. 5:1-5:10, http://dx.doi.org/10.1147/JRD.2012.2186519. [23]

Forbus, K. (2019), *Qualitative Representations: How People Reason and Learn about the Continuous World*, MIT Press, Cambridge, MA. [14]

Forbus, K. and T. Hinrichs (2017), "Analogy and qualitative representations in the companion cognitive architecture", *AI Magazine*, Vol. 38/4, pp. 34-42, https://doi.org/10.1609/aimag.v38i4.2743. [33]

Forbus, K., C. Liang and I. Rabkina (2017), "Representation and computation in cognitive models", *Topics in Cognitive Science*, Vol. 9/3, pp. 694-798, http://dx.doi.org/DOI:10.1111/tops.12277. [11]

Friedman, S. and K. Forbus (2010), "An integrated systems approach to explanation-based conceptual change", *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, GA*, Vol. 24/1, https://ojs.aaai.org/index.php/AAAI/article/view/7572. [34]

Fromherz, M., D. Bobrow and J. de Kleer (2003), "Model-based computing for design and control of reconfigurable systems", *AI Magazine*, Vol. 24/4, https://doi.org/10.1609/aimag.v24i4.1735. [29]

Gentner, D. and A. Stevens (eds.) (1983), *Mental Models*, Lawrence Erlbaum Associates, Hillsdale, NJ. [15]

Gil, Y. and B. Selman (eds.) (2019), "A 20-year community roadmap for artificial intelligence research in the US", *Workshop Report*, Computing Community Consortium (CCC) and the Association for the Advancement of Artificial Intelligence, http://dx.doi.org/arXiv:1908.02624. [30]

Gluck, K. and J. Laird (eds.) (2019), *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions*, MIT Press, Cambridge, MA. [13]

Goodfellow, I., J. Shlens and C. Szegedy (2015), "Explaining and harnessing adversarial examples", *Proceedings of International Conference on Learning Representations 2015*, http://dx.doi.org/arXiv:1412.6572. [4]

Jalali, V. and D. Leake (2018), "Harnessing hundreds of millions of cases: Case-based prediction at industrial scale: 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9-12, 2018, Proceedings", in Cox, M., P. Funk and S. Begum (eds.), *Case-Based Reasoning Research and Development. ICCBR 2018. Lecture Notes in Computer Science*, Springer, Cham, http://dx.doi.org/10.1007/978-3-030-01081-2_11. [21]

Jia, R. and P. Liang (2017), "Adversarial examples for evaluating reading comprehension systems", *arXiv*, Vol. 07328, http://dx.doi.org/arXiv:1707.07328. [7]

Kaushik, D. and Z. Lipton (2018), "How much reading does reading comprehension require? A critical investigation of popular benchmarks", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5010-5015, http://dx.doi.org/10.18653/v1/D18-1546. [8]

Klein, G. (1999), *Sources of Power*, MIT Press, Cambridge, MA. [16]

Laird, J. (2012), *The SOAR Cognitive Architecture*, MIT Press, Cambridge, MA. [32]

LeCun, Y., Y. Bengio and G. Hinton (2015), "Deep learning", *Nature*, Vol. 512, pp. 436-444, https://doi.org/10.1038/nature14539. [1]

Lenat, D. and P. Durlach (2014), "Reinforcing math knowledge by immersing students in a simulated learning by teaching experience", *International Journal of Artificial Intelligence in Education*, Vol. 24, pp. 216-250, https://doi.org/10.1007/s40593-014-0016-x. [36]

Lenat, D. et al. (2010), "Harnessing Cyc to answer clinical researchers' ad hoc queries", *AI Magazine*, Vol. 31/3, pp. 13-32, http://dx.doi.org/10.1609/aimag.v31i3.2299. [35]

Marcus, G. (2018), "Deep learning: A critical appraisal", *arXiv*, Vol. 00631, http://dx.doi.org/arXiv:1801.00631. [3]

Marcus, G. and E. Davis (2020), "GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about", 22 August, MIT Technology Review. [12]

Newcombe, C. et al. (2015), "How Amazon web services uses formal methods", *Communications of the ACM*, Vol. 58/4, pp. 66-73, http://dx.doi.org/10.1145/2699417. [25]

Nguyen, A., J. Yosinski and J. Clune (2015), "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015*, pp. 427-436, http://dx.doi.org/10.1109/CVPR.2015.7298640. [6]

Noy, N. et al. (2019), "Industry-scale knowledge graphs: Lessons and challenges", *acmqueue*, Vol. 17/2, https://queue.acm.org/detail.cfm?id=3332266. [18]

Price, C. (2000), "AutoSteve: Automated electrical design analysis", *ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, Germany, 20-25 August 2000*, https://www.dbai.tuwien.ac.at/event/ecai2000-kbsmbe/papers/w31-10.pdf. [28]

Reeves, B. and C. Nass (2003), *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*, CSLI Publications, Stanford. [37]

Silver, D. et al. (2016), "Mastering the game of Go with deep neural networks and tree search", *Nature*, Vol. 529, pp. 484-489, https://doi.org/10.1038/nature16961. [9]

Silver, D. et al. (2018), "A general reinforcement learning algorithm that masters chess, shogi, and Go", *Science*, Vol. 362/6419, pp. 1140-1144, http://dx.doi.org/0.1126/science.aar6404. [10]

Simonis, H. (2001), "Building industrial applications with constraint programming", in Comon, H., C. Marche and R. Treinen (eds.), *Constraints in Computational Logics: Theory and Applications*, Springer, Lecture Notes in Computer Science. [27]

Tunstall-Pedoe, W. (2010), "True knowledge: Open-domain question answering using structured knowledge and inference", *AI Magazine*, Vol. 31/3, pp. 80-92, https://doi.org/10.1609/aimag.v31i3.2298. [38]

Notes

[1] AlphaZero is more data-efficient, but the 21 million games played during training would still require a person to live close to 4 800 years.

[2] They performed over 150 tests to come to this conclusion, which are described at https://cs.nyu.edu/faculty/davise/papers/GPT3CompleteTests.html. A common problem with the evaluation of AI systems is what is called the *Eliza effect*, i.e. that people are easily fooled by AI systems. People's tendency to humanise technologies has been amply documented (Reeves and Nass, 2003[37]).

[3] Another common term is *knowledge base*. Typically, these terms are used interchangeably, but there is a subtle difference. Knowledge graphs are always constructed out of graphs, i.e. node and link data-structures. While all knowledge graphs are knowledge bases, some approaches try to use distributed representations, e.g. statistical language models, as knowledge bases.

[4] Amazon acquired True Knowledge (Tunstall-Pedoe, 2010[38]) and has been building on it for general question-answering services.

[5] The BBC, for example, has made theirs available on line at www.bbc.com/ontologies.

[6] While many knowledge graphs contain URLs for images, there is little processing of such information to use it for reasoning. They typically display it to users to show what a product looks like.

[7] Vector representations consist of a one-dimensional sequence of numbers. Vector, matrix and tensor mathematical representations are commonly used in neural network models.

[8] Relational knowledge is expressed in the form of statements involving predicates and arguments. Statements are analogous to a verb and its arguments in a natural language sentence, but the predicates are drawn from a formal vocabulary and their meaning is constrained by rules of inference. Logics and procedural formalisms are two examples of representational systems that can be used to encode relational knowledge.

[9] Published examples include answering queries integrated across multiple databases by medical researchers expressed in natural language (Lenat et al., 2010[35]) and intelligent tutoring for mathematics (Lenat and Durlach, 2014[36]).

[10] A roadmap for an Open Knowledge Network is laid out in Gil and Selman (2019[30]), as part of a broader 20-year roadmap for AI in the United States.

[11] This example is correctly handled by Google, Bing, Alexa and Bixby at this writing, but they do not handle all such simple two step inferences.

[12] This example is due to the late Marvin Minsky.

[13] Research on qualitative reasoning has done the most to formalise model formulation (Forbus, 2019[14]), but there is still much more research needed on this topic.

[14] Imagine AI assistants for doctors, lawyers, and engineers that accumulate experiences as they work with people. Now imagine that experience is being accumulated across assistants within an organisation

or within a profession, combined into a massive library of experience that would go beyond what any person could ever experience.

[15] Typical venues include the journal and conference *Advances in Cognitive Systems* (www.cogsys.org) and the Cognitive Systems track at the AAAI conference.

[16] ACT-R, SOAR and Companions have been used in deployed systems but not yet at a scale or frequency to constitute a revolution in the sense used here. On the other hand, today's deployed systems may lead to more experiments in autonomy. In an apocryphal story, an apprentice too lazy to manually supervise the steam engine invented the governor to do it for him. Similarly, there may be stirrings of autonomy in today's AI pipelines as engineers find ways to let the system take on more of their work, but that is operating at the margins.

[17] In Gil and Selman (2019[30]), the Integrated Intelligence workshop report is mostly concerned with such issues [e.g. expanding the kinds of memories available to AI systems (page 25) and reducing the maintenance footprint of AI systems (page 26)]. The Meaningful Interaction workshop report addresses multiple relevant issues, including the integration of diverse natural modalities for communication and challenges involved in collaboration with people. The Self-Aware Learning workshop report addresses both expanding the kinds of knowledge that learning techniques can handle (page 66) and controlling their own learning (i.e. durable machine learning systems, page 74).