AUTOMATING VERSUS AUGMENTING INTELLIGENCE*

William B. Rouse and James C. Spohrer

Abstract

This article addresses the prospects for automating intelligence versus augmenting human intelligence. The evolution of artificial intelligence (AI) is summarized, including contemporary AI and the new capabilities now possible. Functional requirements to augment human intelligence are outlined. An overall architecture is presented for providing this functionality, including how it will make deep learning explainable to decision makers. Three case studies are addressed, including driverless cars, medical diagnosis, and insurance underwriting. Paths to transformation in these domains are discussed. Prospects for innovation are considered in terms of what we can now do, what we surely will be able to do soon, and what we are unlikely to ever be able to do.

INTRODUCTION

The idea that intelligence can be automated, replacing millions of humans in routine jobs, has received an enormous amount of attention, e.g., (Brynjolfsson, & McAfee, 2014; Beyer, 2016; Auerswald, 2017). Various pundits have projected dramatic disruptions of the economy as robots, or equivalent, pervasively provide an increasing range of services. There has been considerable debate about the extent to which completely "hands-off" automation will be possible and how legal issues will be addressed.

Undoubtedly, there are many jobs that involve 100% routine, highly repeatable tasks that will become totally automated. There are many more jobs that are partially routine and partly non-routine that will be amenable to automation that augments humans who are responsible for the non-routine aspects of these jobs. This article addresses the ways in which human intelligence in such situations can be augmented rather than replaced.

First, consider several observations about contemporary artificial intelligence. Machine learning – or deep learning -- applications have demonstrated impressive capabilities to perform tasks such as recognizing pictures and speech, detecting anomalous behaviors, and other pattern-oriented functions. The neural network algorithms underlying machine learning are composed of multiple layers involving both linear and non-linear transformations. Conclusions reached by machine learning are, in general, not explainable in the sense that the computational system cannot explain why it is making particular recommendations.

* To appear in the ***Journal of Enterprise Transformation***.

The implications are fairly clear. To the extent that decisions emanating from machine learning are always 100% correct, then the action systems, human or otherwise, can simply execute the recommended decisions. If recommendations will occasionally be rejected, or should be rejected, then the lack of explanation capabilities will impose responsibilities on humans that will require decision support. This suggests the need for an intelligent interface layer between the machine learning capabilities and the action systems, particularly when human decision makers are ultimately responsible for decision making.

The concept we are proposing includes the following overall functionality. An intelligent interface needs to understand human decision makers' intentions and provide support needed for successful pursuit of these intentions. Humans' intentions are very context dependent and change in time depending on external circumstances and the intentions and actions of a range of stakeholders, e.g., other drivers, patients, customers, and competitors. Consequently, an underlying time-varying workflow model is required that provides explicit representation of humans' goals, plans, and scripts, as well as information and control requirements. These notions come together in an approach to augmenting intelligence that is elaborated in this article.

We proceed as follows. First, we summarize the evolution of artificial intelligence (AI). Then, we discuss contemporary AI and the new capabilities now possible. This leads to consideration of functional requirements to augment human intelligence. We then present an overall architecture for providing this functionality, including how it will make deep learning explainable to decision makers. Three case studies are addressed – driverless cars, medical diagnosis, and insurance underwriting. Paths to transformation in these domains are discussed. The article concludes by considering prospects for innovation in terms of what we can now do, what we surely will be able to do soon, and what we are unlikely to ever be able to do.

EVOLUTION OF AI

One could argue that AI began with Ada Lovelace in the mid 1800s (Isaacson, 2014, Auerswald, 2017). However, many would agree that the field began in earnest with Alan Turing's landmark paper (Turing, 1950). His article on the Imitation Game unveiled his test for a machine's ability to exhibit intelligent behavior. It has remained an important philosophical construct with AI.

The emerging field of Artificial Intelligence (AI) was recognized at the Dartmouth College AI Conference in 1956 led John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester. Marvin Minsky's PhD thesis (Minsky, 1954) led 15 years later to his book, with Seymour Papert, on perceptrons (Minsky & Papert, 1969). Frank Rosenblatt's work on perceptrons appeared soon after Minsky's thesis (Rosenblatt, 1957). Allen Newell, John Shaw, and Herbert Simon published work on the General Problem Solver (1959).

Even this early, differences of approaches were clear. Perceptrons were based on statistical methods for pattern recognition. This approach foreshadowed the success of multiple layer networks of today's deep learning systems that require tremendous computing power and data sets no available in those early days. Symbolic logic was adopted for problem solving or reasoning. This approach presaged the rise of expert systems, and the challenges of manually building and maintaining large rule-based knowledge systems. Of course, recognizing an object with a network model is a rather different task from making a sequence of tests to troubleshoot an electronic circuit with a rule-based knowledge system.

In the 1960s, Joseph Weizenbaum introduced Eliza (1966), which simulated "conversation" by matching patterns and substituting key words that gave users an illusion of understanding, despite the computer having no means for understanding the context of the dialog. A richer approach to language was Roger Schank's Conceptual Dependency Model (1969), which eventually led to major contributions to natural language understanding (Schank & Abelson, 1977).

The 1970s saw applications of AI to enhance medical diagnosis and treatment, starting perhaps with MYCIN (Shortliffe & Buchanan, 1975). However, a report by James Lighthill (1973) criticized AI for articulating and then failing in its pursuit of grandiose objectives. This report, and other forces, led to the First AI Winter, with substantial DARPA funding cuts.

The 1980s saw the growth of expert systems, led by Edward Feigenbaum (1980). These rule-based systems were built from 'knowledge engineering" with subject matter experts. DARPA's Pilot's Associate's Program emerged to leverage expert systems technology (Banks & Lizza, 1991). Our basic research on intelligent interfaces (discussed below) was funded by a variety of agencies; this DARPA program provided the means to bring the pieces together.

John Searle (1980), a philosopher, introduced the Chinese Room during this period. A non-Chinese speaker is locked in a room and provided various rules for translating Chinese stories into English. He argued that while one might think one was communicating to a Chinese person, one is just communicating with an algorithm. This was yet another critique of AI not being capable of creating a mind.

The late 1980s saw the Second AI Winter. The Lisp machine market collapsed. Japan's 5th Generation Project fizzled. DARPA funding cuts happened again. This period saw computing move from Lisp machines to Sun engineering workstations to desktop PCs.

In the 1990s, there were several real applications. In 1991, the ISX Corporation, (a spinoff from Teknowledge, founded by Feigenbaum), created and deployed DART (Dynamic Analysis and Replanning Tool). DART was used by the U.S. military in the Middle East to optimize and schedule the transportation of supplies or personnel and solve other logistical problems (Cross, et al., 1994).

In 1997, IBM's Deep Blue defeated chess master Gary Kasparov. The development of this technology, chronicled by (Hsu, 2002), was not intended for only playing chess. Another major demonstration of IBM's capabilities came when Watson won *Jeopardy!* in 2011. Ferrucci and colleagues (2013) describe the development of this technology.

The 2000s also saw the maturation of deep learning first at universities, and then at Microsoft, Google and other companies (Hof, 2013). Deep learning is discussed in some depth in the next section. It is nevertheless worth noting here a trend over the 60+ years since Turing. Early innovations were associated with people, often individuals at universities. Universities helped create large open datasets and competitions that led to measurable progress and accumulation of knowledge. Teams, often working at large companies, have accomplished later innovations, often with far more computing resources.

Another trend is also of note. Early research focused on two dominant approaches: statistically based learning for pattern recognition, and rules based or symbolic logic for problem solving and reasoning. Big data and almost free computing power have allowed enormous advances in machine-based pattern recognition. We argue later that both are still needed to augment intelligence.

CONTEMPORARY AI

AI based on symbolic logic worked where rules and definitions were very clear in domains such as mathematics and chess. However, the symbolic logic approach was overwhelmed by many pattern recognition tasks. AI based on layered neural nets, now termed deep machine learning, has been successful for speech recognition, image recognition, language translation, and driverless cars. Each subsequent layer looks for patterns in the previous level's patterns. Early on this approach was called connectionism or distributed parallel processing.

Trends

Lloyd's (2016) sees the drivers of the AI revolution as economic – these systems make us more productive, mobile, connected, and able to compete in the global world economy. They also improve safety in hazardous environments and for tedious jobs.

Mittal and colleagues (2017) see the driving forces to include exponential data growth, faster distributed systems (networks), and smarter algorithms. They assert that AI will enable new approaches to customer engagement, amplification of employee skills and intelligence, cultivation of new product and service offerings, and exploration of new business models.

Lewis-Kraus (2016) provides an in-depth report on Google's decision to reorganize Google Translate around AI, taking nine months to succeed and, in the process, growing the Google Brain project. They used deep machine learning to enable an enormous improvement in the performance of Google

Translate. Faced with Searle's question of whether the AI really understands the languages it translates (Searle, 1980), the Google team dismissed the question as irrelevant. Instead, they argue that they are on a path to "Artificial *general* intelligence (that) will demonstrate a facility with the implicit, the interpretative."

Formation of The Partnership on AI was recently announced (Partnership, 2017). A network of companies heavily involved in AI has been recruited with Amazon, Apple, Google, Facebook, IBM, and Microsoft as founding members. Recent new members include eBay, Intel, Salesforce, SAP, Sony, and Zolando, as well as 14 non-profit members. A primary goal of this partnership is collaborations around defining frameworks to build and deploy safe and ethical AI systems.

What do these trends portend? Lenartowicz (2015) considers "a genuine, unstoppable intelligence arising from the accelerating interconnectivity of individuals and their technological extensions." However, social boundaries seem to impede this. She suggests a biological explanation for why this happens. Auerswald (2017) vision is less sweeping. He sees a bifurcation of jobs, involving "discontinuous advances in code (that) creates a new low-cost high-volume option and a high-cost, low-volume option." The former are automated; the latter offer new opportunities for augmenting human intelligence.

Limits

Deep learning works well when trained with large numbers of examples, but this is not feasible for many tasks, e.g., reasoning about and solving novel problems. Further, as the Stanford 100 Year Study notes, "No machines with self-sustaining long-term goals and intent have been developed, nor are they likely to be developed in the near future" (Stanford, 2016).

Finding a large number of training examples can be a challenge. *The Economist* (2017) reports on the use of video games to train AI, for example, to recognize particular features of the environment. Relative to actually learning to play these games, AI has difficulty learning games when early events have no meaning until much later in the game. AI also has great difficulty using knowledge from one game to play another game.

Nguyen, Yosinski and Clune (2015) report that deep neural networks are easily fooled, sometimes making high confidence predictions for unrecognizable images, e.g., reporting that a picture of white noise is a lion. This could be a problem if the neural networks are operating autonomously; less so if they are augmenting human intelligence. Coldewey (2017) provides an amusing example of fooling a driverless car.

Kaplan (2015) asks, "How do we assure that (robots) respect the unstated conventions that people unconsciously follow?" He further notes that "Finding the right balance between our personal interests and the needs of others – or society in general – is a finely calibrated human instinct, driven by a sense of fairness, reciprocity, and common interest."

Driverless Cars

The concept of driverless cars has, of late, received enormous attention. The race is among several very serious players to refine and deploy their cars, with potentially disruptive impacts on the economy, e.g., car insurance and auto finance (Liu, 2017). A range of challenges remains.

Bollier (2017) discusses human interaction with driverless cars and asks whether AI can understand human foolishness. In particular, humans in cars and human pedestrians signal each other in ways difficult for AI to sense. And, of course, who is responsible for any harm caused by driverless cars?

Boudette (2016) discusses five things that give self-driving cars headaches including unpredictable humans, disappearing lines on the road (e.g., due to snow), detours and rerouted roads, puddles that might be too deep to cross, and having to make tough decisions, i.e., when a crash is inevitable. Bershidsky (2015) and Kaplan (2015) discuss the last item, as do many commentators.

Sherman (2015) reports that humans drive 3 trillion miles per year in the US, but driverless cars have only been tested for 1+ million miles. It would take decades or more to gain comparable data for driverless cars. People learn to drive differently in different contexts – day versus night, clear weather versus storms, seasonal variations, type of road variations, highway versus city, levels of congestion variations, etc. There have been several accidents when human driven cars hit the driverless cars, typically in the rear end.

Norman (2014) notes that people cannot successfully supervise automation when it involves long periods of doing nothing. They are supposed to be vigilant, but automation complacency inevitably occurs. When they need to intervene, they have no training and just seconds to react. The best solution is human-automation collaboration or teamwork, with humans in charge.

Healthcare

Bollier (2017) sees a primary role for AI being augmentation of the intelligence and skills of physicians. That was seemingly the intent of the MYCIN project four decades ago (Shortliffe & Buchanan, 1975). Much more recently, IBM's Watson has been targeted to enhance diagnosis and treatment (Galeon & Hauser, 2016). No one, however, anticipates replacing physicians with robots.

However, there are some tasks where deep learning could excel. *The Economist* (2016) discusses deep learning for radiology. Such pattern recognition tasks could greatly benefit from deep learning, although the questions remains of whether people will accept computers telling them, for example, that they have cancer. Augmentation seems much more likely than automation.

Auerswald argues that "The greatest advances in the provision of healthcare services will come from a combination of wearable technologies, diagnostics supported by Big Data applications, peer-to-peer operations, and other

innovations in code that distribute healthcare delivery away from the highly centralized models that came to dominate in the twentieth century." (Auerswald, 2017, pp. 160-161)

Journalism

Bollier (2017) sees AI affecting journalism by enabling automated curation and exclusion of information, automated detection of news trends, and automated fact checking. *The Economist* (2016) discusses automation of market reports and sports reporting.

Our experience with automated content aggregation and text analytics suggests that there is no other practical way to digest millions, or even thousands, of publications on topics of interest (Basole, Seuss & Rouse, 2012; Yu, Serban & Rouse, 2013). The state of the art is very impressive (Seuss, 2011). Admittedly, however, the automation needs help from humans, at least initially, to assure that it correctly understands the targeted domain.

Impact

Brancaccio (2017), interviewing Martin Ford, considers the economic and social impacts of AI. Ford notes that technology has historically made us richer. Agricultural workers lost jobs to mechanization, but then moved to factory jobs. However, now it may be different, as automation technology has become pervasive, almost like electricity. Any job that seems routine and boring is at risk of automation. More specifically, Ng (2016) argues that if a human can perform a mental task in less than one second, that task probably can be automated.

Chui, Manyika and Miremadi (2016) see five factors affecting automation including technical feasibility, costs to automate, relative scarcity of skills and costs of workers, benefits of automation beyond labor-cost substitution, and regulatory and social acceptance considerations. They conclude that more than three quarters of predictable physical work is automatable. Retail is most vulnerable to automation; manufacturing is second. In contrast, jobs involving decision making, planning, and creative work are much less vulnerable. Jobs that involve managing and developing people are least vulnerable.

The Economist (2016) notes that, "What determines vulnerability to automation is not so much whether the work is manual or white-collar but whether or not it is routine." For those affected, it needs to be easier for workers to acquire new skills and change jobs.

Another possibility is that jobs will disappear rather than being automated. A recent study (Liu, 2017) found that the car insurance and auto finance industries could be significantly disrupted by cars that seldom have accidents and, due to high costs of technology laden driverless cars, consumers' shift from car ownership to use of car services. Thus, taxi and truck driving jobs may be replaced while insurance and loan-underwriting jobs will simply disappear.

New Perspective

Fairly recently, a new perspective on these issues is emerging (Spohrer & Banavar 2015; Spohrer 2016). We can view the automation-augmentation continuum as involving mixes of two different types of cognitive systems – biological and digital. Each cognitive system can play a range of roles -- tool, assistant, collaborator, coach, and mediator. The progression from cognition tool to cognitive mediator requires cognitive systems with increasingly sophisticated models of tasks, the world, the user, and the institutional context of the interactions. A digital cognitive mediator does not yet exist, but when it does it will be trusted to make decisions on its user's behalf because of the level of sophistication of its model of its user as well as the laws and institutions of society. Such digital cognitive mediator systems will be designed to behave ethically according to the evolving standards of society.

It is easy to imagine digital cognitive systems serving as tools, assistants, and collaborators. Coaching and mediation abilities are evolving. It is also possible for biological cognitive systems to serve as tools and assistants to digital cognitive systems.

An intriguing example is the use of bald eagles to down illegal drones (Booker, 2016). The digital cognitive system monitors the skies for illegal drones and then dispatches the biological cognitive assistants to catch the interlopers and bring them back, for a rewarding chunk of meat. Special talon protectors have been developed to assure the eagles are not injured.

This brings a new perspective to the automation-augmentation continuum. It is no longer a question of what should humans do and what should computers do. The question now concerns creating the best cognitive team or cognitive organization to address the problems at hand. This portends some creatively different solutions from what have been developed in the past.

Cognitive organizations of biological and digital workers will likely evolve rapidly over the next sixty years driven by Moore's Law. Moore's law can be thought of as reducing the cost of computation by a factor of a thousand every twenty years, a million every forty years, and a billion every sixty years. Figure 1 shows the impact of Moore's Law on the cost of digital workers anticipated over the next sixty years. The figure also shows the increase in GDP (Gross Domestic Product/Employee) anticipated as people and organizations can afford to have more and more digital workers working on their behalf.

GDP was computed using a specific query to WolframAlpha website: "gdp/employees USA from 1950 to 2015." This provided eight rounded data points (1950 $7000; 1960 $10,000; 1970 $15,000; 1980 $33,000; 1990 $55,000; 2000 $78,000; 2010 $116,000; 2015 $127,000), that were then used as input to Microsoft Excel curve fitting to project rounded values (2020 $200,000; 2030 $350,000; 2040 $500,000; 2050 $800,000; 2060 $1,250,000; 2070 $2,000,000;

2080 $3,250,000). Again, the point is merely to show that (1) as the cost of digital workers decreases and (2) GPD per employee increases, then the number of digital workers that people and organizations can afford increases exponentially, for some period of time.

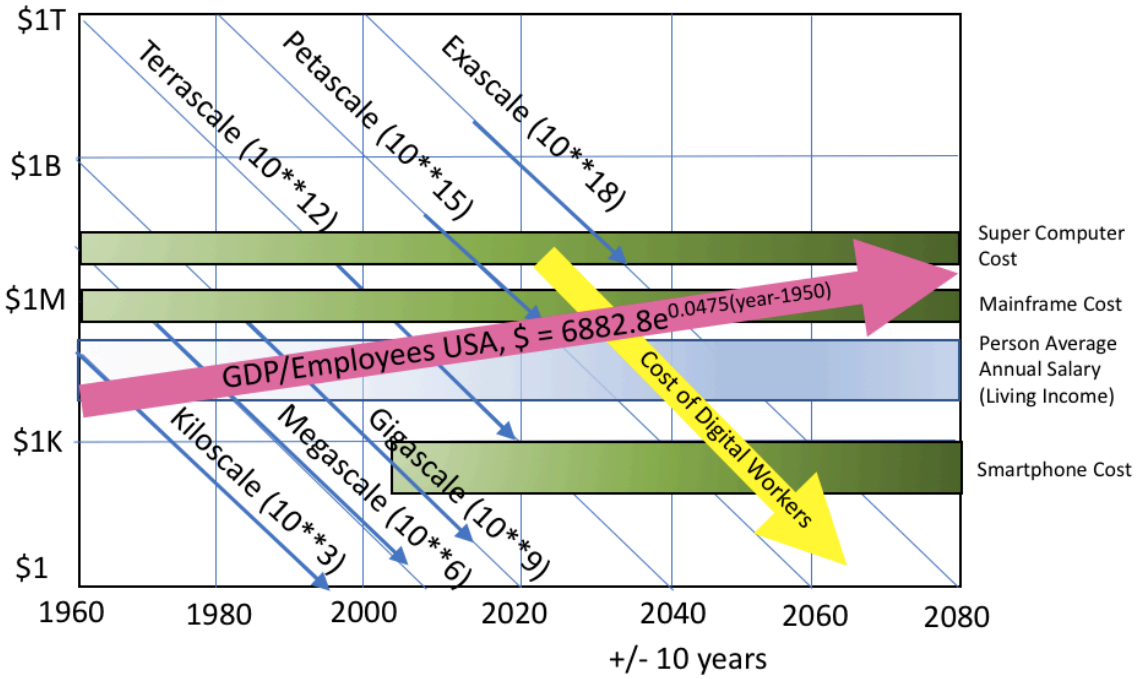

Figure 1: Decreasing Costs of Digital Workers Anticipated From 2020 – 2080, and Increasing GDP/Employee (see service-science.info/archives/4741).

Digital workers are likely to become increasingly intelligent as they move from petascale (narrow AI pattern recognition) to exascale (general or broad AI reasoning) computing capabilities. In 2017, a terascale systems costs about $3K, but twenty years ago such a system would have costs millions. In 2017, a smartphone is a gigascale system. As the cost of digital workers decreases, people and organizations will be able to afford more and more digital workers to work on their behalf, increasing GDP/employees.

Business and governments that have a fiduciary responsibility to shareholders and citizens will likely use an increasing number of digital workers over time to reduce costs. Individuals will also be able to use digital workers that they own on their behalf to lower the cost of family operations as service systems. The implications of cognitive tools, assistants, collaborators, coaches, and mediators as part of smarter and wiser service systems is a new area of exploration for service science (Spohrer, Siddike, & Kohda 2017). Service science studies the evolving ecology of entities with capabilities, constraints, rights, and responsibilities, their value co-creation and capability co-elevation mechanisms. The rapidly dropping cost of digital workers with human-level capabilities will have a dramatic impact on existing service systems, including families,

universities, businesses, and governments – and test if there is a speed limit to progress (Spohrer, Giuiusa, Demirkan & Ing, 2013).

At both microeconomic and macroeconomic scale, technologies that extend human capabilities in one context can also automate and replace them in another context (Markoff, 2016). For example, workers might use technology to perform better at their jobs, then have their jobs go away from a more advanced version of the technology, but then adopt a low-cost version of the technology to help them set up their own business or even continue do the "job" in a do-it-yourself hobby-mode of work. The rise of low cost digital workers may make it easier for the long-tail of hobbies in society to develop into sole-proprietorship businesses (Aubrey, 2010). This trend might be accelerated by basic income guarantees, and a number of nations have begun such experiments (Widerquist & Lewis 2005). Mixes of biological-digital cognitive systems could be facilitated both by low cost digital workers, technology-augmented human resources readily available because of basic income guarantee, and other biological species with augmented capabilities. A wider range of service systems than exist today will likely result.

AUGMENTING INTELLIGENCE

The foregoing sets the stage for our main argument. In many situations, AI will be used to augment human intelligence, rather than being deployed to automate intelligence and replace humans. What functions are needed to augment intelligence?

Information Management

One function will be information management (Rouse, 2007). This involves information selection (what to present) and scheduling (when to present it). Information modality selection involves choosing among visual, auditory, and tactile channels. Information formatting concerns choosing the best levels of abstraction (concept) and aggregation (detail) for the tasks at hand. Artificial intelligence can be used to make all these choices in real time as the human is pursuing the tasks of interest.

Intent Inferencing

Another function is intent inferencing (Rouse, 2007). Information management can be more helpful if it knows both what humans are doing and what they intend to do. Representing humans' task structure in terms of goals, plans, and scripts (Schank & Abelson, 1977) can enable making such inferences. Scripts are sequences of actions to which are connected information and control requirements. When the intelligence infers what you intend to do, it then knows what information you need and what controls you want to execute it.

One of the reasons that humans are often included in systems is because they can deal with ambiguity and figure out what to do. Occasionally, what they

decide to do has potentially unfortunate consequences. In such cases, "human errors" are reported. Errors in themselves are not the problem. The consequences are the problem.

Error Tolerant Interfaces

For this reason, another function is an error tolerant interface (Rouse & Morris, 1987; Rouse, 2007). This requires capabilities to identify and classify errors, which are defined as actions that do not make sense (commissions) or the lack of actions (omissions) that seem warranted at the time. Identification and classification lead to remediation. This occurs at three levels: monitoring, feedback, and control. Monitoring involves collection of more evidence to support the error assessment. Feedback involves making sure the humans realize what they just did. This usually results in humans immediately correcting their errors. Control involves the automation taking over, e.g., applying the brakes, to avoid the imminent consequences.

Adaptive Aiding

The notion of taking control raises the overall issue of whether humans or computers should perform particular tasks. There are many cases where the answer is situation dependent. Thus, this function is termed adaptive aiding (Rouse, 1988, 2007). The overall concept is to have mechanisms that enable real time determination of who should be in control. Such mechanisms have been researched extensively, resulting in a framework for design that includes principles of adaptation and principles of interaction. A First Law of Adaptive Aiding has been proposed – *computers can take tasks, but they cannot give them*.

Intelligent Tutoring

Another function is intelligent tutoring to both train humans and keep them sufficiently in the loop to enable successful human task performance when needed. Training usually addresses two questions: 1) How the system works and, 2) How to work the system. Keeping humans in the loop addresses maintaining competence. Unless tasks can be automated to perfection, humans' competencies need to be maintained. Not surprisingly, this often results in training vs. aiding tradeoffs, for which guidance has been developed (Rouse, 2007).

Example Applications

Many of the earlier research and applications of the notions elaborated in this section focused on operation and maintenance of complex engineered systems such as aircraft, power plants, and factories. The tasks associated with such systems are usually well understood. One application focused on electronic checklists for aircraft pilots (Rouse & Rouse, 1980; Rouse, Rouse & Hammer,

1982; Rouse, 2007). The results were sufficiently compelling to motivate inclusion of some of the functionality on the Boeing 777 aircraft.

A conceptual architecture for intelligent interfaces has been developed and applied several times to tasks that are sufficiently structured to be able to make the inferences needed to support the functionality outlined here (Rouse, Geddes & Curry, 1988; Rouse, Geddes & Hammer, 1990; Rouse, 2007). The notion of augmented intelligence can build on this foundation, with some important extensions due to advances in contemporary AI.

OVERALL ARCHITECTURE

Figure 2 provides an overall architecture for augmenting intelligence. The intelligent interface, summarized above, becomes a component in this broader concept. The overall logic is as follows:

- Humans see displays and controls, and decide and act. Humans need not be concerned with other than these three elements of the architecture. The overall system frames human's roles and tasks, and provides support accordingly.
- The intent inference function infers what task(s) humans intend to do. This function retrieves information and control needs for these task(s). The information management function determines displays and controls appropriate for meeting information and control needs
- The intelligent tutoring function infers humans' knowledge and skill deficits relative to these task(s). If humans cannot perform the task(s) acceptably, the information management function either provides just-in-time training or informs adaptive aiding (see below) of the humans' need for aiding.
- Deep learning neural nets provide recommended actions and decisions. The explanation management function provides explanations of these recommendations to the extent that explanations are requested. This function is elaborated below.
- The adaptive aiding function, within the intelligent interface, determines the human's role in execution. This can range from manual to automatic control, with execution typically involving somewhere between these extremes. The error monitoring function, within the intelligent interface, detects, classifies and remediates anomalies.

Note that these functions influence each other. For example, if adaptive aiding determines that humans should perform task(s), intelligent tutoring assesses availability of necessary knowledge and skills, and determines training interventions needed, and information management provides the tutoring experiences to augment knowledge and skills. On the other hand, if adaptive aiding determines that automation should perform task(s), intelligent tutoring assesses humans' abilities to monitor automation, assuming such monitoring is needed.

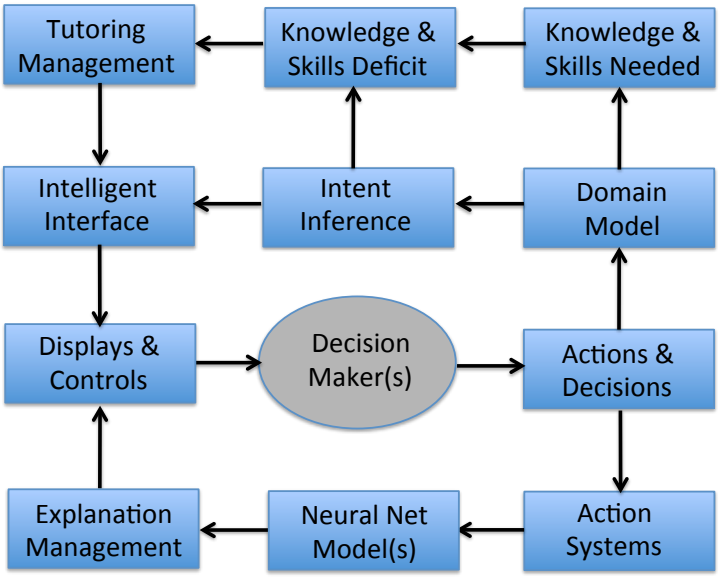

Figure 2. Overall Architecture of Augmented Intelligence

Explanation Management

As discussed in the Introduction, neural network models cannot explain their (recommended) decisions. This would seem to be a fundamental limitation. However, science has long addressed the need to understand systems that cannot explain their own behaviors. Experimental methods are used to develop statistical models of input-output relationships. Applying these methods to neural network models can yield mathematical models that enable explaining the (recommended) decisions as shown in Figure 3.

Given a set of independent variables **X**, a statistical experiment can be designed, e.g., a fractional factorial design, that determines the combinations of values of **X** to be input to the neural net model(s). These models, typically multi-layered, have "learned" from exposure to massive data lakes with labeled instances of true positives, and possibly false positives and false negatives. True negatives are the remaining instances.

The neural net models yield decisions, **D**, in response to the designed combinations of **X**. A model **D(X)**, is then fit to these input-output data sets. Explanation generation then yields explanations **E(D)** based on the attributes and weights in the fitted model. The result is a first-order, i.e., non-deep, explanation of the neural net (recommended) decisions.

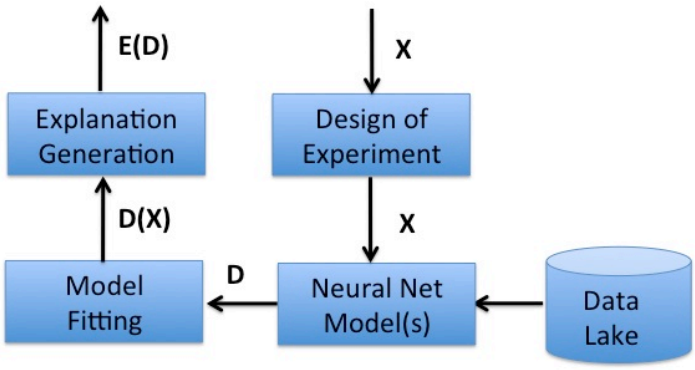

Figure 3. Explanation Management Function

As noted earlier, the paradigm underlying Figure 3 is the standard paradigm of empirical natural science. Thus, it is clear it will work, i.e., yield rule-based explanations, but will it be sufficient to help decision makers understand and accept what the machine learning recommends? We imagine this will depend on the application.

As an example, consider control theory. Optimal stochastic control theory includes both optimal estimation and optimal control. Determining the optimal solution across both estimation and control involves rather sophisticated mathematics. We could apply the method in Figure 3 to the optimal control actions resulting from the solution of this stochastic control problem.

We would not be able to infer the nature of the underlying sophisticated mathematics. Instead, we would likely unearth something akin to classic PID controllers, where the acronym stands for proportional, integral, and derivative attributes of the errors between desired and actual states. It has been shown that this provides a reasonable explanation of optimal control actions.

Learning Loops

Figures 2 and 3 include both explicit and implicit learning loops. The statistical machine-learning loop will be continually refining the relationships in its layers, either by supervised learning or reinforcement learning. This will involve balancing exploration (of uncharted territory) and exploitation (of current knowledge). This may involve human designers and experimenters not included in Figures 2 and 3. Of particular interest is how machine learning will forget older

data and examples that are not longer relevant, e.g., a health treatment that has more recently been shown to be ineffective.

The rule-based learning loops in Figures 2 and 3 are concerned with inferring rule-based explanations of the recommendations resulting from machine learning (Figure 3) and inferring human decision makers' intentions and state of knowledge (Figure 2). Further, learning by decision makers is facilitated by the tutoring function in Figure 2.

Thus, the AI will be learning about phenomena, cues, decisions, actions, etc. in the overall task environment. The decision makers will learn about what the AI is learning, expressed in more readily understandable rule-based forms. The intelligent support system will be learning about the decision makers' intentions, information needs, etc., as well as influencing what the decision makers learn.

THREE SCENARIOS

Given the foregoing discussion of intelligent interfaces and an architecture for augmenting intelligence, this section illustrates additional challenges in creating and deploying such systems.

Human Interactions With Driverless Cars

The dominant perspective on driverless cars is that these technology-intensive vehicles will be sufficiently expensive that most people will be reluctant to buy them and instead will use car services for their transportation needs (Liu, 2017). Thus, it will be like using Uber or Lyft without a human driver. In theory at least, such services will be flawless, economically benefit both passengers and society, and will completely eliminate accidents.

This may be true eventually, but the transition will be extended over decades. During that transition, there will be flaws in the service and occasional accidents. Humans will have to intervene or least want to intervene to ask the question, "Why are you going this way; that's not where I want to go?" How can a driverless car explain itself in response to this question?

We outlined a computational approach in the previous section. The explanations potentially feasible with this approach will have to be integrated into an overall customer experience that allows for different languages and varying levels of comfort with technology. In other words, the driverless car will need an understanding of the passenger, not just alternative routes between A and B.

Medical Diagnosis & Treatment

There have long been prognostications that AI can enhance medical diagnosis and treatment, starting perhaps with MYCIN (Shortliffe & Buchanan, 1975). Much more recently, IBM's Watson has been targeted to enhance diagnosis and

treatment (Galeon & Hauser, 2016). Undoubtedly, this technology will continually improve.

However, will it ever replace human diagnosticians? Perhaps it will in tasks involving complex pattern recognition. However, there is a difference been a radiology scan and the patient, between the sensed pattern and the whole human. This could lead to questions like, “Why do you expect this intervention protocol will succeed? This patient has had a negative reaction to elements of this in the past.

I can imagine the system responding that it was unaware of the previous effects of the intervention in question. Certainly, a lack of complete knowledge is not uncommon, e.g., Boodman (2017). Given a vast amount of information, an AI system can probably digest the full corpus better than a human, but humans are very good at looking at a result and concluding that it does not make sense. The implication is that the explanation management function in Figures 1 and 2 will need capabilities for dialog with the humans it is augmenting.

Insurance Underwriting

Insurance underwriting is an important part of the process associated with any insurance application. When someone applies for insurance coverage, they are requesting the insurance company to assume the potential risk of having to pay a claim in the future. The level of risk assessed determines the premium charged.

The factors considered depend on the type of insurance. For example, age, type of car, and driving record affect risks for auto insurance; age, use of tobacco and alcohol and health records affect risks for life insurance. Insurance companies have enormous data sets that they use to project such risks. With the maturity of machine learning, automated insurance underwriting seems to be quickly maturing as well (Batty & Kroll, 2009).

It is easy to imagine automated and human underwriting conflicting. Upon seeing the AI recommendation, the human might respond, “Why are you pricing this enormous risk so cheaply? This customer has a questionable history of claims.” Since the AI recommendation is based on much more data than the human could ever digest, it is quite likely that, in this case, the human is wrong.

The explanation management function can provide an answer in terms of the data sets used to inform choices and the relative importance of different elements of these data sets. However, what if the human simply disagrees? The intelligent tutoring function might play a role to adjudicate this conflict. Error monitoring might get involved if the human insists on making a bad decision. This runs the risk that the human will feel that they have no choice but agree with the AI. That perception could undermine the symbiosis intended.

Paths to Transformation

How will AI transform enterprises? We expect that two scenarios – incremental change and breakthrough change -- will bound the course of transformation. Incremental change is the norm and is clearly evident for healthcare (Accenture, 2017), automobiles (Plungis, 2017; Liu, 2017), and insurance (Batty & Kroll, 2009). For example, the sensing and control technologies that will enable driverless cars are being increasingly deployed on current vehicles. The last steps to automated driving may not seen so radical after people have become comfortable with the ongoing stream of technological innovations.

In healthcare, AI and analytics are increasingly augmenting many tasks. We do not, by any means, expect this will incrementally lead to clinician-less healthcare. Instead, clinicians will increasingly value and rely upon augmentation, while retaining the central role of clinician-patient interaction. The range of data, information, and knowledge available to support these interactions will continually grow. For example, "precision oncology" will increasingly enable tailoring treatment to individual cancer patients (Grossmueller, 2017). As another example, clinical decision support will help identify patients in need of advanced heart failure therapies (Evans, et al., 2017)

Breakthrough change is less common. Innovations like electricity took many decades before it was available to the majority of citizens in the US. Radio and television were adopted much more quickly because the electrical infrastructure was in place. Similarly, wireless communications networks slowly became pervasive, enabling much more rapid adoption of portable digital devices, epitomized by the iPhone.

Infrastructure dependencies have an enormous effect. The physical infrastructure needed for driverless cars will hinder pervasive change for decades. The integrated information infrastructure needed to transform healthcare delivery is currently being pursued, which should enable much faster adoption of AI augmented clinician and patient support systems.

The information infrastructure needed to enable augmented insurance underwriting is much more under the control of individual companies. This will enable much faster change. Such change is more likely to be on the breakthrough end of the continuum. It is easy to imagine a single human underwriter managing a team of AI underwriters, both to spot anomalies and to continually train these team members.

What are the workforce implications, as AI inventions become market innovations? There is a wide range of commentators on this question. There seems to be agreement that many jobs currently performed by humans will disappear, e.g., routine cognitive jobs (Cross, 2017; Paquette, 2017). Non-routine jobs, and those requiring non-repetitive physical dexterity, are less likely to be automated (Englebert & Hagel, 2017). Jobs involving designing, developing, and managing AI systems are already experiencing very strong demand.

Arthur (2017) argues that this will eventually result in a transition from a production economy to a distributive economy. We will have all the products and services we need without employing all of our citizenry. We will then have to concern ourselves with the rationale and finances for the distribution of goods and services to those not involved in production. This could be a blessing or curse, depending on how creatively we address it.

PROSPECTS FOR INNOVATION

Many impressive innovations have been developed and deployed over the past 20-30 tears. A variety of valuable innovations are soon to come. However, there are quite a few abilities that we do not see computers gaining in the foreseeable future.

What Can We Do Now?

AI is well developed and has matured to accomplish many tasks:

- Retrieving, aggregating and analyzing large numeric, alphabetic, and image data sets
- Recognizing pictures and speech, recognizing patterns in large data sets, both with training
- Problem solving for well-structured tasks, e.g., mathematics, puzzle-like games, troubleshooting
- Robotic storage, retrieval, and manipulation of materials in well-structured environments such as warehouses and factories

This is not an inconsequential set of abilities and applications; they have already revolutionized the workplace.

What Will We Surely Be Able to Do Soon?

Several capabilities are in development and evaluation and are likely to soon be commercially viable:

- Recognizing and analyzing pictures, voice, and video in noisy environments and ambiguous situations
- Understanding and generating natural language, but perhaps not in informal casual situations, e.g., will not deal well with “How was your weekend?”
- Flawlessly driving vehicles within well-defined environments supported by relevant infrastructure
- Summarizing all that is known about a particular topic, e.g., causes and treatments of diabetes

These abilities represent significant advances over what we can do now.

What Are We Unlikely to Ever Be Able to Do?

Several human capacities will be difficult to realize in the foreseeable future:

- Computers taking responsibility for things for which they were not designed and are not responsible, e.g., the factory worker who is injured
- Computers having consciousness and being capable of reflection, e.g., regarding their own capabilities and responsibilities
- Computers having feelings, especially feelings rooted in suffering, without faking them, e.g., experiencing a good, bad, happy, sad, relaxing or stressful day
- Computers enjoying seeing a friend's face, or the feeling of rain or snow, or the smell of fresh cut grass or wood burning in a fireplace

These essentially human capacities very much relate to our being animals in a physical, behavioral, and social world.

CONCLUSIONS

This article summarized the evolution of artificial intelligence (AI), including contemporary AI and the new capabilities now possible. This led to consideration of functional requirements to augment human intelligence. An overall architecture was presented for providing this functionality, including how it will make deep learning explainable to decision makers. Three case studies were addressed – driverless cars, medical diagnosis, and insurance underwriting. Prospects for innovation were considered in terms of what we can now do, what we surely will be able to do soon, and what we are unlikely to ever be able to do.

What can be concluded from the perspective presented in this article? Very pragmatically, we might argue for automating routine, regular tasks; and augmenting non-routine, irregular tasks. Either way, given prevailing values and norms, we have to conclude that responsibility, for both automation and augmentation, remains with humans. This responsibility includes deciding what to work on next (Brynjolfsson & McAfee, 2017):

"We think the biggest and most important opportunities for human smarts in this new age of super powerful ML lie at the intersection of two areas: figuring out what problems to work on next, and persuading a lot of people to tackle them and go along with the solutions."

The humans are likely to always retain decisions to take responsibility for the consequences of agreeing or disagreeing with other people on what is important to do next together.

REFERENCES

Accenture (2017). ***Artificial Intelligence: Healthcare's New Nervous System***. New York: Accenture Corporation.

Arthur, W.B. (2017). Where is technology taking the economy? ***McKinsey Quarterly***, Fall, 33-43.

Aubrey, S.B. (2010). ***The Profitable Hobby Farm: How to Build a Sustainable Local Foods Business***. New York: Wiley.

Auerswald, P.E. (2017). ***The Code Economy: A Forty-Thousand-Year History***. New York: Oxford University Press.

Banks, S., & Lizza, C.S. (1991). Pilot's Associate: A cooperative, knowledge-based system application. ***IEEE Expert: Intelligent Systems and Their Applications***, 6 (3), June, 18-29.

Basole, R.C., Seuss, C.D., & Rouse, W.B. (2012). IT innovation adoption by enterprises: Knowledge discovery through text analytics. ***Decision Support Systems***, 54 (2), 1044-1054.

Batty, M., & Kroll, A. (2009). ***Automated Life Underwriting: A Survey of Life Insurance Utilization of Automated Underwriting Systems***. New York: Deloitte.

Bershidsky, L. (2015). Self-driving cars can't handle moral choices. ***Bloomberg View***, May 10.

Beyer, D. (Ed.).(2016). ***The future of machine intelligence: Perspectives from leading practitioners***. Sebastopol, CA: O'Reilly Media.

Bollier, D. (2017). ***Artificial Intelligence Comes of Age: The Promise and Challenge of Integrating AI Into Cars, Healthcare, and Journalism***. Washington, DC: The Aspen Institute.

Boodman, S.G. (2017). A dog bite sent him to the ER: A cascade of missteps nearly killed him. ***The Washington Post***, June 16.

Booker, C. (2016). Dutch police use eagles to hunt illegal drones. ***PBS NewsHour***, September 16.

Boudette, N.E. (2016). Five things that give self-driving cars headaches. ***New York Times***, June 4.

Brancaccio, D. (2017). The coming age of robo-everything. ***Marketplace***. Washington, DC: National Public Radio, April 3.

Brynjolfsson, E., & McAfee, A. (2014) ***The second machine age: Work progress, and prosperity in a time of brilliant technologies***. New York: Norton.

Brynjolfsson, E., & McAfee, A. (2017). The Business of Artificial Intelligence. The Big Idea. ***Harvard Business Review***. Online. URL: https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Chui, M., Manyika, J., & Miremadi, M. (2016). Where machines could replace humans – and where they can't (yet). ***McKinsey Quarterly***, July.

Coldewey, D. (2017). Laying a trap for self-driving cars. ***TechCrunch***, March 17.

Cross, S.E., Walker, E., (1994). DART: Applying knowledge-based planning and scheduling to crisis action planning (pp. 711–729). In M. Zweben & M.S. Fox, Eds., ***Intelligent Scheduling***. San Mateo, CA: Morgan Kaufmann.

Cross, T. (2017). Human obsolescence: How quickly will machines sweep man aside? ***The Economist***, November 21.

Economist (2016). Automation and anxiety: Will smarter machines cause mass unemployment? ***The Economist***, June 25.

Economist (2017). Why AI researchers like video games. ***The Economist***, May 13.

Engelbert, C., & Hagel, J. (2017). Fulfilling the promise of AI means rethinking the nature of work itself. ***Harvard Business Review***, December 18.

Evans, R.C., et al. (2017). Clinical decision support to efficiently identify patients eligible for advanced heart failure therapies. ***Journal of Cardiac Failure***, 23 (10), 719-726.

Feigenbaum, E.A. (1980). Expert systems: Looking back and looking ahead. In R. Wilhelm, Ed., ***Jahrestagung. Informatik-Fachberichte*** (Vol 33), Berlin: Springer.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., & Mueller, E.T. (2013). Watson: Beyond Jeopardy! ***Artificial Intelligence***, 199, 93–105.

Galeon, D., & Hauser, K. (2016). IBM's Watson AI recommends same treatment as doctors in 99% of cancer cases. ***Futurism***, October 26.

Grossmueller, A. (2017). How precision oncology will use data to advance cancer treatment. ***Health Data Management***, December 11.

Hof, R.D. (2013). Deep learning: With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart. ***Technology Review***, April.

Hsu, F. (2002). ***Behind Deep Blue: Building the computer that defeated the world chess champion.*** Princeton, NJ: Princeton University Press.

Isaacson, W. (2014). ***The Innovators: How a Group of Hackers, Geniuses, and Geeks Created The Digital Revolution***. New York: Simon & Schuster

Kaplan, J. (2015). Is it possible to create an ethical robot? ***Wall Street Journal***, July 25.

Lenartowicz, M. (2015). Mere impediments? A second thought on the role of social boundaries in self-organization of the global collective intelligence on Earth. ***Proceedings of the International Society for Information Studies Summit***, Vienna, June 3-7.

Lewis-Kraus, G. (2016). The great AI awakening. ***New York Times Magazine***, December 14.

Lighthill, J. (1973). Artificial Intelligence: A general survey. ***Symposium on Artificial Intelligence***. UK: Science Research Council.

Liu, C. (2017). ***Enterprise Transformation in the Automobile Ecosystem: How Brands and Technologies Interact with Market and Environment***. PhD Dissertation, School of Systems and Enterprises, Stevens Institute of Technology.

Lloyd's (2016). ***Foresight Review of Robotics and Autonomous*** Systems. London: Lloyd's.

Markoff, J. (2016) ***Machines of Loving Grace: The Quest for Common Ground Between Humans and Robots*.** New York: Ecco/Harper Collins.

Minsky, M. L. (1954). ***Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain Model Problem***, PhD Dissertation, Princeton, NJ: Princeton University.

Minsky, M., & Papert, S. (1969). ***Perceptrons: An introduction to computational geometry***, Cambridge MA: MIT Press.

Mittal, N., Lowes, P., Ronanki, R., Wen, J., & Sharma, S. (2017). ***Machine Intelligence: Technology Mimics Human Cognition to Create Value***. New York: Deloitte University Press.

Newell, A., Shaw, J.C., & Simon, H.A. (1959). Report on a general problem-solving program. ***Proceedings of the International Conference on Information Processing***, 256–264.

Ng, A. (2016). What artificial intelligence can and can't do right now. ***Harvard Business Review***, 2-4.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. ***Proceedings of Computer Vision and Pattern Recognition***, Boston, June 7-12.

Norman, D.A. (2014). The human side of automation. ***Proceedings of the Automated Vehicles Symposium***, San Francisco.

Paquette, D. (2017). Robots could replace nearly a third of the US workforce by 2030. ***The Washington Post***, November 30.

Partnership (2017). Partnership on AI strengthens its network of partners and announces first initiatives. https://www.partnershiponai.org.

Plungis, J. (2017). Self-driving cars: Driving into the future. ***Consumer Reports***, February 28.

Rosenblatt, F. (1957). ***The Perceptron: A Perceiving and Recognizing Automaton***. Buffalo, NY: Cornell Aeronautical Laboratory.

Rouse, S.H., & Rouse, W.B. (1980). Computer-based manuals for procedural information. ***IEEE Transactions on Systems, Man, and Cybernetics***, SMC-10(8), 506-510.

Rouse, S.H., Rouse, W.B., & Hammer, J.M. (1982). Design and evaluation of an onboard computer-based information system for aircraft. ***IEEE Transactions on Systems, Man, and Cybernetics***, SMC-12(4), 451-463.

Rouse, W.B. (1988). Adaptive aiding for human/computer control. ***Human Factors***, 30(4), 431-443.

Rouse, W.B. (2007). ***People and Organizations: Explorations of Human-Centered Design***. New York: Wiley.

Rouse, W.B., Geddes N.D., & Curry, R.E. (1988). An architecture for intelligent interfaces: Outline of an approach to supporting operators of complex systems. ***Human-Computer Interaction***, 3(2), 87-122.

Rouse, W.B., Geddes, N.D., & Hammer, J.M. (1990). Computer-aided fighter pilots. ***IEEE Spectrum***, 27(3), 38-41.

Rouse, W.B., & Morris, N.M. (1987). Conceptual design of a human error tolerant interface for complex engineering systems. ***Automatica***, 23(2), 231-235.

Schank, R. (1969). A conceptual dependency parser for natural language. ***Proceedings of the 1969 Conference on Computational linguistics***, Sång-Säby, Sweden, 1-3.

Schank, R., & Abelson, R. P. (1977). ***Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures***. Hillsdale, NJ: Erlbaum.

Searle, J. (1980). Minds, brains, and programs. ***Behavioral and Brain Sciences***, 3, 417-424.

Seuss, C.D. (2011). In-depth understanding: Teaching search engines to interpret meaning. ***Proceedings of the IEEE***, 99 (4), 531-535.

Shortliffe, E.H., & Buchanan, B.G. (1975). A model of inexact reasoning in medicine. ***Mathematical Biosciences***, 23, (3–4), 351–379.

Spohrer J. (2016). Innovation for jobs with cognitive assistants: A service science perspective. In D. Nordfors & M. Senges, Eds., ***Disrupting Unemployment*** (pp. 157-174). St. Louis, MO: Ewing Marion Kauffman Foundation.

Spohrer J. & Banavar G. (2015). Cognition as a service: an industry perspective. ***AI Magazine***. 36 (4),71-86.

Spohrer J., Giuiusa A., Demirkan H., & Ing, D. (2013). Service science: Reframing progress with universities. ***Systems Research and Behavioral Science***. 30 (5), 561-569.

Spohrer J., Siddike M.A., & Kohda Y. (2017). Rebuilding Evolution: A Service Science Perspective. ***Proceedings of the 50th Hawaii International Conference on System Sciences***, Jan 4.

Stanford (2016). ***One Hundred Year Study on Artificial Intelligence***. Palo Alto: CA: Stanford University.

Sherman, E. (2015). It's impossible to find out if self-driving cars are safe: Report. ***Fortune***, April 12

Turing, A.M. (1950). Computing machinery and intelligence. ***Mind***, 59, 433-460.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. ***Communications of the ACM***, 9, 36–45.

Widerquist K., & Lewis M.A. (2005). ***The Ethics and Economics of the Basic Income Guarantee***. New York: Routledge.

Yu, Z., Serban, N., & Rouse, W.B. (2013). The demographics of change: Enterprise characteristics and behaviors that influence transformation. ***Journal of Enterprise Transformation***, 3 (4), 285-306.