# Deloitte Review

**Issue 20  |  2017**

## Cognitive collaboration

Why humans and computers think better together

**By James Guszcza, Harvey Lewis, and Peter Evans-Greenwood**
**Illustration by Josie Portillo**

# Deloitte.

# Cognitive
# collaboration

## Why humans and computers think better together

**By James Guszcza, Harvey Lewis, and Peter Evans-Greenwood**
**Illustration by Josie Portillo**

## A SCIENCE OF THE ARTIFICIAL

**A**LTHOUGH artificial intelligence (AI) has experienced a number of "springs" and "winters" in its roughly 60-year history, it is safe to expect the current AI spring to be both lasting and fertile. Applications that seemed like science fiction a decade ago are becoming science fact at a pace that has surprised even many experts.

The stage for the current AI revival was set in 2011 with the televised triumph of the IBM Watson computer system over former *Jeopardy!* game show champions Ken Jennings and Brad Rutter. This watershed moment has been followed rapid-fire by a sequence of striking breakthroughs, many involving the machine learning technique known as *deep learning*. Computer algorithms now beat humans at games of skill, master video games with no prior instruction, 3D-print original paintings in the style of Rembrandt, grade student papers, cook meals, vacuum floors, and drive cars.[1]

All of this has created considerable uncertainty about our future relationship with machines, the prospect of technological unemployment, and even the very fate of humanity. Regarding the latter topic, Elon Musk has described AI as "our biggest existential threat." Stephen Hawking warned that "The development of full artificial intelligence could spell the end of the human race." In his widely discussed book *Superintelligence*, the philosopher Nick Bostrom discusses the possibility of a kind of technological "singularity" at which point the general cognitive abilities of computers exceed those of humans.[2]

---

> "Any sufficiently advanced technology is indistinguishable from magic."
>
> —*Arthur C. Clarke's Third Law[3]*

---

Discussions of these issues are often muddied by the tacit assumption that, because computers outperform humans at various circumscribed tasks, they will soon be able to "outthink" us more generally. Continual rapid growth in computing power and AI breakthroughs notwithstanding, this premise is far from obvious.

Furthermore, the assumption distracts attention from a less speculative topic in need of deeper attention than it typically receives: the ways in which machine intelligence and human intelligence *complement* one another. AI has made a dramatic comeback in the past five years. We believe that another, equally venerable, concept is long overdue for a comeback of its own: *intelligence augmentation*. With intelligence augmentation, the ultimate goal is not building machines that think like humans, but designing machines that help humans think better.

## THE HISTORY OF THE FUTURE OF AI

AI as a scientific discipline is commonly agreed to date back to a conference held at Dartmouth University in the summer of 1955. The conference was convened by John McCarthy, who coined the term "artificial intelligence," defining it as the science of creating machines "with the ability to achieve goals in the world."[4] The Dartmouth Conference was attended by a who's who of AI pioneers, including Claude Shannon, Alan Newell, Herbert Simon, and Marvin Minsky.

Interestingly, Minsky later served as an advisor to Stanley Kubrick's adaptation of the Arthur C. Clarke novel *2001: A Space Odyssey*. Perhaps that movie's most memorable character was HAL 9000: a computer that spoke fluent English, used commonsense reasoning, experienced jealousy, and tried to escape termination by doing away with the ship's crew. In short, HAL was a computer that implemented a very general form of human intelligence.

The attendees of the Dartmouth Conference believed that, by 2001, computers would implement an artificial form of human intelligence. Their original proposal stated:

> The study is to proceed on the basis of the conjecture that *every aspect of learning or any other feature of intelligence* can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves [emphasis added].[5]

As is clear from widespread media speculation about a "technological singularity," this original vision of AI is still very much with us today. For example, a *Financial Times* profile of DeepMind CEO Demis Hassabis stated that:

> At DeepMind, engineers have created programs based on neural networks, modelled on the human brain. These systems make mistakes, but learn and improve over time. They can be set to play other games and solve other tasks, so the intelligence is general, not specific. This AI "thinks" like humans do.[6]

Such statements mislead in at least two ways. First, in contrast with the artificial *general* intelligence envisioned by the Dartmouth Conference participants, the examples of AI on offer—either currently or in the foreseeable future—are all examples of *narrow* artificial intelligence. In human psychology, general intelligence is quantified by the so-called "*g* factor" (aka IQ), which measures the degree to which one type of cognitive ability (say, learning a foreign language) is associated with other cognitive abilities (say, mathematical ability). This is not characteristic of today's AI applications: An algorithm designed to drive a car would be useless at detecting a face in a crowd or guiding a domestic robot assistant.

Second, and more fundamentally, current manifestations of AI have little in common with the AI envisioned at the Dartmouth Conference. While they do manifest a narrow type of "intelligence" in that they can solve problems and achieve goals, this does not involve implementing human psychology or brain science. Rather, it involves machine learning: the process of fitting highly complex and powerful—but typically uninterpretable—statistical models to massive amounts of data.

For example, AI algorithms can now distinguish between breeds of dogs more accurately than humans can.[7] But this does not involve algorithmically representing such concepts as "pinscher" or "terrier." Rather, deep learning neural network models, containing thousands of uninterpretable parameters, are trained on large numbers of digitized photographs that have already been labeled by humans.[8] In a similar way that a standard regression model can predict a person's income based on various educational, employment, and psychological details, a deep learning model uses a

photograph's pixels as input variables to predict such outcomes as "pinscher" or "terrier"—without needing to understand the underlying concepts.

The ambiguity between general and narrow AI—and the evocative nature of terms like "neural," "deep," and "learning"—invites confusion. While neural networks are loosely inspired by a simple model of the human brain, they are better viewed as generalizations of statistical regression models. Similarly, "deep" refers not to psychological depth, but to the addition of structure ("hidden layers" in the vernacular) that enables a model to capture complex, nonlinear patterns. And "learning" refers to numerically estimating large numbers of model parameters, akin to the "β" parameters in regression models. When commentators write that such models "learn from experience and get better," they mean that more data result in more accurate parameter estimates. When they claim that such models "think like humans do," they are mistaken.[9]

In short, the AI that is reshaping our societies and economies is far removed from the vision articulated in 1955 at Dartmouth, or implicit in such cinematic avatars as HAL and Lieutenant Data. Modern AI is founded on computer-age statistical inference—not on an approximation or simulation of what we believe human intelligence to be.[10] The increasing ubiquity of such applications will track the inexorable growth of digital technology. But they will not bring us closer to the original vision articulated at Dartmouth. Appreciating this is crucial for understanding both the promise and the perils of real-world AI.

## LICKLIDER'S AUGMENTATION

FIVE years after the Dartmouth Conference, the psychologist and computer scientist J. C. R. Licklider articulated a significantly different vision of the relationship between human and computer intelligence. While the general AI envisioned at Dartmouth remains the stuff of science fiction, Licklider's vision is today's science fact, and provides the most productive way to think about AI going forward.[11]

Rather than speculate about the ability of computers to *implement* human-style intelligence, Licklider believed computers would *complement* human intelligence. He argued that humans and computers would develop a symbiotic relationship, the strengths of one counterbalancing the limitations of the other:

> Men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations. Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions in technical and scientific thinking. . . . The symbiotic partnership will perform intellectual operations much more effectively than man alone can perform them.[12]

We humans must settle for solutions that "satisfice" rather than optimize because our memory and reasoning ability are limited.

This kind of human-computer symbiosis already permeates daily life. Familiar examples include:

- Planning a trip using GPS apps like Waze

- Using Google Translate to help translate a document

- Navigating massive numbers of book or movie choices using menus of personalized recommendations

- Using Internet search to facilitate the process of researching and writing an article

In each case, the human specifies the goal and criteria (such as "Take me downtown but avoid highways" or "Find me a highly rated and moderately priced sushi bar within walking distance"). An AI algorithm sifts through otherwise unmanageable amounts of data to identify relevant predictions or recommendations. The human then evaluates the computer-generated options to arrive at a decision. In no case is human intelligence mimicked; in each case, it is augmented.

Developments in both psychology and AI subsequent to the Dartmouth Conference suggest that Licklider's vision of human-computer symbiosis is a more productive guide to the future than speculations about "superintelligent" AI. It turns out that the human mind is less computer-like than originally realized, and AI is less human-like than originally hoped.

### LINDA, *C'EST MOI*

AI algorithms enjoy many obvious advantages over the human mind. Indeed, the AI pioneer Herbert Simon is also renowned for his work on bounded rationality: We humans must settle for solutions that "satisfice" rather than optimize because our memory and reasoning ability are limited. In contrast, computers do not get tired; they make consistent decisions before and after lunchtime; they can process decades' worth of legal cases, medical journal articles, or accounting regulations with minimal effort; and they can evaluate five hundred predictive factors far more accurately than unaided human judgment can evaluate five.

This last point hints at a transformation in our understanding of human psychology, introduced by Daniel Kahneman and Amos Tversky well after the Dartmouth Conference and Licklider's essay. Consider the process of making predictions: Will this job candidate succeed if we hire her? Will this insurance risk be profitable? Will this prisoner recidivate if paroled?

Intuitively, it might seem that our thinking approximates statistical models when making such judgments. And indeed, with training and deliberate effort, it can—to a degree. This is what Kahneman calls "System 2" thinking, or "thinking slow."[13]

But it turns out that most of the time we use a very different type of mental process when making judgments and decisions. Rather than laboriously gathering and evaluating the relevant evidence, we typically lean on a variety of mental rules of thumb (heuristics) that yield narratively plausible, but often logically dubious, judgments. Kahneman calls this "System 1," or "thinking fast," which is famously illustrated by the "Linda" experiment. In an experiment with students at top universities, Kahneman and Tversky described a fictional character named Linda: She is very intelligent, majored in philosophy at college, and participated in the feminist movement and anti-nuclear demonstrations. Based on these details about Linda's college days, which is the more plausible scenario involving Linda today?

1. Linda is a bank teller.

2. Linda is a bank teller who is active in the feminist movement.

*Algorithms are reliable only to the extent that the data used to train them are sufficiently complete and representative of the environment in which they are to be deployed.*

Kahneman and Tversky reported that 87 percent of the students questioned thought the second scenario more likely, even though a moment's thought reveals that this could not possibly be the case: Feminist bank tellers are a subset of all bank tellers. But adding the detail that Linda is still active in the feminist movement lends narrative coherence, and therefore intuitive plausibility, to the (less likely) second scenario.

Kahneman calls the mind "a machine for jumping to conclusions": We confuse the easily imaginable with the highly probable,[14] let emotions cloud judgments, find patterns in random noise, tell spuriously causal stories about cases of regression to the mean, and overgeneralize from personal experience. Many of the mental heuristics we use to make judgments and decisions turn out to be systematically biased. Dan Ariely's phrase "predictably irrational" describes the mind's systematic tendency to rely on biased mental heuristics.

Such findings help explain a phenomenon first documented by Kahneman's predecessor Paul Meehl in the 1950s and subsequently validated by hundreds of academic studies and industrial applications of the sort dramatized in Michael Lewis's *Moneyball*: The predic-

tions of simple algorithms routinely beat those of well-informed human experts in a wide variety of domains. This points to the need for human-computer collaboration in a way that even Licklider himself probably didn't imagine. It turns out that minds need algorithms to debias our judgments and decisions as surely as our eyes need artificial lenses to see adequately.

## I'M SORRY, DAVE. I'M AFRAID I CAN'T DO THAT.

WHILE it is easy to anthropomorphize self-driving cars, voice-activated personal assistants, and computers capable of beating humans at games of skill, we have seen that such technologies are "intelligent" in essentially the same minimal way that credit scoring or fraud detection algorithms are. This means that they are subject to a fundamental limitation of data-driven statistical inference: Algorithms are reliable only to the extent that the data used to train them are sufficiently complete and representative of the environment in which they are to be deployed. When this condition is not met, all bets are off.

To illustrate, consider a few examples involving familiar forms of AI:

- During the *Jeopardy!* match with Watson, Jennings, and Rutter, Alex Trebek posed this question under the category "US cities": "Its largest airport is named for a World War II hero; its second largest, for a World War II battle." Watson answered "Toronto."[15]

- One of us used a common machine translating service to translate the recent news headline "Hillary slams the door on Bernie" from English into Bengali, then back again. The result was "Barney slam the door on Clinton."[16]

- In 2014, a group of computer scientists demonstrated that it is possible to "fool" state-of-the-art deep learning algorithms into classifying unrecognizable or white noise images as common objects (such as "peacock" or "baseball") with very high confidence.[17]

- On May 7, 2016, an unattended car in "autopilot" mode drove underneath a tractor-trailer that it did not detect, shearing off the roof of the car and killing the driver.[18]

None of these stories suggest that the algorithms aren't highly useful. Quite the contrary. IBM's Watson did, after all, win *Jeopardy!*; machine translation and image recognition algorithms are enabling new products and services; and even the self-driving car fatality must be weighed against the much larger number of lives likely to be saved by autonomous vehicles.[19]

Rather, these examples illustrate another point that Licklider would have appreciated: Certain strengths of human intelligence can counterbalance the fundamental limitations of brute-force machine learning.

Returning to the above examples:

- Watson, an information retrieval system, would have responded correctly if it had access to, for example, a Wikipedia page listing the above facts about Chicago's two major airports. But it is unable to use commonsense reasoning, as answering "Toronto" to a question about "US cities" illustrates.[20]

- Today's machine translation algorithms cannot reliably extrapolate beyond existing data (including millions of phrase pairs from documents) to translate novel combinations of words, new forms of slang, and so on. In contrast, a basic phenomenon emphasized by Noam Chomsky in linguistics is the ability of young children to acquire language—with its infinite number of possible sentences—based on surprisingly little data.[21]

- A deep learning algorithm must be trained with many thousands of photographs to recognize (for example) kittens—and even then, it has formed no conceptual understanding. In contrast, even small children are actually very good at forming hypotheses and learning from a small number of examples.

- Autonomous vehicles must make do with algorithms that cannot reliably extrapolate beyond the scenarios encoded in their databases. This contrasts with the ability of human drivers to use judgment and common sense in unfamiliar, ambiguous, or dynamically changing situations.

In short, when routine tasks can be encoded in big data, it is a safe bet that algorithms can be built to perform them better than humans can. But such algorithms will lack the conceptual understanding and commonsense reasoning needed to evaluate novel situations. They can make inferences from structured hypotheses but lack the intuition to prioritize which hypothesis to test in the first place. The cognitive scientist Alison Gopnik summarizes the situation this way:

> One of the fascinating things about the search for AI is that it's been so hard to predict which parts would be easy or hard. At first, we thought that the quintessential preoccupations of the officially smart few, like playing chess or proving theorems—the corridas of nerd machismo—would prove to be hardest for computers. In fact, they turn out to be easy. Things every dummy can do, like recognizing objects or picking them up, are much harder. And it turns out to be much easier to simulate the reasoning of a highly trained adult expert than to mimic the ordinary learning of every baby.[22]

Just as humans need algorithms to avoid "System 1" decision traps, the inherent limitations

of big data imply the need for human judgment to keep mission-critical algorithms in check. Neither of these points were as obvious in Licklider's time as they are today. Together, they imply that the case for human-computer symbiosis is stronger than ever.

## GAME OVER?

CHESS provides an excellent example of human-computer collaboration—and a cautionary tale about over-interpreting dramatic examples of computers outperforming humans. In 1997, IBM's Deep Blue beat the chess grandmaster Garry Kasparov. A major news magazine made the event a cover story titled "The brain's last stand." Many observers proclaimed the game to be over.[23]

Eight years later, it became clear that the story is considerably more interesting than "machine vanquishes man." A competition called "freestyle chess" was held, allowing any combination of human and computer chess players to compete. The competition resulted in an upset victory that Kasparov later reflected upon:

> The surprise came at the conclusion of the event. The winner was revealed to be not a grandmaster with a state-of-the-art PC but a pair of amateur American chess players using three computers at the same time. Their skill at manipulating and "coaching" their computers to look very deeply into positions effectively counteracted the superior chess understanding of their grandmaster oppo-

nents and the greater computational power of other participants. Weak human + machine + better process was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process . . . Human strategic guidance combined with the tactical acuity of a computer was overwhelming.[24]

"Freestyle *x*" is a useful way of thinking about human-computer collaboration in a variety of domains. To be sure, some jobs traditionally performed by humans have been and will continue to be displaced by AI algorithms. An early example is the job of bank loan officer, which was largely eliminated after the introduction of credit scoring algorithms. In the future, it is possible that jobs ranging from long-haul truck driver to radiologist could be largely automated.[25] But there are many other cases where variations on "freestyle *x*" are a more plausible scenario than jobs simply being replaced by AI.

For example, in their report *The future of employment: How susceptible are jobs to computerization?*, the Oxford University business school professors Carl Benedikt Frey and Michael Osbourne list "insurance underwriters" as one of the top five jobs most susceptible to computerization, a few notches away from "tax preparers." Indeed, it is true that sophisticated actuarial models serve as a type of AI that eliminates the need for manual underwriting of standard personal auto or homeowners insurance contracts.

Consider, though, the more complex challenge of underwriting businesses for commercial liability or injured worker risks. There are fewer businesses to insure than there are cars and homes, and there are typically fewer predictive data elements common to the wide variety of businesses needing insurance (some are hipster artisanal pickle boutiques; others are construction companies). In statistical terms, this means that there are fewer rows and columns of data available to train predictive algorithms. The models can do no more than mechanically tie together the limited number of risk factors fed into them. They cannot evaluate the accuracy or the completeness of this information, nor can they weigh it together with various case-specific nuances that might be obvious to a human expert, nor can they underwrite new types of businesses and risks not represented in the historical data. However, such algorithms can often automate the underwriting of small, straightforward risks, giving the underwriter more time to focus on the more complex cases requiring commonsense reasoning and professional judgment.

Similar comments about job loss to AI can be made about fraud investigators (particularly in domains where fraudsters rapidly evolve their tactics, rendering historical data less relevant), hiring managers, university admissions officers, public sector case workers, judges making parole decisions, and physicians making medical diagnoses. In each domain, cases fall on a spectrum. When the cases are frequent, unambiguous, and similar across time and context—and if the downside costs of a false prediction are acceptable—algorithms can presumably automate the decision. On the other hand, when the cases are more complex, novel, exceptional, or ambiguous—in other words, not fully represented by historical cases in the available data—human-computer collaboration is a more plausible and desirable goal than complete automation.

The current debates surrounding self-driving cars illustrate this spectrum. If driving environments could be sufficiently controlled—for example, dedicated lanes accessible only to autonomous vehicles, all equipped with interoperable sensors—level 5 autonomous vehicles would be possible in the near term.[26] However, given the number of "black swan"-type scenarios possible (a never-before-seen combination of weather, construction work, a mattress falling off a truck, and someone crossing the road—analogous to the example of translating "Hillary slams the door on Bernie" into Bengali), it is unclear when it will be possible to dispense entirely with human oversight and commonsense reasoning.

## BRIDGING THE EMPATHY GAP

FOR the reasons given above, and also because of its inherent "human element," medicine is a particularly fertile domain for "freestyle *x*" collaboration. Paul

Just as it is overly simplistic to think that computers are getting smarter than humans, it is probably equally simplistic to think that only humans are good at empathy.

Meehl realized 60 years ago that even simple predictive algorithms can outperform unaided clinical judgment.[27] Today, we have large databases of lifestyle data, genomics data, self-tracking devices, mobile phones capable of taking medical readings, and Watson-style information retrieval systems capable of accessing libraries of continually updated medical journals. Perhaps the treatment of simple injuries, particularly in remote or underserved places, will soon be largely automated, and certain advanced specialties such as radiology or pathology might be largely automated by deep learning technologies.

More generally, the proliferation of AI applications in medicine will likely alter the mix of skills that characterize the most successful physicians and health care workers. Just as the skills that enabled Garry Kasparov to become a chess master did not guarantee dominance at freestyle chess, it is likely that the best doctors of the future will combine the ability to use AI tools to make better diagnoses with the ability to empathetically advise and comfort patients. Machine learning algorithms will enable

physicians to devote fewer mental cycles to the "spadework" tasks computers are good at (memorizing the *Physicians' Desk Reference*, continually scanning new journal articles) and more to such characteristically human tasks as handling ambiguity, strategizing treatment and wellness regimens, and providing empathetic counsel.

However, just as it is overly simplistic to think that computers are getting smarter than humans, it is probably equally simplistic to think that only humans are good at empathy. There is evidence that AI algorithms can play a role in promoting empathy. For example, the Affectiva software is capable of inferring people's emotional states from webcam videos of their facial expressions. Such software can be used to help optimize video content: An editor might eliminate a section from a movie trailer associated with bored audience facial expressions. Interestingly, the creators of Affectiva were originally motivated by the desire to help autistic people better infer emotional states from facial expressions. Such software could be relevant not only in medicine and marketing, but in the

broader business world: Research has revealed that teams containing more women, as well as team members with high degrees of social perception (the trait that Affectiva was designed to support), exhibit higher group intelligence.[28]

There is also evidence that big data and AI can help with both verbal and nonverbal communications between patients and health care workers (and, by extension, between teachers and students, managers and team members, salespeople and customers, and so on). For example, Catherine Kreatsoulas has led the development of algorithms that estimate the likelihood of coronary heart disease based on patients' own descriptions of their symptoms. Kreatsoulas has found evidence that men and women tend to describe symptoms differently, potentially leading to differential treatment. It's possible that well-designed AI algorithms can help avoid such biases.[29]

Regarding nonverbal communication, Sandy Pentland and his collaborators at MIT Media Lab have developed a wearable device, known as the "sociometer," that can measure patterns of nonverbal communication. Such devices could be used to quantify otherwise intangible aspects of communication style in order to coach health care workers on how to cultivate a better bedside manner. This work could even bear on medical malpractice claims: There is evidence that physicians who are perceived as more "likable" are sued for malpractice less often, independently of other risk factors.[30]

## ALGORITHMS CAN BE BIASED, TOO

ANOTHER type of mental operation that cannot (and must not) be outsourced to algorithms is reasoning about fairness, societal acceptability, and morality. The naive view that algorithms are "fair" and "objective" simply because they use hard data is giving way to recognition of the need for oversight. In a broad sense, this idea is not new. For example, there has long been legal doctrine around the socially undesirable disparate impact that hiring and credit scoring algorithms can potentially have on various classes of individuals.[31] More recent examples of algorithmic bias include online advertising systems that have been found to target career-coaching service ads for high-paying jobs more frequently to men than women, and ads suggestive of arrests more often to people with names commonly used by black people.[32]

Such examples point to yet another sense in which AI algorithms must be complemented by human judgment: If the data used to train an algorithms reflect unwanted pre-existing biases, the resulting algorithm will likely reflect, and potentially amplify, these biases.

The example of judges—and sometimes algorithms—making parole decisions illustrates the subtleties involved. In light of the work of Meehl, Kahneman, and their collaborators, there is good reason to believe that judges should consult algorithms when making parole decisions. A well-known study of judges

making parole decisions indicates that, early in the morning, judges granted parole roughly 60 percent of the time. This probability would shrink steadily to near zero by mid-morning break, then would shoot up to roughly 60 percent after break, shrink steadily back to zero by lunch time, jump back to 60 percent after lunch, and so on throughout the day. It seems that blood sugar level significantly affects these hugely important decisions.[33]

Such phenomena suggest that *not* considering the use of algorithms to improve parole decisions would be morally questionable. Yet a recent study vividly reminds us that building such algorithms is no straightforward task. The journalist Julia Angwin, collaborating with a team of data scientists, reported that a widely used black-box recidivism risk scoring model mistakenly flags black defendants at roughly twice the rate as it mistakenly flags white people. A few months after Angwin's story appeared, the Wisconsin Supreme Court ruled that while judges could use risk scores, the scores cannot be a "determinative" factor in whether or not a defendant is jailed. In essence,

> Just as the best cars are ergonomically designed to maximize the driver's comfort and control, so decision-support algorithms must be designed to go with the grain of human psychology, rather than simply bypass human psychology altogether.

the ruling calls for a judicial analog of freestyle chess: Algorithms must not automate judicial decisions; rather, judges should be able to use them as *tools* to make better decisions.[34]

An implication of this ruling is that such algorithms should be designed, built, and evaluated using a broader set of methods and concepts than the ones typically associated with data science. From a narrowly technical perspective, an "optimal" model is often judged to be the one with the highest out-of-sample accuracy. But from a broader perspective, a usable model must balance accuracy with such criteria as omitting societally vexed predictors, avoiding unwanted biases,[35] and providing enough transparency to enable the end user to evaluate the appropriateness of the model indication in a particular case.[36]

The winners of the freestyle chess tournament have been described as "driving" their computer algorithms in a similar way that a person drives a car. Just as the best cars are ergonomically designed to maximize the driver's comfort and control, so decision-support algorithms

must be designed to go with the grain of human psychology, rather than simply bypass human psychology altogether. Paraphrasing Kasparov, humans plus computers *plus a better process* for working with algorithms will yield better results than either the most talented humans or the most advanced algorithms working in isolation. The need to design those better processes for human-computer collaboration deserves more attention than it typically gets in discussions of data science or artificial intelligence.

## DESIGNING THE FUTURE

THERE is a coda to our story. One of Licklider's disciples was Douglas Engelbart of the Stanford Research Institute (SRI). Two years after Licklider wrote his prescient essay, Engelbart wrote an essay of his own called "Augmenting human intellect: A conceptual framework," which focused on "increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems."[37] Like Licklider's, this was

a vision that involved keeping humans in the loop, not automating away human involvement.

Engelbart led the Augmentation Research Center at SRI, which in the mid-1960s invented many of the elements of the modern personal computer. For example, Engelbart conceived of the mouse while pondering how to move a cursor on a computer screen. The mouse—along with such key elements of personal computing as videoconferencing, word processing, hypertext, and windows—was unveiled at "the mother of all demos" in San Francisco in 1968, which is today remembered as a seminal event in the history of computing.[38]

About a decade after Engelbart's demo, Steve Jobs purchased the mouse patent from SRI for a reported $40,000. Given this lineage, it is perhaps no accident that Jobs memorably articulated a vision of human-computer collaboration very close in spirit to Licklider's:

> I think one of the things that really separates us from the high primates is that we're tool builders. I read a study that measured the efficiency of locomotion for various species on the planet. . . . Humans came in with a rather unimpressive showing, about a third of the way down the list, but somebody at *Scientific American* had the insight to test the efficiency of locomotion for a man on a bicycle. . . . A human on a bicycle blew the condor away, completely off the top of the charts. And that's what a computer is to me . . . it's the most remarkable tool that we've ever come up with; it's the equivalent of a bicycle for our minds.[39]

Consistent with this quote, Jobs is remembered for injecting human-centric design thinking into personal computer technology. We believe that fully achieving Licklider's vision of human-computer symbiosis will require a similar injection of psychology and design thinking into the domains of data science and AI.

## "THIS HAS TO BE A HUMAN SYSTEM WE LIVE IN": SANDY PENTLAND ON ARTIFICIAL INTELLIGENCE

*Alex "Sandy" Pentland is the Toshiba Chair in Media Arts and Sciences at MIT (a chair previously held by Marvin Minsky), one of the founders of the MIT Media Lab, and one of the most cited authors in computer science. He is a pioneer in the burgeoning interdisciplinary field of computational social science, and is the author of the recent book* Social Physics. *Last summer, Pentland kindly agreed to be interviewed about big data, computational social science, and artificial intelligence. A portion of this interview appears below.*

**Jim Guszcza:** Sandy, one of your predecessors at MIT was Marvin Minsky, who was one of the founders of artificial intelligence back in the 1950s. What do you think of current developments in AI?

**Sandy Pentland:** The whole AI thing is rather overhyped. It is going to be a blockbuster economically—that is clear. It is all based on correlation, where you can tune systems with examples of all the types of variations you are interested in, so they can interpret things. However, there are basically zero examples of AI extrapolating new situations. To do that, you have to understand the causal structure of what is going on.

**JG:** Which is what Minsky wanted originally, right?

**SP:** That is right. Marvin was advocating what's called "commonsense reasoning." Machines have shown essentially no examples of doing that. Therefore, they are complements to people. People are actually not so bad at that. However, they are somewhat lousy at tuning things and keeping exact accounts of stuff. Machines are good at that.

That gives the idea that there could be a human-machine partnership, and there are examples of that. A middle-class chess player with a middle-class machine beats the best chess machine and



Photo by Ben Gebo

the best chess human. I think we see a lot of examples coming up where the human does the strategy, the machine does the tactics, and when you put them together, you get a world-beater.

**JG:** This can make the human more human. For example, a doctor could use information retrieval like IBM Watson to call up documents based on the symptoms. You could use deep learning to do medical imaging, leaving the health care worker more time to empathize and the patient more time to strategize.

Originally, Minsky, Simon, and Newell wanted strong AI (general AI). Right now we've got narrow AI. Do you see us going back to that research program that Newell and Simon wanted?

**SP:** No, I think that was a mistake in many ways. Perhaps it was a tactical win. However, this human-machine system thing is a much better idea for a lot of reasons. One is the complementary side of it, but the other thing is that this has to be a human system we live in. Otherwise, why are we doing it? One of the big problems with big data and AI is how to keep human values as central. If you think of it as a partnership, then there is a natural way to do that. If you think of AI as replacing people, you end up in all sorts of nightmare scenarios.

**JG:** This is keeping humans in the loop.

**SP:** But as partners, as symbiotes—not as just extras. **DR**

---

*James Guszcza* is the US chief data scientist for Deloitte Consulting LLP.

*Harvey Lewis* is a director in Deloitte MCS Limited and lead for cognitive computing in the Technology Consulting practice in the United Kingdom.

*Peter Evans-Greenwood* is a fellow at the Deloitte Centre for the Edge, Australia, supported by Deloitte Consulting Pty. Ltd.

**Endnotes**

1.  Watson's triumph on *Jeopardy!* was reported by John Markoff, "Computer wins on 'Jeopardy!': Trivial, it's not," *New York Times*, February 11, 2011. The algorithmically generated Rembrandt was reported by Chris Baraniuk, "Computer paints 'new Rembrandt' after old works analysis," *BBC News*, April 6, 2016. For robot chefs, see Matt Burgess, "Robot chef that can cook any of 2000 meals at tap of button to go on sale in 2017," *Factor Tech*, April 14, 2015. Regarding self-driving cars, see Cecilia Kang, "No driver? Bring it on. How Pittsburgh became Uber's testing ground," *New York Times*, September 10, 2016.

2.  Regarding technological unemployment, a recent World Economic Forum report predicted that the next four years will see more than 5 million jobs lost to AI-fueled automation and robotics. See World Economic Forum, *The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution*, January 2016, http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf. Regarding Musk and Hawking on AI as an existential threat, see "Elon Musk: Artificial intelligence is our biggest existential threat," *The Guardian*, October 27, 2014, and "Stephen Hawking warns artificial intelligence could end mankind," *BBC News*, December 2, 2014. In his book *Superintelligence* (Oxford University Press, 2014), Nick Bostrom entertains a variety of speculative scenarios about the emergence of "superintelligence," which he defines as any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.

3.  The Arthur C. Clarke Foundation, "Sir Arthur's quotations," http://www.clarke-foundation.org/about-sir-arthur/sir-arthurs-quotations/, accessed October 24, 2016.

4.  In more detail: McCarthy defined artificial intelligence as "the science and engineering of making intelligent machines, especially intelligent computer programs" and defined intelligence as "the computational part of the ability to achieve goals in the world." He noted that "Varying kinds and degrees of intelligence occur in people, many animals and some machines." See John McCarthy, "What is artificial intelligence?," http://www-formal.stanford.edu/jmc/whatisai/whatisai.html, accessed October 24, 2016.

5.  The original proposal can be found in John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A proposal for the Dartmouth Summer Research Project on Artificial Intelligence," *AI Magazine* 27, no. 4 (2006), http://www.aaai.org/ojs/index.php/aimagazine/article/view/1904/1802. Regarding the time frame, the proposal went on to state, "We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it for a summer." (!) In hindsight, this optimism might seem surprising, but it is worth remembering that the authors were writing in the heyday of both behaviorist psychology, led by B. F. Skinner, and the logical positivist school of philosophy. Our understanding of both human psychology and the challenges of encoding knowledge in logically perfect languages has evolved considerably since the 1950s.

6.  "Demis Hassabis, master of the new machine age," *Financial Times*, March 11, 2016, https://www.ft.com/content/630bcb34-e6b9-11e5-a09b-1f8b0d268c39#axzz42jQkBo00. This is not an isolated statement. Two days earlier, the *New York Times* carried an opinion piece by an academic who stated that "Google's AlphaGo is demonstrating for the first time that machines can truly learn and think in a human way." Howard Yu, "AlphaGo's success shows the human advantage is eroding fast," *New York Times*, March 9, 2016, http://www.nytimes.com/roomfordebate/2016/03/09/does-alphago-mean-artificial-intelligence-is-the-real-deal/alphagos-success-shows-the-human-advantage-is-eroding-fast.

7.  Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," February 6, 2015, https://arxiv.org/pdf/1502.01852v1.pdf.

8.  It is common for such algorithms to fail in certain ambiguous cases that can be correctly labeled by human experts. These new data points can then be used to retrain the models, resulting in improved accuracy. This virtuous cycle of human labeling and machine learning is called "human-in-the-loop computing." See, for example, Lukas Biewald, "Why human-in-the-loop computing is the future of machine learning," *Computerworld*, November 13, 2015, http://www.computerworld.com/article/3004013/robotics/why-human-in-the-loop-computing-is-the-future-of-machine-learning.html.

9.  In a recent IEEE interview, the Berkeley statistician and machine learning authority Michael Jordan comments that "Each neuron [in a deep learning neural net model] is really a cartoon. It's a linear-weighted sum that's passed through a nonlinearity. Anyone in electrical engineering would recognize those kinds of nonlinear systems. Calling that a neuron is clearly, at best, a shorthand. It's really a cartoon. There is a procedure called logistic regression in statistics that dates from the 1950s, which had nothing to do with neurons but which is exactly the same little piece of architecture." Lee Gomes, "Machine-learning maestro Michael Jordan on the delusions of big data and other huge engineering efforts," *IEEE Spectrum*, October 20, 2014, http://spectrum.ieee.org/robotics/

artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts. For technical details of how deep learning models are founded on generalized linear models (a core statistical technique that generalizes both classical and logistic regression), see Shakir Mohamed, "A statistical view of deep learning (I): Recursive GLMs," January 19, 2015, http://blog.shakirm.com/2015/01/a-statistical-view-of-deep-learning-i-recursive-glms/.

10. *Computer Age Statistical Inference* is the title of a new monograph by the eminent Stanford statisticians Brad Efron and Trevor Hastie. This book presents a unified survey of classical (e.g., maximum likelihood, Bayes theorem), "mid-century modern" (e.g., empirical Bayes, shrinkage, ridge regression), and modern (e.g., lasso regression, tree-based modeling, deep learning neural networks) statistical methods. See Bradley Efron and Trevor Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* (Cambridge University Press, 2016).

11. While not a household name, Licklider occupies an important place in the history of computing, and has been called "computing's Johnny Appleseed." During his tenure as research director at the US Department of Defense's Advanced Research Projects Agency (ARPA), Licklider wrote a memo, fancifully entitled "Memorandum for members and affiliates of the intergalactic computer network," outlining a vision that led to the creation of ARPANet, the forerunner of the Internet. See the Wikipedia page on Licklider, https://en.wikipedia.org/wiki/J._C._R._Licklider, for references.

12. J. C. R. Licklider, "Man-computer symbiosis," *IRE Transactions on Human Factors in Electronics*, March 1960, http://worrydream.com/refs/Licklider%20-%20Man-Computer%20Symbiosis.pdf. In this essay, Licklider analogized the relationship of humans and computers with that of the fig wasp and the fig tree.

13. Kahneman outlines the so-called dual process theory of psychology (System 1 versus System 2 mental operations) in his book *Thinking, Fast and Slow* (Farrar, Straus, and Giroux, 2013). This book contains an account of the "Linda" experiment discussed in the following paragraph.

14. It is helpful to keep this point, known as the "availability heuristic," in mind when considering the likelihood of apocalyptic scenarios of AI technologies run amok.

15. Michelle Castillo, "Why did Watson think Toronto is a US city on 'Jeopardy!'?," *TIME*, February 16, 2011, http://techland.time.com/2011/02/16/why-did-watson-think-toronto-is-a-u-s-city-on-jeopardy/. The correct answer is Chicago.

16. The news story was from March 16, 2016. Repeating the experiment on October 2, 2016 yielded the same result. Note that "Barney" approximates a Bengali pronunciation of the name Bernie.

17. Ahn Nguyen, Jason Yosinski, and Jeff Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images" in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, https://arxiv.org/pdf/1412.1897.pdf.

18. Bill Vlasic and Neal E. Boudette, "As US investigates fatal Tesla crash, company defends autopilot system," *New York Times*, July 12, 2016, http://www.nytimes.com/2016/07/13/business/tesla-autopilot-fatal-crash-investigation.html?_r=0.

19. The Obama administration recently released a set of autonomous vehicle guidelines and articulated the expectation that autonomous vehicles will "save time, money, and lives." See Cecilia Kang, "Self-driving cars gain powerful ally: The government," *New York Times*, September 16, 2016, http://www.nytimes.com/2016/09/20/technology/self-driving-cars-guidelines.html.

20. Canadians could be forgiven for thinking this is a mistake many US citizens might also make. The Open Mind Common Sense Project, initiated by Marvin Minsky, Robert Speer, and Catherine Havasi, attempts to "crowdsource" common sense by using Internet data to build network graphs that represent relationships between concepts. See Catherine Havasi, "Who's doing common-sense reasoning and why it matters," *TechCrunch*, August 9, 2014, https://techcrunch.com/2014/08/09/guide-to-common-sense-reasoning-whos-doing-it-and-why-it-matters/.

21. Chomsky introduced the "poverty of the stimulus argument" for why the ability to acquire language must be an innate capability "hard-wired" into the human brain.

22. Alison Gopnik, "Can machines ever be as smart as three-year-olds?" edge.org, https://www.edge.org/response-detail/26084, accessed October 24, 2016. This point is sometimes called Moravec's Paradox after Hans Moravec, who wrote in his book *Mind Children* (Harvard University Press, 1990): "It is comparatively easy to make computers exhibit adult-level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility."

23. Adrian Crockett, "What investment bankers can learn from chess," Fix the Pitch, http://fixthepitch.pellucid.com/what-investment-bankers-can-learn-from-chess/, accessed October 24, 2016.

24. Garry Kasparov, "The chess master and the computer," *New York Review of Books*, February 11, 2010, http://www.nybooks.com/articles/2010/02/11/the-chess-master-and-the-computer/.

25. Regarding truck drivers, see "Self-driving trucks: What's the future for America's 3.5 million truckers?," *Guardian*, June 17, 2016, https://www.theguardian.com/technology/2016/jun/17/self-driving-trucks-impact-on-drivers-jobs-us. Regarding radiologists, see Ziad Obermeyer and Ezekiel Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *New England Journal of Medicine* 375 (September 29, 2016): pp. 1,216–1,219. The authors of the latter piece comment that, because their work largely involves interpreting digitized images, "machine learning will displace much of the work of radiologists and anatomical pathologists."

26. See Gill Pratt's comments in the Aspen Ideas Festival discussion, "On the road to artificial intelligence," http://www.aspenideas.org/session/road-artificial-intelligence. For a primer on the National Highway Traffic Safety Administration's levels for self-driving cars, see Hope Reese, "Autonomous driving levels 0 to 5: Understanding the differences," *Tech Republic*, January 20, 2016, http://www.techrepublic.com/article/autonomous-driving-levels-0-to-5-understanding-the-differences/. Briefly, level 0 means traditional cars with no driver assistance; today's cars with "autopilot" mode are considered level 2; level 4 means "fully autonomous"; and level 5 means no steering wheel—in other words, full automation with no possibility of human-computer collaboration.

27. See *Thinking, Fast and Slow* by Daniel Kahneman. Interestingly, Kahneman comments here that Meehl was a hero of his. Michael Lewis's celebrated book *Moneyball* (W. W. Norton & Company, 2004) can be viewed as an illustration of the phenomenon Meehl discovered in the 1950s, and that Kahneman and Tversky's work helped explain. In a profile of Daniel Kahneman, Lewis commented that he was unaware of the behavioral economics implications of his story until he read a review of his book by the behavioral economics pioneers Richard Thaler and Cass Sunstein. See Richard Thaler and Cass Sunstein, "Who's on first," *New Republic*, August 2003, https://newrepublic.com/article/61123/whos-first; and Michael Lewis, "The king of human error," *Vanity Fair*, December 2011, http://www.vanityfair.com/news/2011/12/michael-lewis-201112.

28. For more information on Affectiva, see Raffi Khatchadourian, "We know how you feel," *New Yorker*, January 19, 2015, http://www.newyorker.com/magazine/2015/01/19/know-feel. For more information on measuring group intelligence and the relationship between social perception and group intelligence, see James Guszcza, "From groupthink to collective intelligence: A conversation with Cass Sunstein," *Deloitte Review* 17, July 2015, http://dupress.deloitte.com/dup-us-en/deloitte-review/issue-17/groupthink-collective-intelligence-cass-sunstein-interview.html.

29. For a description of Kreatsoulis's work, see Abhinav Sharma, "Can artificial intelligence identify your next heart attack?," *Huffington Post*, April 29, 2016, http://www.huffingtonpost.com/abhinav-sharma/can-artificial-intelligen_2_b_9798328.html.

30. For information on sociometric badges, see Alex "Sandy" Pentland's book *Social Physics* (Penguin Books, 2015) and his April 2012 *Harvard Business Review* article "The new science of building great teams," https://hbr.org/2012/04/the-new-science-of-building-great-teams. For a discussion of research linking physicians' communication styles with the likelihood of being sued for malpractice, see Aaron Carroll, "To be sued less, doctors should consider talking to patients more," *New York Times*, June 1, 2015, http://www.nytimes.com/2015/06/02/upshot/to-be-sued-less-doctors-should-talk-to-patients-more.html.

31. For a brief introduction to the legal doctrine of disparate impact, see Ian Ayres, "Statistical methods can demonstrate racial disparity," *New York Times*, April 27, 2015, http://www.nytimes.com/roomfordebate/2015/04/27/can-discrimination-exist-without-clear-intent/statistical-methods-can-demonstrate-racial-disparity. Ayres, a Yale Law School professor, has authored and co-authored several law review articles exploring the concept.

32. Amit Datta, Michael Carl Tschantz, and Anupam Datta, "Automated experiments on ad privacy settings," *Proceedings on Privacy Enhancing Technologies 2015*, no. 1: pp. 92–112, https://www.degruyter.com/view/j/popets.2015.1.issue-1/popets-2015-0007/popets-2015-0007.xml; Latanya Sweeney, "Discrimination in online ad delivery," January 28, 2013, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240.

33. "I think it's time we broke for lunch . . . ," *Economist*, April 14, 2011, http://www.economist.com/node/18557594. In the October 2016 *Harvard Business Review* article "Noise: How to overcome the high, hidden cost of inconsistent decision making" (https://hbr.org/2016/10/noise), Daniel Kahneman and several co-authors discuss the ubiquity of random "noise" resulting in inconsistent decisions in both business and public policy. The authors discuss the benefits of using algorithms as an intermediate source of information in a variety of contexts, including jurisprudence. They comment, "It's obvious in [the case of making parole decisions] that human judges must retain the final authority for the decisions: The public would be shocked to see justice meted out by a formula."

34. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Angwin discusses the Wisconsin Supreme Court decision in "Make algorithms accountable," *New York Times*, August 1, 2016, http://www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html.

35. An academic paper that appeared after both the Angwin article and the Wisconsin Supreme Court decision proves that no realistic scoring model can simultaneously satisfy two highly intuitive concepts of "fairness." Continuing with the recidivism example: A predictive model is said to be "well-calibrated" if a particular model score implies the same probability of re-arrest regardless of race. The recidivism model studied by the Angwin team did (by design) satisfy this concept of fairness. On the other hand, the Angwin team pointed out that the false-positive rate for blacks is higher than that of whites. In other words, the model judges blacks who are not re-arrested to be riskier than whites who are not re-arrested. Given that the fact that the overall recidivism rate for blacks is higher than that of whites, it follows by *mathematical necessity* that a well-calibrated recidivism model will fail the Angwin team's criterion of fairness. See Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent trade-offs in the fair determination of risk scores," September 19, 2016, https://arxiv.org/abs/1609.05807.

36. For a complementary perspective, see Kate Crawford and Ryan Calo, "There is a blind spot in AI research," *Nature*, October 13, 2016, http://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805. The authors comment that "Artificial intelligence presents a cultural shift as much as a technical one . . . We need to ensure that [societal] changes are beneficial, before they are built further into the infrastructure of everyday life." In a *Wired* magazine interview with Barack Obama, MIT Media Lab director Joi Ito expresses the view that artificial intelligence is "not just a computer science problem," but requires input from a broader cross-section of disciplines and perspectives. Scott Dadich, "Barack Obama, neural nets, self-driving cars, and the future of the world," *Wired*, November 2016, https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/.

37. Douglas C. Engelbart, *Augmenting human intellect: A conceptual framework*, October 1962, http://www.dougengelbart.org/pubs/augment-3906.html.

38. Nicely dating the story is the fact that Stuart Brand, the editor of the *Whole Earth Catalog*, was the event's cameraman. See Dylan Tweney, "Dec. 9, 1968: The mother of all demos," *Wired*, December 9, 2010, https://www.wired.com/2010/12/1209computer-mouse-mother-of-all-demos/. This demo "also premiered 'what you see is what you get' editing, text and graphics displayed on a single screen, shared-screen videoconferencing, outlining, windows, version control, context-sensitive help and hyperlinks."

39. Jobs made this comment in the 1990 documentary film *Memory & Imagination: New Pathways to the Library of Congress*. Clip available at "Steve Jobs, 'Computers are like a bicycle for our minds.' - Michael Lawrence Films," YouTube video, 1:39, posted by Michael Lawrence, June 1, 2006, https://www.youtube.com/watch?v=ob_GX50Za6c.

**Deloitte.**
University Press

Follow @DU_Press   #DeloitteReview

Sign up for Deloitte University Press
updates at DUPress.com.

Get every
issue on your
mobile device

Deloitte
Review