Human-Like Implemented A.I. and Human Problem-Solving A.I.

Russ McBride
Department of the Management of Complex Systems
Gallo School of Management
University of California, Merced
United States

Abstract

Artificial intelligence is cycling into another peak of enthusiasm right now with long-time evangelists like Kurzweil redoubling their hyperbole and Elon Musk suggesting that the odds are "a billion to one" that we are not already living in an AI simulation matrix. There is a new wave of attention within the formerly silent fields of economics, strategy, and management from those like Agrawal, Gans, and Goldfarb (2018) who see the lower prices of AI-powered "prediction" reshaping entire industries. Others more skeptical see genuine human reasoning of the kind needed to, e.g., make firm-level strategic decisions, impossible for machines to duplicate and safe from encroachment for the foreseeable future. In this paper, I look at whether we have achieved human-like implemented AI (HLI-AI), a question which requires an exploration of what cognition and intelligence fundamentally are. I then look at what would be needed to build it, and then suggest a distinction between HLI-AI and human problem solving AI (HPS-AI)—the former is what we do not currently have, the latter is implemented in a wide variety of (non human-like) techniques that solve human-relevant problems. Finally, I suggest a way forward for reasonable expectations of the role for HPS-AI in the socio-techno world.

Introduction

We are in another peak of enthusiasm for artificial intelligence. AI-startups are pouring into entrepreneurship incubators like TechStars and Y-Combinator. AI-based companies are getting venture capital funding at record rates. And the popular press is awash in stories about the forthcoming mass extinction of jobs by AI systems, Kurtzweil's (2001) prediction of the merging of human consciousness and AI systems by 2045, and Elon Musk's declaration that we are already living a simulation matrix, all fueled the seemingly incontrovertible evidence of self-driving cars, object identification systems that surpass human object identification, IBM's Watson winning Jeopardy, and Google's AlphaGo beating the best Go players in the world. Surely, we are on the very cusp of AI not just facilitating most aspects of our everyday interactions, but reshaping economic landscapes, strategic decision-making, and everything from our sex lives with sex-specialist robots to the absorption of our very consciousness at the point of Singularity. This seeming inevitability has spurred worries of a Terminator-like future and solving the problem of how to program morality into our AI offspring is seen to be of critical importance before they gain full independence from us.

According to Lotfi Zadeh, the founding father of one AI technique known as "fuzzy logic", in 1956 a cover story in the NY Times predicted that every home would have a robot attending to the dishes and the laundry within five years. Since then the predictions and the hyperbole have never stopped and cycled through peaks and troughs over the decades. Alan Turing (1950) said, "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted." In 1965, Herbert Simon said, "Machines will be capable, within twenty years, of doing any work that a man can do." And, Marvin Minsky suggested (1965) that, "within a generation… the problem of creating 'artificial intelligence' will be substantially solved." Turing, Simon, and Minsky turned out to have overestimated AI and overestimated the path of progress from Turing's Universal Turing Machine.

The refrain remains the same: "Yes, past predictions were overly optimistic but *this time* we are certain that AI will take over the world." Famously, the CIA during the Cold War was deeply interested in anything that might help automate the translation of decrypted Soviet messages, but early attempts at translating a sentence like, "The spirit is willing but the flesh is weak", resulted in "The vodka is good but the meat is rotten". The failed return on the huge investments in AI by the US and British governments led to a report to the UK Parliament from Sir James Lighthill in 1973, dubbed "The Lighthill Report", that clarified that depth and extent of AIs failures and all but entirely ended funding for AI and precipitated the first AI Winter of the 1980's. "Artificial Intelligence" would remain a dirty term for decades to come. No self-respecting researcher would describe their work as "artificial intelligence" but rather as

pattern recognition, object identification, search algorithms, Bayes Nets, evolutionary algorithms, convolutional neural networks, back propogation optimization, or any of a variety of other more specific techniques.
AI is once again sexy and startups are no longer shy about using the term in their mission statements and self-descriptions. Indeed, whereas previously *nothing* was AI, now all those computational techniques are unabashedly AI.
Have we achieved human-like AI? If you take the popular press seriously the answer is, "yes". fMRI machines can now "read your thoughts", Facebook built two AI systems that "developed their own secret language" only they could understand, AlphaGo has not just matched human creativity but exceeded it with novel Go moves, and we already have self-driving vehicles with fewer accidents per mile than humans.

In what follows I want to step back and look at whether we have actually attained human-like implemented AI (HLI-AI) and suggest, along with others like the world's leading roboticist Rodney Brooks of MIT, that we have not. Doing so requires getting a bit clearer about what exactly human intelligence and cognition actually are in the first place. Toward that end I will go through a primer on what may be the best theory of the human mind so far. Doing so will show clearly that we haven't achieved HLI-AI and what it will take to achieve HLI-AI in the future. It will also make clear the source of the common errors among experts and layperson alike. The most common errors are thinking that the human mind is just a computer (not an infrequent refrain even among neuroscientists), thinking that we have AI systems that are built "just like" the human mind, or that the human mind has a one-dimensional scale of power (in the form of bits processed per second) whichAI systems will surpass with CPU speed improvements. Additionally, it will make clear the naive optimism of Doug Lenat and data mining proponents who say that, "all we need is more data". Rather, we don't have a quantity problem; we have a theory quality problem on both the computer side and the human side of the fence.

Still, AI has made strides recently and the responsibility for those advances does, it is true, lie in both the huge increases in data and the modest increase in CPU speeds. But there has been no advance in algorithms for about 35 years, or as Marcus (2016) said more bluntly, "there has been no progress" [in the quest to create HLI-AI full stop]. The advances we see are not HLI-AI, they are advances in our computational efforts to solve problems that humans care about, they are advances in human problem-solving AI (HPS-AI) which utilize an enormously wide variety of relatively old computational techniques. IBM's Watson is a veritable everything-but-the-kitchen sink approach to building an AI system, although unfortunately one that hasn't generalized to a successful and profitable medical diagnosis system as IBM had hoped. We have hacked and brute-forced our way into better object recognition and better self-driving cars. But there are limits to these hacks and limits to HPS-AI in general and it's important to understand what they are.

After showing a path forward for the achievement of true HLI-AI in the future, I attempt to locate a more realistic and less hyperbolic place for HPS-AI and find a useful place for it in a socio-techno integrated world where AI systems are less like independent and self-reliant persons and more like goal-specific systems that become deeply integrated into the habits of our lives, our firms, and the social systems that we are enmeshed within.

The Computational Theory of Mind

We know surprisingly little about how the mind works. The currently most popular theory of the mind began as a guess using the computer as a powerful metaphor. The universal Turing machine was seen not just as an answer to the Hilbert's Entscheidungsproblem, or a precise specification of 'computation' for which mathematics was in desperate need, but almost immediately viewed both as something that could manifest human-like intelligence one day and conversely a powerful formalism that extended beyond mere metaphor, but *actually* describe what the mind was—literally a computer.

This wasn't the first metaphor used to understand the mind. The early Myth of Golem suggested that humans were clay with souls magically infused by a wise blacksmith or God himself. Later metaphors grew from advanced hydraulic mechanical system with microscopic levers and tubes that drove the "robots" in the Palace of Versailles, creatures that appeared to be alive. The advance of deterministic science reinforced this idea of the mind as some kind of deterministic device too and set the stage for the Computational theory of mind where the "soul" is the software and the body, made of clay, hydraulics, or other mechanical components is mostly irrelevant.

The other available option for a theory of the mind at the time of Turing's invention was Norbert Weiner's "cybernetics" theory (1965) but most found the hundreds of pages of mathematic integrals overwhelming while the universal Turing machine, by contrast, was based upon simple first order logic. A universal Turing machine could be implemented on almost any type of hardware—brain tissue, tin cans, erasable tape, or silicon. Similarly, the thinking went, the human mind needed some kind of hardware upon which to be implemented but the critical work was done by the software.

The Computational Theory of Mind implied a future where with enough work we could decode the software that constituted the "code" of the mind. Just as a computer manipulates representational symbols, the mind manipulates representational symbols. Those symbols are processed according to rules and those rules were taken to be the very same rules of first order logic. It's all very neat and tidy and doesn't suffer from the messy mathematics of Weiner's cybernetics.

The Computational Theory of Mind (CTM) proved too simple and attractive for the world to ignore. Most of the greatest thinkers of the day couldn't resist its appeal and enthusiastically and explicitly embraced it as a grand unifying theory of the mind and cognition that swept across psychology, cognitive science, linguistics, organizational behavior, economics, and affected almost every known field in some way or other.

What exactly does CTM postulate?

1) That the mind is a computer.
2) As such, the hardware implementation of the mind is mostly irrelevant; it's the software that counts.
3) The fundamental unit that is manipulated by the software is the representation.
4) Representations are discrete entities that can isomorphic ally map to objects in the real world.
5) Human action is the result of retrieving the relevant representations and performing inferences upon them in a way so as to plan the next behavior.
6) As such, humans are essentially information processing devices too.

The Embodied Cognition Theory of Mind

Unfortunately, the last few decades have proven the Computational Theory of Mind to be false. To Quote Robert Epstein (2016): "Your brain does not process information, retrieve knowledge or store memories. In short: your brain is not a computer."

Here is what we are not born with: information, data, rules, software, knowledge, lexicons, representations, algorithms, programs, models, memories, images, processors, subroutines, encoders, decoders, symbols, or buffers – design elements that allow digital computers to behave somewhat intelligently. Not only are we not born with such things, we also don't develop them – ever.

We don't store words or the rules that tell us how to manipulate them. We don't create representations of visual stimuli, store them in a short-term memory buffer, and then transfer the representation into a long-term memory device. We don't retrieve information or images or words from memory registers. Computers do all of these things, but organisms do not (p2).

Epstein's claims might appear bold but are part of the larger, evidence-based theory—the embodied cognition theory of mind—that rejects CTM and the fundamental analogy that the mind is to the software as the body is to the hardware (Lakoff 1987; Brooks, 1992; Lakoff and Johnson 1999; Varela, Thompson, and Rosch, 1991; Talmy, 2000; Barsalou, 1999; Talmy, 2000; Nöe, 2004; Matlock, 2004; Gallagher, 2005; Gallese& Lakoff, 2005; Wheeler, 2005; Feldman, 2006; Dreyfus, 1979, 1992, 2007; Nanay, 2016).

Embodied cognition does not have a unified formalism like CTM does with the Turing Machine. It rejects the fundamental tenets of CTM

1) The mind is *not* a computer.
2) As such, the hardware implementation of the mind is *not* irrelevant.
3) Representations are important but whatever they turn out to be they are not simple, discrete units manipulatable by some "language of thought" or simple logic.
4) Representations are *not* discrete entities that isomorphic ally map to objects in the real world.
5) Human action is *not* mererly the result of retrieving the relevant representations and performing inferences upon them in a way so as to plan the next behavior.
6) Humans trade in information but are not merely or essentially information processing devices.

The fundamental unit is not the abstract representational symbol, but rather human experience instantiated in neural activation patterns and it is from human experience and neural activity that the vast and diverse canopy of human cognition derives. There is a broad range of empirical work supporting this claim. Buccino G., Riggio L., Melli G., Binkofski, F. , Gallese V., and Rizzolatti G. (2005), Pulvermueller, F., M. Haerle, & F. Hummel (2001), and Tettamanti, M., Buccino, G., Saccuman, M.C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S.F. and Perani, D. (2005), have all found that thinking of some motor activity activates areas of the motor cortex that fire when that activity is actually executed. Thinking about chewing activates the motor cortex involved in actual chewing. Thinking about kicking activates the portion active when actually kicking.

And reading about grabbing activates the motor cortex actually involved when grabbing. There is also extensive work built from response-time tests that supports this view (e.g., Stanfield &Zwaan, 2001; Zwaan, Stanfield, Yaxley, 2002).

This should strike us as nothing short of profoundly shocking. We *understand* a word that we read, hear, or speak because that word stimulates the very neural activity involved in the activity described by the word. We *understand* the word's concept because it is grounded in human experience of the activity described by the concept.

We *simulate* the experience. This is the simulation theory of semantics, or 'simulation semantics' for short. If I ask you whether it's possible to fit 3 chicken eggs in your mouth you will (if you're like most people) *imagine* yourself attempting to put them in your mouth and in doing so you will activate the very same neural structures that are active when you are actually doing it, even though you've never done it before. Simulation semantics is the fundamental tenet of Embodied Cognition and makes it clear why the body is not merely a "hardware layer" but critically important to cognition—the body is the most prominent conduit for human experience that is then later simulated when only imagined or cognized instead of acted. Indeed, one view of emotions (Prinz, 2004) is that they are merely a generalized result of the state of one's body and that when we are thinking about such emotions we are re-imagining such states.

A common retort is: "That's fine when one is *thinking about bodily activities* but the vast majority of thinking involves high-level, abstract concepts that have nothing to do with bodily experience, like unemployment, irrational numbers, hierarchical governance, or the weak electromagnetic force." As it turns out extensive research has been done that shows just how high-level concepts are ultimately grounded, mostly through neurally instantiated metaphorical mapping, in human bodily experience (especially Lakoff & Johnson, 1987, 1999, and Lakoff and Nuñez's *Where Mathematics Comes From*, 2001). A sentence like, "the relationship was moving to fast for me" rely on a metaphorical mapping of the abstract concept, 'relationship', to vehicle speed for which we have extensive bodily experience. "France fell into a recession and couldn't find a way out" maps the high level concept of 'recession' to the experience of falling into a physical depression.

Others, like Nanay (2016) have argued that decision-making itself is fundamentally a simulation process too. When deciding whether you want to take a new job and move to a different city you *imagine* yourself in the new city and compare that *experience* to your experience of your current city. If not just bodily concepts, but abstract concepts, reasoning, and decision-making are fundamentally based upon *human experience* then we face an obvious hurdle to any AI system built upon representational symbols that are *not* grounded in such experiences—such symbols will have no meaning for them.

Simulation semantics' greatest strength is precisely AI's greatest weakness in that there is no way to explain how discrete, disconnected symbols in a computational processing device get any *meaning*. This has come to be known as the "symbol grounding problem", a variant of which is known as the "frame problem". Google's AI language translator that translates from Mandarin to English doesn't *understand* Mandarin. It's just manipulating symbols according to pre-established rules. Searle's Chinese Room Argument showed (1980) that you can't get from syntax to semantics and no seemingly brilliant translation computer understands *anything* that it is "saying". It's not just that the human mind is not computational but rather that the fundamental organizing principle of the human mind bares no relation to the fundamental process of computation—the manipulation of discrete representations that are not grounded in human experience. Far from being an irrelevant implementation layer, the "bodily hardware" is the indispensable and the essential means through which all of mind and cognition is structured.

Reflect for a moment upon how we learn best. We learn easily and most rapidly when we are conscious and having a vivid experience of the material we're studying. Those cassette tapes that attempted to teach you a language while you were sleeping aren't around anymore because it turns out that learning while you are sleeping isn't so effective. Most of our cognitive work is unconscious but most of those unconscious abilities and knowledge got their through conscious experience. Conscious human experience is critical for learning, structuring our knowledge, making decisions, engaging socially, and achieving intellectual feats.

How to Build Real HLI-AI

It should now be readily apparent then that no research lab or firm's R&D department has built human-like instantiated AI (HLI-AI) since no computer has human-like experience. We have not replicated human-like cognition. For all the novels and movies and stories about a sentient artificial systems, they simply do not exist. One might think that all we need to do is replicate the very neural mechanisms that cause human experience in humans (and other animals which presumably have such experiences too). There's just one problem—we currently do not have the slightest clue how human experience comes to be. In philosophy this is known as "the problem of qualia" or simply as "the hard problem" or "the problem of consciousness". Many have tried to find neural correlates of consciousness but those who have dedicated their careers to this problem have either given up in the face of abject failure or gone partially mad trying.

So, how then might we begin to approach the problem of building HLI-AI? Well, we might not have the faintest clue how to build an experiential machine but we can build the correlates of the conduits of bodily experience—namely the sensory modalities.

We can't build actual vision or haptics or hearing or taste or sensorimotor proprioception into a machine but we can build the their correlates in the form of a wide array of electromagnetic radiation sensors, sound sensors, etc. This has started to happen with self-driving vehicles and robots which are employing more and more sensor systems of different types. The more sensory systems the more the AI system can use those diverse modalities to triangulate upon features, patterns, and objects in the external world and begin to utilize that diversity in its computations. Mosquitos can detect CO2 emissions. Bats can echolocate with sonar. Humans have a wider array of sensory modalities but not those. Generally, evolution has seen fit to endow the creatures with the greatest number of sensory inputs the greatest degree of qualitative experience. Amoeba and protozoa have limited input systems and (as far as we can tell) limited experience, while spiders and centipedes are more advanced, and mammals like dogs, dolphins, and humans at the far end of the spectrum, lacking things like infrared detection but possessing a wide array of sensory systems. Why does a greater diversity of input systems seem to correlate with a higher degree of conscious experience? We have no clue; it just seems to be one of those empirical facts the cause of which we will probably only understand when we understand *what* the neural correlate of conscious experience actually is.

So, adding diverse input systems into an AI engine, unlike with living organisms, won't get us very far along the path to true human experience but it will at least allow us to establish one of the empirically necessary stepping stones to move in the direction of a future HLI-AI.

The Most Common Errors in AI

Given the fundamental structural difference between AI and human cognition we can see that the two most egregious errors are 1-thinking that the human mind is a computer, i.e., that CTM (the Computational Theory of Mind) is correct; and 2-thinking that a computer is a human mind, i.e., that it is engineered the same way that the human mind is, or implements a "partial" human mind. Neither is true and until we see a way forward to engineering human experience from scratch they will remain false and HLI-AI nothing but a distant dream.

There are, however, many who believe that first-order logic *is* how the mind works, that the mind has discrete symbols, and that all we need is *enough data* and a computational system will magically sprout genuine human intelligence (HLI-AI). The implicit premise behind this claim is that we *already* know exactly how the mind works and it works more or less like a computer, a computer hungry for more data (coincidentally enough). Kurzweil has said (2001), in a book entitled, *The Age of Spirtual Machines* that "We have already reverse engineered the cerebellum, and by 2029, reverse engineering of the brain will be complete." Someone should have told the neuroscientists that we already finished the cerebellum and that they can continue "reverse engineering" the rest of the brain.

If you believe that we already understand how the mind works and that it works like a computer than you wind up saying some deeply dubious things. Doug Lenat said, "once you have a truly massive amount of information integrated as knowledge, then the human-software system will be superhuman" (Lenat, 2001). Lenat has received millions of dollars of investments for CYC and expended millions of man-hours inputting sentences, translated into first order logic, from newspapers and other media like, "cats have four legs", "the German economy is on the rise", and "Albany is the capital of New York". The hope? That once some threshold is crossed—perhaps at 1 billon or maybe 1 trillion sentences?—that genuine consciousness will magically spring to life… in some extended network that lacks eyes, ears, emotions, proprioceptions, or any sensory inputs at all.

It's tempting to think that this is an isolated incident of deep confusion by some extremist but current discussions of AI are awash in the glory of big data with many suggesting that AI has only failed so far because we have not had the data or the CPU power to process that data (using exactly the same algorithms that we've historically used to process that data). Kurzweil (2001) has famously suggested that "brain power" can be represented in terms of how many "instructions" per second can be processed. "MIPS" is "millions of instructions per second" which computers achieved around 1970. By 2020 he says that computer will achieve human intelligence because they will exceed the speed of human information processing (see figure 1).

Unfortunately, he never says *how* it's possible to quantify the number of instructions per second a brain can process. To think that the brain can be analyzed in terms of discrete processing instructions is to presume that the human brain is simply a Universal Turing machine. Of course, this is false. Figure 1 (from Kurzweil, 2001)

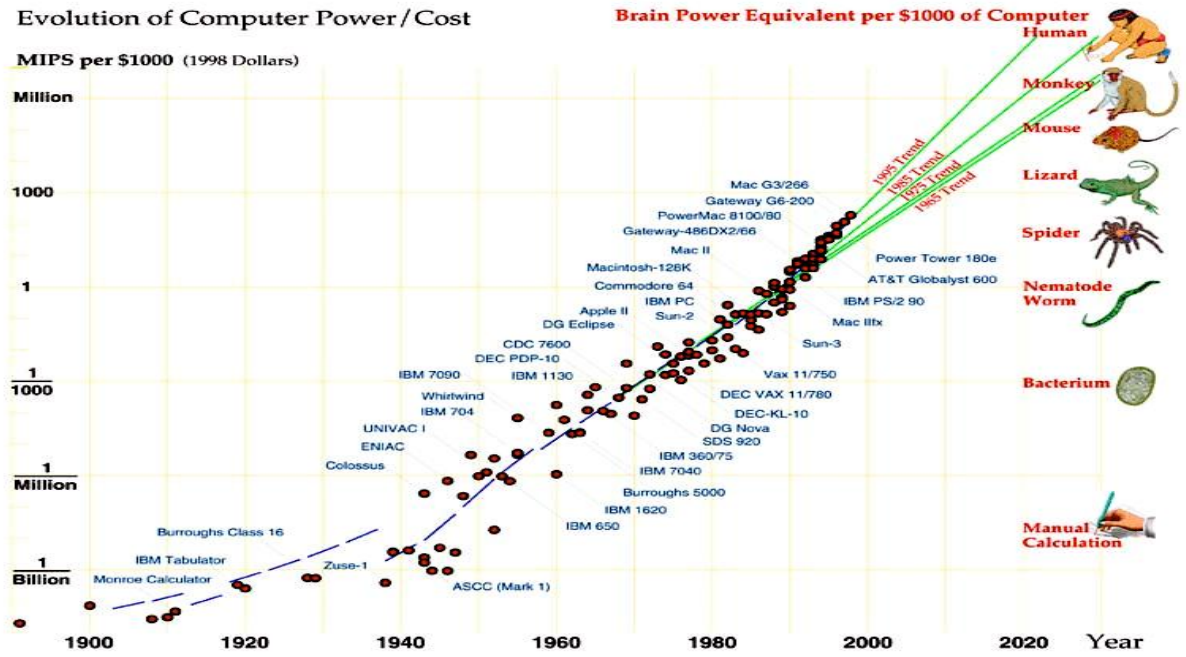

Figure 1 (from Kurzweil, 2001)

We know the brain is essential to cognition and consciousness but there is broad misunderstanding about how advanced neuroscience is. Neuroscience has no clue how we experience the taste of black pepper or the mechanisms by which general anesthetics render us unconscious but also nothing about how "computations" are carried out… other than a claim that neurotransmission seem to be important. Even here some in neuroscience suggest that it's not the neurons we should be focusing on but rather the otherwise ignored glial cells!

There is no doubt that computers can do interesting things and have advanced our productivity in all sorts of ways. And that recent advances have been made possible by better access to bigger data. But processing speed has long since fallen short of Moore's Law that predicted a doubling of processing speed every 18 months, and we have not had any real breakthrough in new algorithms for 35 years.

Object identification systems in images have made great strides recently (see figure 2) with the best systems seen as having surpassed the "5% human error rate for object identification" in 2015. It turns out the claimed 2% error rate of such AI systems is not quite accurate. To "correctly identify" an object on the standard Imagenet recognition test a system just needs to list the correct object as one of five possible objects in order of priority, so when shown an image of a camel the system's response is deemed correct if the system responds with: "penguin, butcher knife, dog, camel, hay bail"—the 4th out of 5 suggestions.

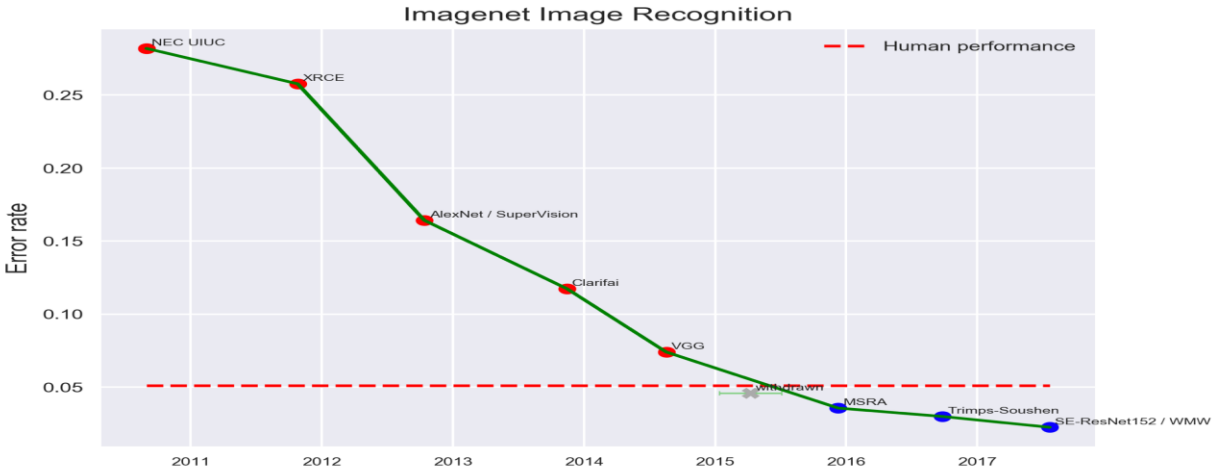

Figure 2 (image credit: the Electronic Frontier Foundation)

Further, "adversarial attacks" on image classification systems have shown that the classification techniques employed by AI systems bear no resemblance whatsoever to human classification. A image correctly classified as a "camel" when modifications to its pixels are made that are utterly undetectable to humans might get classified by the AI system as a "tree". Undetectable pixel modifications caused a system to classify a school bus and a temple both as an "ostrich". These same kind of adversarial attacks can occur with pattern recognition systems that process audio or text or, frighteningly enough, the image received by a self-driving car from a stop sign to make it look to the system like a green stop-light instead.

Additional data does not help block such adversarial attacks. Indeed, additional data can make it worse and reveal biases in the data pool that was not otherwise known, as when a Google classifier embarrassingly mis-categorized the image of an African American male as a "gorilla", an error that stemmed from the paucity of African Americans in the large Facebook derived data set Google utilized. "We're still very, very far from visual intelligence, understanding scenes and actions the way humans do" (Ali Farhadi of the University of Washington, quoted in Lohr, 2016).

Part of the problem is that these classification systems rely upon a single input source, an image in a binary file format, while humans have a rich multi-modal and experience-laden history with most images that it encounters which give humans a panoply of other ways of triangulating upon an object so as to more likely identify it.

Computers do not "think" like humans. And more data will not magically enable computers to think like humans. It is not a quantity problem. It is a quality problem—a problem of understanding exactly how human cognition works. Right now, we only have one or two pieces of the puzzle of human cognition in place and both point to multi-modal sensory-rich human *experience* being critical, something that no AI system possesses, or has any hope of possessing in the near future. The obsession among software engineers to build human-like AI does nothing to help actually build it since in order to build it, we must first understand human cognition and that is not the work of engineers but rather the purview of cognitive scientists, neuroscientists, psychologists, philosophers, and anthropologists.

Human-Like Implemented AI (HLI-AI) vs. Human Problem-Solving AI (HPS-AI)

HL-AI is AI implemented *in the very same way* as human cognition. Getting clear on the question about whether any AI is implemented the way human cognition is implemented necessitated a dive into the structure of human cognition. We know very little about it, but do know that it's structured around the diversity and multi-modality of human experience.

This point alone highlights the fact that HLI-AI does not exist and that we are not close to a time when it will exist (contrary to Kurzweil, Lenat, the singularity evangelists, and others). There has, however, been great success at solving problems that we as humans are interested in (HPS-AI). The range of problems solved spans the mundane, like text editors, to advanced credit card fraud detection systems, self-driving vehicles, and translation engines. These HPS-AI systems are similar to enormously useful inventions like the steam engine which generated power in a way we thought only horses could previously, or the breakthrough in extracting oil from shale reserves. HPS-AI, similarly, offers a way to achieve some goal we deem relevant. But there's no confusion that the steam engine actually replicates the human ATP energy production system, or that shale oil extraction works the same way that we squeeze juice from an orange.

Similarly, there should not be any confusion that Google's language translation engine works like human translation, or that AlphaGo is implemented the way humans play Go. The goals are the same across machine and human; the implementations are radically distinct, and that's perfectly fine. A steam engine beat John Henry over a hundred years ago. We shouldn't be surprised that AlphaGo recently beat the world's best Go players.

Tremendously useful goals can be achieved by manufactured systems—whether the steam engine or AlphaGo. But, for whatever reason, goals that are more "intellectual" when achieved by artificial systems are immediately seen by media and experts alike as indicative of "true" human intelligence. Engineers keep trying everything they can to squirm away from the most obvious fact about human cognition—that it involves human experience caused (mysteriously) by the interactions of our bodies in the environment, driven by wet grey matter. Our current AI systems, HPS-AI, are not HLI-AI and we don't yet have a path forward for how they could become HLI-AI.

Finding Discriminating Alignment for HPS-AI

HLI-AI is off the table. So, instead of conflating HLI-AI and HPS-AI, how might we understand HPS-AI better and more reasonably so as to understand its proper domain of application and find discriminating alignment among the various goals that we as humans find relevant on the one hand, and the *kinds of goals* that HPS-AI systems have a shot of achieving?

After all, firms need to make strategic decisions about the future of their businesses, investors need to make investment decisions, Engineers need to choose projects that have a chance of completion, and consumers need to choose products that work. Some domains of application offer better chances of success for HPS-AI (which I will now simply refer to as 'AI') systems than others. We need to understand where it will excel and where it won't.

Agrawal et al. (2018) offer an economist's, and less hyperbolic, view of AI suggesting that its fundamental effect is a reduction in the cost of "prediction". As the cost of paying a translator to convert a page of Spanish to English falls from $300 to $0 this has enormous effects on a wide range of economic activities. "Prediction costs", they claim, will continue to fall. But there are a couple of oddities here: a) They have an idiosyncratic view of 'prediction' such that any system that reduces knowledge uncertainty, not just future uncertainty, is engaged in what they call 'prediction'. b) There are no suggestions about what domains will see the greatest decline in "prediction costs"; they offer no discriminating alignment.

Let's attempt to provide some discriminating alignment for AI systems, but without the baggage of errors in thinking that such systems are implementing *human-like* AI, HLI-AI.

System That Will Excel

1- Information processing and vs. environmental interaction.

One need only look at the state of robots today compared to the grand predictions of the 1950's to see how difficult progress has been here. Turing would've never believed that our best selling robots two decades into the third millennium would only be a floor cleaner and some assembly line machines. Interacting with the real physical world is extraordinarily difficult. The invention of a robot cockroach that could forage for food, evade predators, reproduce, and engage in rudimentary communication with its fellow iCockroaches would garner a science medal and be the invention of the century. Despite trillions of dollars of precisely built manmade roadways, providing an incredibly controlled and limited environment (unlike the cockroach's environment), self-driving vehicles are still not quite yet fully automated. The greatest recent success in environmental interactions, arguably, has been SpaceX's vertically landing rockets—it took 70 years from the time of the invention of the rocket to get it to reliably land vertically so it could be reused. So, for the foreseeable future, systems that do information processing (broadly speaking) will continue to radically outpace those that attempt any kind of advanced interactions with environment. The challenge for those information processing systems is not being able to rely on environmental interactions and all the benefits that provides for true understanding of the world.

2- Environment-interacting systems must have multi-modal inputs.

If you do need to make a bet on some form of AI that is attempting to interact with the environment then, roughly speaking, it must have a wide array of multi-modal input sensors. This is by no means a guarantee of success but rather a necessary prerequisite for its success.

3- Goals that afford clear algorithms.

Some goals afford clear algorithmic steps toward their attainment; others don't. Turing designed the computer as an algorithmic processing machine and they are algorithm processors par excellence. DeepBlue beating Gary Kasporov in chess was seen as a clear sign at AI was actually HLI-AI, but of course chess is a game with precise rules, a precise and abstract table environment, and clear movement options. Winning chess is, in other words, an algorithmic and near ideal goal for an AI system, which Turing himself knew when he programmed the first chess playing software into the giant Bletchy Park vacuum tube computer used for decrypting the Nazi Enigma codes in the 1940's.

It is, of course, often difficult to determine which goals are amenable to algorithmic approaches and which are not, especially when probabilistic and statistical techniques are employed with greater frequency. But insofar as this is possible then those goals amenable to such algorithmic approaches will fare better.

4- Reliable statistical regularities.

The ever-increasing use of statistical techniques, and inductive techniques in general, need phenomena that is reliably approachable via induction and statistics. Famously, regression techniques afford manipulatable grey areas where whether or not there is a reliable correlation can be swayed by the kind of analysis chosen. The housing market crash of 2009 wasn't foreseen precisely because the regression performed saw statistical regularity and overlooked the interdependent nature of the different housing markets. So, inductive formalisms are difficult to employ successfully but the generalization remains true—Finch bird mating patterns are highly regular; bond price time series are highly irregular. Systems built around regularities in the former are much more likely than the latter to succeed.

Systems Less Likely to Excel

The ground work done above also helps us draw boundaries around some areas of *non-alignment* where AI systems will be less likely to achieve interesting goals. Of course most of the goals that we find most interesting are the least likely achievable. Alibaba founder, Jack Ma, has said that "In 30 years, a robot will likely be on the cover of Time Magazine as the best CEO". This is not an uncommon refrain from techno-evangelists—that soon AI systems will be "just like" humans only better in all sorts of ways.

John McCarthy coined the term, "artificial intelligence" in 1955 when a year later he helped organize a 2 month summer computing workshop at Dartmouth College. The workshop, which included other luminaries like Claude Shannon and Marvin Minsky believed that they would have human "common sense" more or less completely reverse engineered and programmable into machines within 8 weeks. More than a half a decade later and we are still nowhere close to "common sense" or "general artificial intelligence" as it is sometimes called. Why not?

1-HPS-AI won't succeed in domains that require possession of human-like concepts.
Humans concepts are the building blocks from which human reasoning, judgment, and decision-making are built. How close are we to human-like concepts? We already know that such concepts are structured around human-like experience. This was the fundamental result of the embodied cognition research movement. But perhaps we have a few rudimentary concepts in place at this point in our enormous technologically advanced state.

The Image net results illustrate otherwise. The currently most advanced AI systems do not and can not yet possess a relatively rudimentary concept like, 'rose'. And we know that they don't possess a human version of the concept of 'rose', despite the 98% object identification rate because modifying a few undetectable pixels in the image does not change our assessment of it as a 'rose' but for an AI classification system that might be enough to change its assessment to 'camel', 'ostrich', or 'school bus'. Now think about the kinds of concepts required to make successful decisions as a CEO, concepts like 'increasing market share', 'collapsing economy', 'employee unrest', 'negotiating motivations', or 'competitive advantage'. If the robot CEO can't deal with the concept of 'rose' without short-circuiting then how well will it do with any of those more abstract concepts? And we haven't even gotten the point of facile manipulation or modification or communication of those concepts. iCEO has a long way to go before it sits in an actual board room.
2- HPS-AI won't succeed in domains that require human-like experience.
This point should be obvious but it's also a less helpful guideline because the question under issue is precisely: "Which goal domains will *necessarily* require human-like experience?" Self-driving vehicles have made good progress without building any actual human experience into the decision chain anywhere.

Conclusion

Arguably, the two best investment systems in the world involve one human (and associates)—Warren Buffet—who trades by simple and intuitive guidelines about the perceived value of a company vs its trading price, where determining value is more or less non-algorithmic. The other is Ray Dalio's Blackwater hedge fund which is fully automated. But here, the rules and the concepts like "rising economy" and "likelihood of the prime interest rate going up" were definitively human and the automated system was built to mimic those concepts through proxies (hacks) as best as possible. With enough work it's possible to hack one's way to an automated version of Buffet's "quality" trading approach as well.

But the point of the paper was to make a muddied distinction clear, the distinction between HLI-AI and HPS-AI. Despite endless proclamations to the contrary we haven't come anywhere near achieving HLI-AI. HLI-AI requires, fundamentally, human experience, if the system is to possess human cognition and human concepts that are structured by such experience.

Still, HPS-AI has made interesting and useful progress. It will continue to do so in domains where the emphasis is on information processing, rather than environmental interaction, although almost *any* progress on the latter affords huge opportunities. Those environmentally interactive systems must, for any chance of success, deploy a wide range of environmental sensors. Obviously, the systems with the best chance will tackle problems that are clearly algorithmic. And, finally, with the ever-increasing use of statistical techniques, successful systems will target domains where there are reliable inductive regularities.

Laying out these areas of discriminating alignment should, it is hoped, offer a more reasonable path forward for understanding the areas where AI will afford a useful advantage and find a more firm position in the socio-techno environment.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, pp 577-660.

Buccino G., Riggio L., Melli G., Binkofski, F. , Gallese V., and Rizzolatti G. (2005) Listening to action-related sentences modulates the activity of the motor system: a combined TMS and behavioral study. Cog. Brain Res. 24: 355-363.

Brooks, R. (1992). Intelligence without representation. *Foundations of Artificial Intelligence*, vol. 47, 139-159.

Dreyfus, H.L. (1979). *What Computers Can't Do*. New York: Harper & Row.

Dreyfus, H.L. (1992). *What Computers Still Can't Do*. New York: Harper & Row.

Dreyfus, H.L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian, *Philosophical Psychology*. 20:2, 247-268.

Gallagher, Shaun. (2005). *How the Body Shapes the Mind*. Oxford: Clarendon Press.

Gallese, V. and Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology*, Vol. 22, no. 3-4, pp.455-479.

Feldman, J. (2006). *From Molecule to Metaphor*. Cambridge, MA: MIT Press.

Kurzweil, R. (2001). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. Penguin Books.

Lakoff, George. (1987). *Woman, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

Lakoff, George & Johnson, Mark. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: Basic Books, a member of the Perseus Books Group.

Lakoff, G. &Nuñez, R. (2001). *Where Mathematics comes from: How the Embodied Mind brings Mathematics into Being*. New York: Basic Books, a member of the Perseus Books Group.

Lohr, S. (2016). A lesson of Tesla crashes? Computer vision can't do it all yet. New York Times Science section, September 19th, 2016.

Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory and Cognition*, 32:8, pp1389-1400.

Nanay, B. (2016). The role of imagination in decision-making. *Mind & Language*, 31:1, pp127-143.

Noë, Alva. (2004). *Action in Perception*. A Bradford Book.

Prinz. J. (2005). *Gut Reactions: A Perceptual Theory of Emotion.* MIT Press.

Pulvermueller, F., M. Haerle, & F. Hummel (2001). Walking or Talking?: Behavioral and Neurophysiological Correlates of Action Verb Processing. *Brain and Language, 78*, 143–168.

Searle, J., 1980, 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3: 417–57

Stanfield, R. & R. Zwaan (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*, 153-156.

Talmy, L. (2000). *Toward a Cognitive Semantics*. 2 Volumes. Cambridge, MA: MIT Press.

Tettamanti, M., Buccino, G., Saccuman, M.C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S.F. and Perani, D. (2005). Listening to action-relatedsentencesactivatesfronto-parietalmotorcircuits. J Cogn. Neurosci. 17: 273-281.

Varela, F.J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind.* Cambridge, Mass: MIT Press.

Weiner, N. (1965). *Cybernetics or Control and Communication in the Animal and the Machine.* MIT Press.

Wheeler, Michael. (2005). *Reconstructing the Cognitive World: The Next Step*. MIT Press.

Zwaan, R, R. Stanfield, & R. Yaxley (2002). Do language comprehenders routinely represent the shapes of objects? *Psychological Science, 13*, 168-171.