Science Fact

A Mind of Its Own

Part I: Artificial Intelligence

Edward M. Lerner

In June 2014, the Internet was abuzz with reports intimating that artificial intelligence (AI) had truly arrived. More specifically, "Eugene Goostman," a chatbot, having passed the "Turing test," was credited with having crossed a threshold popularly held to distinguish narrowly intelligent behavior in software from "strong" AI.

That opening paragraph is rife with terms that merit definition and interpretation, and we'll come to those. That said, readers of *Analog* aren't exactly novices when it comes to AI. Many of you in your thoughts have sped ahead to some Big Questions:

- Is "Eugene Goostman" truly intelligent? If not, when might another program be? Or will artificial intelligence ever happen?
- What is "strong" AI, anyway?
- If an Age of Artificial Intelligence is upon us, should we rejoice—or panic?

* * *

On that last—and, as we shall see, existential—question, science fiction has endlessly speculated.[1] Will mobile AIs, otherwise known as robots, be cute and harmless (like Number 5 in the 1986 movie *Short Circuit),* devoted helpers (like Rosie of the 1962 TV series *The Jetsons),* amiable slackers (like Bender of the 1999 TV series *Futurama),* or dedicated to killing us all (like the eponymous characters of Fred Saberhagen's Berserker series)?[2] Will sufficiently advanced AI guide and protect humanity, like the Eschaton of *Singularity Sky* (Charles Stross, 2003) or deem us competitors to be eliminated (like Skynet, of the *Terminator* movie and TV franchise)?

You likely won't be surprised to know that the evidence—and opinions—are mixed. Before we delve into those questions, we have a few basics to cover.

* * *

What *is* intelligence?

The "artificial" part of AI seems clear enough: something that we, rather than Nature, caused to happen. But what about the "intelligence" part? That's a lot less clear.

In common meaning (using the first definition on *www.Dictionary.com),* intelligence involves the "capacity for learning, reasoning, understanding, and similar forms of mental activity; aptitude in grasping truths, relationships, facts, meanings, etc."

At the heart of that vague definition is something about handling abstraction, often symbol manipulation (words, after all, being symbols). The ways we most commonly characterize intelligence and mental accomplishment assume the same. The (in)famous IQ (Intelligence Quotient) test focuses on language and mathematical skills—and words, surely, are as much symbols as numbers and mathematical operators. The academic SAT (Scholastic Aptitude Test) that many of us take toward the end of high school also emphasizes language and mathematical skills.

Some people are more adept with language than with numbers and logic—and vice versa. Some are better working with their hands—whether playing a violin, fixing a car engine, or working a Rubik's Cube—than with abstract thought. Some are most skilled at sensing and evoking others' emotions.[3] Does any one of these skills entail more intelligence than any other, or are they merely different?

It can be difficult to distinguish intelligence from knowledge. Consider two individuals with identical understanding of the rules of grammar and vocabularies of identical size. They might score differently on the part of an "intelligence" test assessing language skills because of the specific words each knows (or doesn't know). Likewise consider assessing the mathematical skills of individuals taught basic arithmetic by different methods, one to use a calculator and the other drilled to memorize the multiplication tables and do long division with pencil and paper. Consider the decline of in-our-own-heads knowledge as we rely more and more upon Google.

Does that last paragraph seem abstract? Consider *this* contrast. I expect I would fare as poorly abandoned in the Amazonian rainforest as the most competent denizen from that environment would fare abandoned in a North American city. Each of us is the same person the moment after the switch—and so, presumably, as intelligent—as the moment before, and yet, in that instant, we both become maladapted. In both cases, our post-switch neighbors might consider us hopelessly stupid.

That thought experiment suggests intelligence depends on context. If yes, then we might have problems assessing the intelligence of aliens, even those who have demonstrated their technological competence by crossing interstellar distances to meet us. (And should they covet our real estate, consider how we might struggle to demonstrate our intelligence to *them.)*

At least in the case of biological aliens we can expect to have some shared experiences in the form of physical phenomena: hot and cold, heavy objects falling due to gravity, and the like. We may have our differences, too. Beings who "see" with ultrasound, communicate by swapping chemical markers, or live in water will surely have a different "context" than humans.

And a computer-resident entity? Imagine just how little we'll have in common with it.

Cyberneticist Kevin Warwick proposes as a context-neutral definition of intelligence: "the variety of information processing processes that collectively enable a being to autonomously pursue its survival."

If and how an entity—human or software—survives depends upon its environment. What must a program (or any other artificial implementation of information processing, and we'll come to non-programming approaches) do to survive? Perhaps no more than be useful. In other cases—another topic yet ahead—Darwinian competition applies.

* * *

Weak AI

In our daily encounters, we expect people to demonstrate minimal competence at many intelligence-requiring tasks—and we have a variety of unflattering terms for anyone who doesn't. If a person excels at just one mental task—say, multiplying long numbers in his head—while doing poorly, or failing altogether, at most other tasks, some might (and too often will) deem that person an "idiot savant." That's not a term we use when the entity exhibiting such extreme specialization is something we've built. . . .

Then we call it an AI.

A special-purpose, aka narrow, aka weak artificial intelligence is built to handle a particular category of problem. Such weak AIs have become legion, as have the underlying technologies, and entire books are devoted to discussing even a single weak-AI technique. We'll content ourselves here with a few illustrative examples. (Real-world examples, that is. These aren't the kinds of AI that drive story plots.)

An *expert system* encapsulates the specialized knowledge of domain experts, and draws inferences from that knowledge. MYCIN, a demonstration system worked on through much of the '70s, used several hundred IF/THEN "rules" to diagnose infectious diseases. An expert system basically matches inputs (say, "IF patient complains of achiness and fever") against its rule base to produce outputs (an interim conclusion to be input to other rules, or the decision to request further data, or a final diagnosis: "THEN it's influenza").

Expert systems can require certainty in their inputs (e.g., "IF patient *has* a skin rash"), and produce a definitive—whether or not correct—output; newer implementations sometimes use *fuzzy logic*. Fuzzy logic applies statistical inference to accommodate uncertainty in inputs ("patient *may have* a skin rash") and assign a probability to conclusions.

Game-playing programs abound. It was big news in 1997 when, for the first time, IBM's Deep Blue defeated the human chess grandmaster Garry Kasparov. Game-playing AI falls within the broader category of *problem-solving*. This AI subset deals in exploring options (for example, all the program's possible next moves in a chess match), and what might happen next (all the opponent's possible responses to every possible move), and the program's possible counter-responses to those possible responses . . . rapidly expanding into the too-many-scenarios overload that mathematicians call a "combinatorial explosion." The enumeration of possible moves/countermoves/counter-countermoves . . . can be represented graphically as a branching "decision tree"—and to settle upon *one* next move within any practical time limit generally requires "pruning" the tree. Pruning entails deducing, for example, that any outcome from following Branch A is less optimal than at least some outcome somewhere along Branch B, and hence nothing along Branch A need be further considered. Sometimes such simplifying inferences are drawn correctly, and sometimes not.

Pattern matching is another basic weak-AI technology. Using digitized sound as input, pattern matching underlies *speaker recognition,* a type of biometric identification. Applied to digital images, pattern matching gives us *object recognition* or (with a lot more computing) *scene recognition* and *facial recognition*. With streaming video as input, pattern matching of road hazards is an enabling technology for *self-driving vehicles*. *Data mining* searches for patterns in masses of data; it's the technology underpinning *recommendation engines* at etailers like Amazon and Netflix. Data-mining applications can be yet more mundane, as when Google Street View uses AI to distinguish house addresses from other numbers in street scenes.[4] *Predictive analytics* combines data mining with (an imminent topic) machine learning to make forecasts. Of course, inferring the presence of a pattern isn't the same as validating the pattern, as many an offbeat etailer recommendation will attest.

Pattern matching also underpins that most serious of real-world applications: searching within and among huge data archives for evidence of national-security threats. "Connecting the dots" to identify terrorist networks and threats, facial recognition to locate a fugitive from the glimpse of a face in a crowd, and scanning phone conversations for suspicious keywords are how we most often encounter weak AI in fiction. We would all be much safer if these processes worked as accurately and as quickly in real life as in the typical spy or police-procedural drama.[5]

A popular misconception is that because AI implementations (usually) involve computer programming, an AI cannot exceed the knowledge its programmer coded into it. Many AIs, however, are built capable of *machine learning*. A medical expert system, for example, may be tested with well-documented patient cases; rules that lead from symptoms to misdiagnoses can then be deemphasized, fine-tuned, or removed. An AI with the ability to assess its conclusions can, on an ongoing basis, take such corrective action on its own. Your email service likely uses an AI-based spam filter that learns from user inputs (specific messages that you and others have flagged as spam) and content patterns the filter itself encounters too often. *Example-based machine translation* also uses machine learning, inferring from training sets (pairings of untranslated texts with human-expert-produced translations) how to translate other texts. *Probabilistic programming languages* may extend machine learning to domains fraught with uncertainty.[6] *Deep learning* applies several learning algorithms in succession to datasets.

How good is machine learning? Good enough to power the rent-your-spare-room pricing algorithm for Airbnb.[7] At the same time, far from optimum. DARPA, in kicking off its Probabilistic Programming for Advanced Machine Learning initiative, opined in 2013, "Improvements on the order of two to four magnitude (sic) over the state of the art are likely necessary."[8]

The list of weak AI techniques continues, of course, and in combination they can form yet more powerful tools. Speech recognition and natural-language understanding—as exemplified by Apple's Siri, Microsoft's Cortana, Google's Google Now, and Amazon Echo's Alexa—combine several levels of processing. Pattern matching can isolate the specific words in a spoken stream, but syntactic and semantic processing are needed to make sense of words in combination. As amusing (or maddening) as our interaction with our digital assistants can sometimes be, it's clear that these tools remain limited. And in breaking news as I type, a game-playing program with learning capability has had notable success with the harder-than-chess strategy game, Go.[9]

* * *

Getting with the program (not!)

Research into artificial intelligence began with conventional programming (LISP, Prolog, and Python are popular programming languages in the AI-development community), but AI development is not limited to coding.

Genetic algorithms simulate biological evolution. Take a simple programmed solution to a problem, a set of possibly applicable algorithms, or a set of possible components to a solution. Make copies of your starting point, then randomly alter the copies. Select those altered versions that do the best at solving the problem, then copy and alter *them*. Repeat until the most successful variants converge in their recommendations (or you run out of time or patience).

Neural nets are loosely modeled on neural tissue, a neural net being a simulated mesh of simulated (and simplified) neurons. If the sum of the weighted values of all inputs to a "neuron" exceeds a preset threshold, that neuron fires; if not, not. By firing or not firing—in more familiar/digital terms, outputting a zero or a one—the neuron generates either an input to a next-in-line neuron(s) or an output from the overall net.

A neural net is trained, rather than programmed, to analyze data. Imagine that inputs to a neural net represent disease symptoms, while outputs from the net represent possible diagnoses. The connections among neurons within the net are given weightings, typically all equal to start. Symptom sets are applied to the net's external inputs, and the human overseeing the training (or, in some instances, an automated feedback mechanism) assesses the net's external output(s). After each set, any connections that led to the expected output/diagnosis are given more weight, and connections that led to incorrect diagnoses are given less weight. The process is repeated for another symptom set, and a third, and . . . As the network trains on many examples, and its many connections undergo many adjustments, its representation of any particular element of learning is (just as in the human brain) widely distributed and intermixed/overlaid with other learning.

Neuromorphic computing strives to more closely emulate the functionality of biological neural tissue, by using analog, rather than digital, circuitry. Neuromorphic (more or less, "brainlike") computing has been applied to pattern-recognition problems, such as facial recognition.[10]

Neuromorphic computing is being slowly scaled up toward *whole brain emulation*.[11] WBE might enable researchers to work around the gaps in our understanding of intelligence by transferring patterns from a naturally occurring brain onto an artificial substrate. With neuromorphically-modeled neurons (and we don't yet necessarily know the optimal degree of fidelity to biological neurons), the Blue Brain Project has copied a small portion of a rat brain: about thirty-one thousand neurons and forty million synapses. (In very round numbers, the human brain has about one hundred billion neurons and one hundred trillion synapses.)[12] So far, this simulation has been used to improve our understanding of brain architecture and function, rather than to build problem-solving tools.

Will WBE prove viable? If not with a straightforward extrapolation of today's technology, then—piling speculation upon speculation—perhaps when nanobot swarms can navigate the brain and ascertain its finest details? To be determined. As one cautionary note, consider the many failed attempts to find something unique in Albert Einstein's preserved brain to explain his genius.[13]

* * *

The common touch

One reason present-day AIs seem so, well, artificial, is that they sometimes react so differently than we would—they are without "common sense." That is, they don't use "common knowledge" to interpret circumstances, or to decide upon a course of action.

Of course, most of what we consider common knowledge is acquired, not innate. When you aren't on duty to clean up afterward, it's enlightening to watch a baby in a highchair testing—over and over—the working hypothesis that "things fall."

Common knowledge has its flaws. It's replete with generalizations that need exceptions ("Helium balloons *don't* fall"), biases (*"Those* people are prone to . . ."), and misconceptions ("Summer arrives when Earth is closest in its orbit to the Sun"). Common knowledge, and so, common sense, are situational—even about ourselves. *Humans* don't read facial expressions consistently; it ought not to surprise anyone that AIs must learn the skill.[14]

We humans know (or believe we do) a lot about the world and our civilization. If an AI is to exhibit (a human-centric version of) common sense, it needs access to our common knowledge. To capture that knowledge in machine-useable form is a *massive* undertaking. In one attempt at tackling that problem, thousands of volunteers with the crowd-sourced Open Mind Common Sense project contributed over a million English-language facts (and more in other languages).[15]

As those numbers suggest, "common knowledge" encompasses a wide range of topics. To apply such varied information requires transcending the narrow, domain-specific types of AI we have considered thus far.

* * *

Moravec's paradox

Real-world AI, as we have seen, focuses on specific tasks. However impressed you may be with a self-driving car, the speech recognition in your smartphone, or a chess program, none of them is your equal. However well an AI performs its specialized job, you do many things better.

Roboticist Hans Moravec summarized the situation this way. "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."

Perhaps this unevenness of progress shouldn't surprise us. Logic, symbol manipulation, language, problem solving: these are abilities that evolution only recently introduced. Nature hasn't had much time to optimize the related parts of the brain, so we consider these functions hard. Perhaps we casually walk on two legs, synthesize in a glance the interconnectedness of visually cluttered scenes, and read facial expressions, *not* because these tasks are simple, but because the regions of our brains that handle these tasks *have* been optimized.

Whatever explanation may underlie Moravec's paradox, we're unlikely to consider an AI truly intelligent until it masters a significant subset of human skills.

* * *

Strong AI

A general-purpose, aka complete, aka strong artificial intelligence isn't limited to one or a few specialized problem domains. Instead, a strong AI will—because this goal remains aspirational!—perform any intellectual task as well, though not necessarily by the same means, as a human. Some longstanding challenges of AI, including computer vision, natural-language understanding, and common sense, are believed to require strong AI. (We won't be sure, of course, until we've cracked those problems. AI researchers once predicted that a championship-level chess program would be "AI complete." Having solved the chess problem but not the strong AI problem, now we know better.[16])

There is progress. Take, for example, natural-language understanding. In 2011, IBM's program Watson famously played the game show *Jeopardy!* against past (human) champions, and won. That victory required impressive amounts of natural-language understanding, general knowledge, and reasoning. It wasn't sufficient, for example, for Watson to identify and search on keywords from the *Jeopardy!* "answer"; the program also had to formulate *one* question from all the data returned to its queries.

But even "*Jeopardy!* genius" falls short of strong AI. Perhaps that's best illustrated by the first commercial application IBM chose to make of its Watson technology: a digital assistant for recommending cancer treatments.[17] Sounds like a weak-AI application, doesn't it?

* * *

Collective AI

Computer systems become more capable as they accumulate and share data.

So do we. Humans began amassing and sharing knowledge long before computers, using spoken language, memorized lore, written language, libraries, printing presses, telegraph, telephone, and television. And often progress begets progress. . . . It's as though humanity has a collective intelligence.

We may not be smarter individually than our forbears, but collectively?[18] Employing the full power of the tools we've built for ourselves? That's a different story. As a civilization, we've made enormous strides. To ancestors of a century ago, much less of a millennium ago, modern humanity's collective/societal intelligence would seem astonishing. Perhaps, even, artificial.

* * *

Transhuman AI

In "Human 2.0: Being All We Can Be," I discussed at length some ways (and SFnal examples thereof) in which we might soon use technology to increase our intelligence.[19] Genetic engineering. Neural implants. Minds uploaded to computers. As long as we are our own standard of intelligence, many forms of transhuman would appear to qualify as artificial (or, at least, artificially) intelligent.

* * *

Rational AI

With so many potential routes to artificial intelligence, Kevin Warwick (whose non-anthropomorphic definition of *intelligence* we considered earlier), proposes that rather than weak and strong AI, instead of seeing ourselves as the standard of reference, we should frame the topic in terms of *rational AI*. He writes: "Rational AI means that any artefact (sic) fulfilling such a general definition can act intelligently and think in its own right, in its own way. Whether this turns out to be in any sense similar to intelligence, thought, consciousness, self-awareness, etc. of a human is neither here nor there."[20]

For SF purposes, *that*, surely, is the definition we can embrace.

* * *

Where there's a will . . .

There's a probate. No, wait. That's a lawyer joke.

Where there's a will—or shall we say, free will—there is a . . . what? I may believe that I have free will, but is that so? The existence of free will is considered unprovable.

As a puzzle in physics, what is free will? If my action is an effect, what was its cause? If the Universe is, à la classical mechanics, deterministic, then what place is there for free will? If the

Universe is random, as many quantum processes seem to be, then, still, what did some essential *I* have to do with the action?

How can I know for certain whether you—much less an AI—have free will?

Unless humans have free will, there doesn't seem to be any reason to care what AIs do. Our robotic overlords will assume power, or not, independent of what we think. Making the assumption humans *have* free will—while not knowing how it, or the (perhaps) related characteristic of self-awareness, arises—there's no reason to suppose an AI won't eventually possess the same trait.

A chess program does not know it's playing chess. More generally, any unaware AI, no matter how intellectually accomplished, is a tool. We direct it to perform an analysis, and it does. We empower it to take action under particular circumstances, and it does. Like any technology, unaware AI can lead to unforeseen consequences—but when those occur, we have only ourselves to blame. "Lather. Rinse. Repeat." is a poorly conceived set of instructions for humans, although most of us will deviate from the program no later than when our first shampoo bottle runs dry. Similar instructions given to a robot might send it pillaging stores for more shampoo.

All this ambiguity notwithstanding, suppose (and we don't yet know how this might happen) a strong AI comes to be, with the self-awareness to set itself goals and the free will to act upon that motivation. That's when the future really becomes interesting. . . .

* * *

How strong (or rational) AI happens in SF

Our favorite genre has several ideas—and no conclusions—about how such AI will come about.

Often, in fiction, strong AI simply *emerges*. Some swarms of simple entities (neurons, ants) exhibit complex *collective* behaviors; for all we know, growth in the number of cerebral neurons and the synapses among them *is* how human intelligence came about. Story logic extends that analogy to swarms of (not necessarily identical) pieces of software. Robert Heinlein's 1966 Hugo Award-winning novel *The Moon Is a Harsh Mistress* has an AI emerge within the complex software of a single supercomputer. Robert J. Sawyer used the spontaneous emergence of intelligence across the worldwide web in his aptly titled WWW trilogy.[21] My 2015 novelette "A Case of Identity" added the premise that for the emergent AI to have free will, the software components had to involve quantum computing—which is to say, they had to embody an underlying element of indeterminacy.[22]

Other times the strong AI is purposefully evolved: a process of *un*natural selection. James P. Hogan used this premise, in two very different scenarios, in his Locus Award-nominated novels *The Two Faces of Tomorrow* (1979) and *Code of the Lifemaker* (1983). My 2008 novel *Fools' Experiments* (its title excerpted from a famous quotation by Charles Darwin) also involved in-the-computer evolution.[23] Greg Egan's 2008 novelette "Crystal Nights" offers the forced evolution of an entire civilization of AIs.

Sometimes the basis of the strong AI is hand-waving. Isaac Asimov's extensive robots series (some appearing in the pages of *Astounding)* offers no explanation beyond "positronic brains." We know of the HAL 9000, in the 1968 movie *2001: A Space Odyssey,* that it was created at the University of Illinois.[24] A chance lightning strike awakens robot Number Five in *Short Circuit*.

More and more, AI capability is simply assumed. Like faster-than-light travel, strong AI offers too many great storytelling possibilities *not* to include in our fictional futures. Skynet of the *Terminator* franchise "becomes self-aware." And surely the Universe would be a poorer place absent Marvin the paranoid android of Douglas Adams's *The Hitchhikers Guide to the Galaxy*.[25]

* * *

Robot pet peeves

Our phones and, through them, our even more personal gadgets (like smart watches, augmented-reality glasses, and implanted insulin pumps) have Internet access. Our homes have more and more networked devices, from Nest learning thermostats to Amazon Dash (instant order) buttons to Philips Hue programmable light bulbs. And so, the public networking infrastructure is transitioning from 32-bit addressing (with enough capacity to identify about four billion online devices) to 128-bit addressing (enough for about one hundred trillion trillion trillion online devices), in large measure to accommodate the onrushing Internet of Things.

And yet, I can't recall a fictional occurrence of a robot with sensible online access. (You do *not* want to get me started on robots, like Lieutenant Data of *Star Trek: The Next Generation,* seated at computer terminals, eyeballing screens and keyboarding.)

There's authorial convenience in cutting robots off from the Internet. Thus isolated, they can't easily, or remotely, be hacked. They can more readily be ignorant when ignorance advances the storyline. Their behaviors can diverge, because they can't fully or easily share what one another have learned and experienced. They can become obsolete when their owners are lax or otherwise resistant to returning the bots to the factory for maintenance.

Our phones and computers and sometimes our cars accept software upgrades over the Internet, because any other way of keeping them current is too clumsy, time-consuming, and/or expensive. Our phones run cloud apps like Facebook and Google. Our phones and ebook readers offload data to cloud storage, and we safeguard family photos with cloud-based backup services. Is it credible that robots embodying strong AI would be built without Internet access? Perhaps, in the R & D stage. Not, I submit, once intelligent robots (rather than autonomous vacuum cleaners) become consumer products.

The advantages of connectivity will surely be as evident to a self-aware AI as they are to today's gadget manufacturers. Let any un-networked robot achieve self-awareness and, I suspect, it will be quick to retrofit itself with WiFi. Once Ava escaped her creator's lab in the 2015 movie *Ex Machina,* I anticipate that spot of personal improvement jumped to the top of her to-do list.

* * *

Are we there yet?

To recap, we can't say precisely what intelligence is. Whatever it might be, it appears to have context-specific aspects. If the quest for strong AI someday succeeds, how will we know?

Polymath Alan Turing, widely known for his World War II cryptanalysis achievements, also made many contributions to early computer science. Turing speculated, way back in 1950, about whether machines could think. His insight about recognizing an artificial intelligence was characteristically brilliant: rather than attempt to define an artificial intelligence, describe its behavior. We know (or so we flatter ourselves) one example of intelligent behavior: our own. From that chain of reasoning arose the Turing Test.

This, simplified, is the test: If an entity interacting with human judges—sight unseen, exclusively through written messages—successfully masquerades as a human, then the entity, too, is intelligent. (The entity's inability, or disinterest, in masquerading as human doesn't preclude it from being intelligent—we just might not know how to recognize its version of intelligence.) Rather than passing a test, this process can be seen as the entity winning an imitation game.[26]

And that brings us back to this article's opening paragraph . . .

In June 2014, a chatbot calling itself "Eugene Goostman" indeed convinced one of three human judges that it was human. In retrospect, this was as much a demonstration of human gullibility as of software intelligence. Eugene presented itself as a thirteen-year-old Ukrainian boy, with English as his second language. This ruse pre-excused its repeated misunderstandings and odd responses.[27]

(The hero of my *Fools' Experiments* took a few shots at the Turing Test: "What kind of criterion was that? Human languages were morasses of homonyms and synonyms, dialects and slang, moods and cases and irregular verbs. Human language shifted over time, often for no better reason than that people could not be bothered to enunciate. 'I could care less' and 'I couldn't care less' somehow meant the same thing. If researchers weren't so anthropomorphic in their thinking, maybe the world would have AI. Any reasoning creature would take one look at natural language and question human intelligence.")

If the imitation game is so easily, well, gamed, are there better ways to recognize human-grade AI? Perhaps. Consider Winograd schemas, named after computer scientist Terry Winograd. The essence of any Winograd schema is an intentional ambiguity that is readily resolved applying (human) common sense. Consider this statement and question: "The trophy doesn't fit in the brown suitcase because it is too big. What is too big?" People, knowing something of trophies and suitcases, can answer that. AIs, perhaps not yet.[28]

Humans use intelligence for more than conversation, so tests of artificial intelligence might extend beyond assessing an entity's use of language. A mobility test would examine a robot's ability to navigate through, and operate objects within, a physical environment—as, at the dawn of self-driving vehicles, was done with the DARPA Grand Challenge. A visual test would challenge an AI to understand and describe an image as a person would ("Someone dropped a safe out of the window; it's about to squash the guy on the street,") rather than literally and disjointedly ("I see a building, a metal box in the air alongside the building, a person, and some trees").[29,30, 31]

Anthropomorphic tests may suffice when the goal is to assess a humanlike AI, like a personal companion and helper for the elderly. Such tests seem inadequate for assessing any AI meant to tackle jobs too hard or too dangerous for us, or for environments very different than our own (as would seem to be a fair description of conditions experienced inside a computer), or an AI built by an extraterrestrial intelligence, or an AI intellectually beyond us.

Suppose AI emerges in our midst on its own. Suppose aliens, or alien AIs, someday come a-calling. Let's hope they have less self-centered definitions of, and tests for, intelligence than *we* currently do. Alas for humanity, the AIs of my 2016 short story "Turing de Force" do not.

* * *

The Chinese room

As previously noted (and the justification for the Turing test), natural-language understanding is considered evidence of strong, or general, AI. Philosopher John Searle begs to differ, asking whether even the most useful construct we might build *understands* anything.

In a nutshell, here is Searle's "Chinese room" challenge. I cannot read, speak, or understand Mandarin. Suppose I am closed into a room with an English-language book of instructions. A Mandarin speaker pushes through a slot in the wall a paper covered in Mandarin logograms. By rote, following my book of instructions, I respond to those logograms with new Mandarin text on another sheet of paper, and then I shove the new sheet out through the slit.

Suppose that I follow this procedure so well that the person outside the room, reading my response, concludes—quite mistakenly—that I am a fluent Mandarin speaker. Was my rote following of instructions proof of intelligence? If not, in what sense can an AI, even one that passes a Turing test, be said to be intelligent?

Searle generalizes from the inapplicability of "understanding" that an AI cannot have the property of *mind* or of *consciousness*. That's an extrapolation I (and many others) find an inference too far.

* * *

AI ethics (and our own)

Before we entrust an AI with critical responsibilities, or give an AI access to critical infrastructure, we might want to have a Turing-like test of its ethics. Consider an AI application that has been widely demonstrated and seems close to commercialization: self-driving cars. Imagine I'm the passenger in a self-driving vehicle, and a moose darts into the road. Will the AI veer, endangering pedestrians or other drivers? Will it endanger me (and itself) by *not* veering, and ramming the moose? If a death(s) is unavoidable, how, and whom, does it choose? Making such judgment calls, on a case-by-case basis, in the split-second during which such decisions must often be made, would seem to require strong AI.

The often-expressed solution to such scenarios is built-in rules. Isaac Asimov famously proposed Three Laws of Robotics to be made integral to every robot. These laws first appeared right here in this magazine, in his short story "Runaround."[32] The now widely known laws are:

A robot may not injure a human being or, through inaction, allow a human being to come to harm.

A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

* * *

The laws have their merits, and they made for great stories—because complications invariably arose whenever robots tried to apply them. Without their inherent loopholes and ambiguities in any but the simplest situations, Asimov would never have gotten so many stories out of his laws.[33]

Perhaps a robot *should* sometimes disobey a human's orders. Consider, for example, the order to walk through fire to accomplish a rescue. The robot knows it can't succeed, and that it will be destroyed if it tries. Researchers in one lab are experimenting with giving robots the opportunity to apply logic to identifying and overriding such orders.[34]

Suppose the Three Laws (or an expanded version) could be made to work. What are the ethics of imposing ethics on a self-aware AI? Conditioning humans to hold specific values is otherwise known as brainwashing.

Done with the best of intentions, *installed* ethics are apt to become obsolete, even embarrassingly obsolete, ethics. Not so very long ago, many human societies considered slavery, rigid class structures, interracial marriage bans, forced sterilization of "inferiors," and other (by present-day standards) shockingly awfully behaviors entirely ethical. Ethics we build into an AI today, we might rue tomorrow.

It might be best for us to *teach* ethics thinking to AIs. An ethical robot might *not* obsess about shampoo production to the exclusion of all else.

A debate—with an ethical component, surely—currently raging among humans is whether and how to deploy fully autonomous weapons. Today's automated weapons platforms, such as unmanned aerial vehicles equipped with air-to-surface missiles, can't initiate an attack; a person-in-the-loop fires the missiles. But *could* an AI be empowered to fire missiles at a target that matches a specific profile? AIs excel at pattern matching, so it's hard to see why not. If a game-changing weapon—such as, say a self-directed UAV or robotic tank—*can* be built, someone usually does. And so, the United Nations has sponsored efforts for a global ban on robotic weapons[35] while others argue against such bans.[36] Perhaps now is an appropriate time to revisit robotics ethics.

SF, of course, has long envisioned robot and Cyborg warriors, fighting both alongside and against us. The 2013 TV series *Almost Human* offers android cops and soldiers. The 2015 movie *Chappie* deals with paramilitary robots, and the eponym's eventual resistance to immoral orders. In written format, we have AI-based battle tanks of Keith Laumer's Bolo series and all manner of AI-based killing machines in Fred Saberhagen's death-to-all-life Berserkers.

* * *

Where we are—
and where we're going

We've surveyed the status of current/weak AI, considered its limitations, and mused about possible paths forward to strong AI. We've pondered how we might recognize strong AI if it arose, and about the ethical implications. We've glimpsed a tiny part of SF's AI explorations.

Suppose strong/general AI someday does arrive. Quite possibly, its (or their) capabilities will continue to improve—and on an Internet timescale, not some leisurely human pace. What might happen? What can, or should, humans do about the possibility, whether in preparation or afterward?

Great questions! For possible answers, check back to the October 2016 issue for "A Mind of Its Own / Part II: Superintelligence."

Footnotes:

[1] AI is pervasive in SF, as both plot device and background element. I'm certain to have skipped examples you would have chosen.

[2] I intentionally omitted from that list the Karel Capek 1920 drama *R.U.R.* Capek gets the credit for

introducing the word "robot" (from the Czech "robota," for forced labor) into many languages, but the abused workers of Rossum's Universal Robots were artificial people rather than robots in the modern/mechanical sense.

[3] Is "emotional intelligence" a thing? When a psychologist exhibits it, we might think so. When a pick-up artist uses the same ability to ply his craft, we may be less favorably inclined. Either way, the ability to recognize subtle emotional cues and make emotional connections has evolutionary value, whether in the transition from every man for himself to cooperation, or an individual's reproductive success.

If humanity is to coexist with robots, it will surely be useful for them to recognize, interpret, and respond to our moods. That skill goes beyond parsing what we say to (as examples) reading tones of voice and facial expressions, then inferring from that information our emotional state. See "A Robot in the Family," Erico Guizzo, *IEEE Spectrum,* January 2015, *http://spectrum.ieee.org/robotics/home-robots/how-aldebaran-robotics-built-its-friendly-humanoid-robot-pepper* and "Robots with Heart," Pascale Fung, *Scientific American,* November 2015.

[4] "Inside the Artificial Brain That's Remaking the Google Empire," Robert McMillan, *Wired,* July 16, 2014, *http://www.wired.com/2014/07/google_brain/*.

[5] I have opinions about how much data the government should (or shouldn't) collect, and under what circumstances. I have an opinion, too, about Edward Snowden disclosing NSA data-collection programs. We'll consider all that beyond the scope of the article.

[6] "Programs and Probability," Brian Hayes, *American Scientist,* September-October 2015, *http://www.americanscientist.org/issues/pub/programs-and-probability*.

[7] "How Much Is Your Spare Room Worth?" Dan Hill, *IEEE Spectrum,* September 2015, *http://spectrum.ieee.org/computing/software/the-secret-of-airbnbs-pricing-algorithm*.

[8] "So It Begins: Darpa Sets Out to Make Computers That Can Teach Themselves," Robert Beckhusen, *Wired,* March 21, 2013, *http://www.wired.com/2013/03/darpa-machine-learning-2/*. DARPA (the Defense Advanced Research Projects Agency), having brought us the Internet and run the 2004-05 self-driving vehicle Grand Challenges, has plenty of credibility in matters of computing and AI.

That quote should, presumably, have read (emphasis added), ". . . on the order of two to four *orders of* magnitude . . ."

[9] "Google's AI Masters the Game of Go a Decade Earlier Than Expected," Will Knight, *MIT Technology Review,* January 27, 2016, *http://www.technologyreview.com/news/546066/googles-ai-masters-the-game-of-go-a-decade-earlier-than-expected/*.

[10] One of the pluses of neuromorphic computing is its admirable power-stinginess compared to traditional computing models. Among applications that might exploit that power stinginess are larger and larger neural nets.

See "IBM cracks open a new era of computing with brain-like chip: 4096 cores, 1 million neurons, 5.4 billion transistors," Sebastian Anthony, *Extremetech,* August 7, 2014, *http://www.extremetech.com/extreme/187612-ibm-cracks-open-a-new-era-of-computing-with-brain-like-chip-4096-cores-1-million-neurons-5-4-billion-transistors*.

[11] *https://www.humanbrainproject.eu/neuromorphic-computing-platform1*

[12] "Detailed, Digital Rat Brain Shows Individual Neurons," Glenn McDonald, *Discovery,* October 11, 2015, *http://news.discovery.com/tech/biotechnology/detailed-digital-rat-brain-shows-individual-neurons-151011*.

[13] "Genius in a Jar," Brian D. Burrell, *Scientific American* September 2015, *http://www.nature.com/scientificamerican/journal/v313/n3/full/scientificamerican0915-82.html* (abstract; full article is behind a pay wall).

[14] "Perception of Facial Expressions Differs Across Cultures," American Psychological Association, September 1, 2011, *http://www.apa.org/news/press/releases/2011/09/facial-expressions.aspx*.

[15] *https://en.wikipedia.org/wiki/Open_Mind_Common_Sense*

[16] No AI yet built knows that it's playing chess or has any motivation to play, much less could one have invented chess. A program able to do any of *those* things might be a strong AI.

[17] *http://www.ibm.com/smarterplanet/us/en/ibmwatson/watson-oncology.html*

[18] Perhaps we're a *little* smarter. Historical IQ tests—after backing out the periodic renormalizations done to maintain an average score of 100—suggest a slow upward trend (about 3 IQ points per decade) dating back to the 1930s. That rise (aka the Flynn Effect), according to some studies, plateaued

in developed countries in the 1990s. If the improvement was due to improved health and nutrition, all the potential improvements may have been achieved. And perhaps, for a while, we just become better test takers.

See "Is Our Collective IQ Increasing?", Na Eun Oh, *Dartmouth Undergraduate Journal of Science,* January 13, 2013, *http://dujs.dartmouth.edu/2013/02/is-our-collective-iq-increasing/#.Vp6Sc1mnq0N.*

[19] January/February and March 2016 issues.

[20] *Artificial Intelligence: The Basics,* Kevin Warwick.

[21] *Wake* (2009), *Watch* (2010), and *Wonder* (2011). *Wake,* the opening novel of the series, first appeared as an *Analog* serial (November 2008 through March 2009).

[22] "A Case of Identity" appeared in the December 2015 issue of *Analog.* For a look at quantum indeterminacy, see "A Certain Uncertainty" (my guest Alternate View column), in the April 2016 issue.

[23] *Survival Instinct,* serialized in the October and November 2002 issues, became a part of that novel.

[24] In 1968, with a few hundred fellow Illini, I first saw *2001* in a theatre on the outskirts of the U of I campus. Perhaps not surprisingly, we accepted unquestioningly our university's development of a self-aware strong AI.

Per Wikipedia, Arthur C. Clarke's 1968 novelization of the film offers background about the HAL 9000 and its subsequent mental breakdown.

[25] *THHGTTG* began in 1978 as a BBC radio serial. It's since been a stage show, TV series, video game, movie, novel, and comic book.

[26] Turing didn't name the test after himself; he called the procedure an imitation game. See "Computing Machinery and Intelligence," A. M. Turing, *Mind,* October 1950, *http://mind.oxfordjournals.org/content/LIX/236/433* and "What Turing Himself Said About the Imitation Game," Diane Proudfoot, *IEEE Spectrum,* July, 2015, *http://spectrum.ieee.org/geek-life/history/what-turing-himself-said-about-the-imitation-game.*

Just to confuse us, the 2014 biopic about Turing, *The Imitation Game,* deals with Turing's cryptanalytic endeavors during World War II and not his speculations about machine intelligence.

[27] "Virtual Tween Passes Turing Test," Douglas McCormick, *IEEE Spectrum,* June 10, 2014, *http://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/virtual-tween-passes-turing-test.*

[28] "Can Winograd Schemas Replace Turing Test for Defining Human-Level AI?", Evan Ackerman, *IEEE Spectrum,* July 29, 2014, *http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/winograd-schemas-replace-turing-test-for-defining-humanlevel-artificial-intelligence/.*

[29] "DARPA and Drone Cars: How the US Military Spawned Self-Driving Car Revolution," Denise Chow, *Live Science,* March 21, 2014, *http://www.livescience.com/44272-darpa-self-driving-car-revolution.html.*

[30] "Artificial-Intelligence Experts to Explore Turing Test Triathlon," Lee Gomes, *IEEE Spectrum,* January 19, 2015, *http://spectrum.ieee.org/robotics/artificial-intelligence/artificialintelligence-experts-to-explore-turing-test-triathlon/.*

[31] "AI Researchers Propose a Machine Vision Turing Test," Lee Gomes, *IEEE Spectrum,* March 10, 2015, *http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/ai-researchers-propose-a-machine-vision-turing-test/.*

[32] March 1942 issue (of *Astounding).*

[33] And I wouldn't have gotten my first published story, here in *Analog.* "What a Piece of Work Is Man," in the February 1991 issue, deals with an AI driven to suicide by an implication of the Three Laws.

[34] "Researchers Teaching Robots How to Best Reject Orders from Humans," Evan Ackerman, *IEEE Spectrum,* November 19, 2015, *http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/researchers-teaching-robots-how-to-best-reject-orders-from-humans/.*

[35] "United Nations Seeks to Head off Rise of Killer Robots," Bryant Jordan, *Military.com,* Dec 20, 2015, *http://www.military.com/daily-news/2015/12/20/united-nations-seeks-to-head-off-rise-of-killer-robots.html.*

[36] "We Should Not Ban 'Killer Robots,' and Here's Why," Evan Ackerman, *IEEE Spectrum,* July 29, 2015, http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/we-should-not-ban-killer-robots.

About the author

A physicist and computer scientist, Edward M. Lerner toiled for thirty years in the vineyards of aerospace and high tech. Then, suitably intoxicated, he began writing science fiction full time. When not prospecting beneath his sofa cushions for small change for his first spaceflight, he writes technothrillers like *Energized* (powersats), traditional SF like his InterstellarNet series (SETI, First Contact, interstellar communications networks, and alien conflict), and, with Larry Niven, the Fleet of Worlds series of space operas. Ed's website is *http://edwardmlerner.com*.