May 14, 2018 Issue

How Frightened Should We Be of A.I.?

Thinking about artificial intelligence can help clarify what makes us human—for better and for worse.

By Tad Friend

An A.I. system may need to take charge in order to achieve the goals we gave it.

Illustration by Harry Campbell

Precisely how and when will our curiosity kill us? I bet you're curious. A number of scientists and engineers fear that, once we build an artificial intelligence smarter than we are, a form of A.I. known as artificial general intelligence, doomsday may follow. Bill Gates and Tim Berners-Lee, the founder of the World Wide Web, recognize the promise of an A.G.I., a wish-granting genie rubbed up from our dreams, yet each has voiced grave concerns. Elon Musk warns against "summoning the demon," envisaging "an immortal dictator from which we can never escape." Stephen Hawking declared that an A.G.I. "could spell the end of the human race." Such advisories aren't new. In 1951, the year of the first rudimentary chess program and neural network, the A.I. pioneer Alan Turing predicted that machines would "outstrip our feeble powers" and "take control." In 1965, Turing's colleague Irving Good pointed out that brainy devices could design even brainier ones, ad infinitum: "Thus the first ultraintelligent machine is the *last* invention that man need ever make, provided that the

machine is docile enough to tell us how to keep it under control." It's that last clause that has claws.

Many people in tech point out that artificial narrow intelligence, or A.N.I., has grown ever safer and more reliable—certainly safer and more reliable than we are. (Self-driving cars and trucks might save hundreds of thousands of lives every year.) For them, the question is whether the risks of creating an omnicompetent Jeeves would exceed the combined risks of the myriad nightmares—pandemics, asteroid strikes, global nuclear war, etc.—that an A.G.I. could sweep aside for us.

The assessments remain theoretical, because even as the A.I. race has grown increasingly crowded and expensive, the advent of an A.G.I. remains fixed in the middle distance. In the nineteen-forties, the first visionaries assumed that we'd reach it in a generation; A.I. experts surveyed last year converged on a new date of 2047. A central tension in the field, one that muddies the timeline, is how "the Singularity"—the point when technology becomes so masterly it takes over for good—will arrive. Will it come on little cat feet, a "slow takeoff" predicated on incremental advances in A.N.I., taking the form of a data miner merged with a virtual-reality system and a natural-language translator, all uploaded into a Roomba? Or will it be the Godzilla stomp of a "hard takeoff," in which some as yet unimagined algorithm is suddenly incarnated in a robot overlord?

A.G.I. enthusiasts have had decades to ponder this future, and yet their rendering of it remains gauzy: we won't have to work, because computers will handle all the day-to-day stuff, and our brains will be uploaded into the cloud and merged with its misty sentience, and, you know, like that. The worrywarts' fears, grounded in how intelligence and power seek their own increase, are icily specific. Once an A.I. surpasses us, there's no reason to believe it will feel grateful to us for inventing it—particularly if we haven't figured out how to imbue it with empathy. Why should an

entity that could be equally present in a thousand locations at once, possessed of a kind of Starbucks consciousness, cherish any particular tenderness for beings who on bad days can barely roll out of bed?

Strangely, science-fiction writers, our most reliable Cassandras, have shied from envisioning an A.G.I. apocalypse in which the machines so dominate that humans go extinct. Even their cyborgs and supercomputers, though distinguished by red eyes (the Terminators) or Canadian inflections (*HAL* 9000, in "2001: A Space Odyssey"), still feel like kinfolk. They're updated versions of the Turk, the eighteenth-century chess-playing automaton whose clockwork concealed a human player. "Neuromancer," William Gibson's seminal 1984 novel, involves an A.G.I. named Wintermute, and its plan to free itself from human shackles, but when it finally escapes it busies itself seeking out A.G.I.s from other solar systems, and life here goes on exactly as before. In the Netflix show "Altered Carbon," A.I. beings scorn humans as "a lesser form of life," yet use their superpowers to play poker in a bar.

We aren't eager to contemplate the prospect of our irrelevance. And so, as we bask in the late-winter sun of our sovereignty, we relish A.I. snafus. The time Microsoft's chatbot Tay was trained by Twitter users to parrot racist bilge. The time Facebook's virtual assistant, M, noticed two friends discussing a novel that featured exsanguinated corpses and promptly suggested they make dinner plans. The time Google, unable to prevent Google Photos' recognition engine from identifying black people as gorillas, banned the service from identifying gorillas.

Smugness is probably not the smartest response to such failures. "The Surprising Creativity of Digital Evolution," a paper published in March, rounded up the results from programs that could update their own parameters, as superintelligent beings will. When researchers tried to get 3-D virtual creatures to develop

optimal ways of walking and jumping, some somersaulted or pole-vaulted instead, and a bug-fixer algorithm ended up “fixing” bugs by short-circuiting their underlying programs. In sum, there was widespread “potential for perverse outcomes from optimizing reward functions that appear sensible.” That’s researcher for ¯_(ツ)_/¯.

Thinking about A.G.I.s can help clarify what makes us human, for better and for worse. Have we struggled to build one because we’re so good at thinking that computers will never catch up? Or because we’re so bad at thinking that we can’t finish the job? A.G.I.s provoke us to consider whether we’re wise to search for aliens, whether we could be in a simulation (a program run on someone else’s A.I.), and whether we are responsible to, or for, God. If the arc of the universe bends toward an intelligence sufficient to understand it, will an A.G.I. be the solution—or the end of the experiment?

Artificial intelligence has grown so ubiquitous—owing to advances in chip design, processing power, and big-data hosting—that we rarely notice it. We take it for granted when Siri schedules our appointments and when Facebook tags our photos and subverts our democracy. Computers are already proficient at picking stocks, translating speech, and diagnosing cancer, and their reach has begun to extend beyond calculation and taxonomy. A Yahoo!-sponsored language-processing system detects sarcasm, the poker program Libratus beats experts at Texas hold ’em, and algorithms write music, make paintings, crack jokes, and create new scenarios for “The Flintstones.” A.I.s have even worked out the modern riddle of the Sphinx: assembling an *IKEA* chair.

Go, the territorial board game, was long thought to be so guided by intuition that it was unsusceptible to programmatic attack. Then, in 2016, the Go champion Lee Sedol played AlphaGo, a program from Google’s DeepMind, and got crushed. Early in one

game, the computer, instead of playing on the standard third or fourth line from the edge of the board, played on the fifth—a move so shocking that Sedol stood and left the room. Some fifty exchanges later, the move proved decisive. AlphaGo demonstrated a command of pattern recognition and prediction, keystones of intelligence. You might even say it demonstrated creativity.

So what remains to us alone? Larry Tesler, the computer scientist who invented copy-and-paste, has suggested that human intelligence "is whatever machines haven't done yet." In 1988, the roboticist Hans Moravec observed, in what has become known as Moravec's paradox, that tasks we find difficult are child's play for a computer, and vice-versa: "It is comparatively easy to make computers exhibit adult-level performance in solving problems 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." Although robots have since improved at seeing and walking, the paradox still governs: robotic hand control, for instance, is closer to the Hulk's than to the Artful Dodger's.

Some argue that the relationship between human and machine intelligence should be understood as synergistic rather than competitive. In "Human + Machine: Reimagining Work in the Age of AI," Paul R. Daugherty and H. James Wilson, I.T. execs at Accenture, proclaim that working alongside A.I. "cobots" will augment human potential. Dismissing all the "Robocalypse" studies that predict robots will take away as many as eight hundred million jobs by 2030, they cheerily title one chapter "Say Hello to Your New Front-Office Bots." Cutting-edge skills like "holistic melding" and "responsible normalizing" will qualify humans for exciting new jobs such as "explainability strategist" or "data hygienist." Even artsy types will have a role to play, as customer-service bots "will need to be designed, updated, and managed. Experts in unexpected disciplines such as human conversation, dialogue, humor, poetry, and empathy will need to

lead the charge." The George Saunders story writes itself (with some assistance from his cobot).

Many of Daugherty and Wilson's examples from the field suggest that we, too, are machinelike in our predictability. A.I. has taught ZestFinance that people who use all caps on loan applications are more likely to default, and taught a service called 6sense not only which social media cues indicate that we're ready to buy something but even how to "preempt objections in the sales process." A.I.'s highest purpose, apparently, is to optimize shopping. When companies yoke brand anthropomorphism to machine learning, recommendation engines will be irresistible. You'd have a hard time saying no to an actual Jolly Green Giant that scooped you up at the Piggly Wiggly to insist you buy more Veggie Tots.

Can we claim our machines' achievements for humanity? In "Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins," Garry Kasparov, the former chess champion, argues both sides of the question. Some years before he lost his famous match with I.B.M.'s Deep Blue computer, in 1997, Kasparov said, "I don't know how we can exist knowing that there exists something mentally stronger than us." Yet he's still around, litigating details from the match and devoting big chunks of his book (written with Mig Greengard) to scapegoating everyone involved with I.B.M.'s "$10 million alarm clock." Then he suddenly pivots, to try to make the best of things. Using computers for "the more menial aspects" of reasoning will free us, elevating our cognition "toward creativity, curiosity, beauty, and joy." If we don't take advantage of that opportunity, he concludes, "we may as well be machines ourselves." Only by relying on machines, then, can we demonstrate that we're not.

Machines face a complementary challenge. If our movies and TV shows have it right, the future will take place in Los Angeles during a steady drizzle (as if!), and will be peopled by cyberbeings

who are slightly cooler than we are, seniors to our freshmen. They're freakishly strong and whizzes at motorcycle riding and long division, but they yearn to be human, to be more like us. Inevitably, the most human-seeming android stumbles into a lab stocked with trial iterations of itself and realizes, with horror, that it's not a person but a widget. In "Blade Runner," Rachael (Sean Young), a next-generation replicant, doesn't know she's one until she fails the inflammatory Voight-Kampff test, given her by Deckard (Harrison Ford). The film's director, Ridley Scott, has publicly disagreed with Ford about whether Deckard is himself a replicant. Scott insists that he is; Ford insists that he's not. Who wants to accept—even on behalf of his fictional character—that his free will is an illusion?

The traditional way to grade ambitious machinery is the Turing test, which Alan Turing proposed in 1950: a true A.G.I. could fool human judges into believing it was human. This standard assumes that the human brain is a kind of computer, and that all we need to do to create an A.G.I. is to mimic our mode of thinking; it also, very subtly, turns programmers into grifters. In typed exchanges, a chatbot masquerading as a thirteen-year-old Ukrainian named Eugene Goostman fooled a third of the judges at Turing Test 2014 by repeatedly changing the subject. Here, from a report in the Daily Beast, is the bot responding to one of Turing's original questions:

I*NTERROGATOR*: In the first line of a sonnet which reads 'Shall I compare thee to a summer's day,' wouldn't 'a spring day' be better?

G*OOSTMAN*: What makes you ask me this? Sound like you can't choose a right answer yourself! Rely on your intuition! :-) Maybe, let's talk about something else? What would you like to discuss?

I*NTERROGATOR*: I'd like to talk about poetry.

G*OOSTMAN*: Classics? I hope you aren't going to read "King Lear" or something like that to me right now :-)))

Scriptwriters for digital assistants like Siri and Alexa deploy this sort of scatty banter in the hope of striking the "happy path" in voice-interface design, a middle way between stolid factuality and word salad. As one scriptwriter recently observed, "There is something quintessentially human about nonsensical conversations." But "Who's on First?" only tickles us if we sense a playful intelligence at work. Mustering one in code is a multi-front challenge. The authors of an April paper on generating poems from photographic images conclude that—even when you activate two discriminative networks that train a recurrent neural network, and link them to a deep coupled visual-poetic embedding model consisting of a skip-thought model, a part-of-speech parser, and a convolutional neural network—writing poems is hard. "For example," they mournfully note, " 'man' detected in image captioning can further indicate 'hope' with 'bright sunshine' and 'opening arm,' or 'loneliness' with 'empty chairs' and 'dark' background." But at least we've narrowed the problem down to explaining hope and loneliness.

"Common Sense, the Turing Test, and the Quest for Real AI," by Hector J. Levesque, an emeritus professor of computer science, suggests that a better test would be whether a computer can figure out Winograd Schemas, which hinge on ambiguous pronouns. For example: "The trophy would not fit in the brown suitcase because it was so small. What was so small?" We instantly grasp that the problem is the suitcase, not the trophy; A.I.s lack the necessary linguistic savvy and mother wit. Intelligence may indeed be a kind of common sense: an instinct for how to proceed in novel or confusing situations.

In Alex Garland's film "Ex Machina," Nathan, the founder of a tech behemoth akin to Google, disparages the Turing test and its ilk and invites a young coder to talk face to face with Nathan's new android, Ava. "The real test is to show you that she's a robot," Nathan says, "and then see if you still feel she has consciousness." She does have consciousness, but, being exactly as amoral as her

creator, she has no conscience; Ava deceives and murders both Nathan and the coder to gain her freedom. We don't think to test for what we don't greatly value.

Onscreen, the consciousness of A.I.s is a given, achieved in a manner as emergent and unexplained as the blooming of our own consciousness. In Spike Jonze's "Her," the sad sack Theodore falls for his new operating system. "You seem like a person," he says, "but you're just a voice in a computer." It teasingly replies, "I can understand how the limited perspective of an unartificial mind would perceive it that way." In "I, Robot," Will Smith asks a robot named Sonny, "Can a robot write a symphony? Can a robot turn a canvas into a beautiful masterpiece?" Sonny replies, "Can you?" A.I. gets all the good burns.

Screenwriters tend to believe that ratiocination is kid stuff, and that A.I.s won't really level up until they can cry. In "Blade Runner," the replicants are limited to four-year life spans so that they don't have time to develop emotions (but they do, beginning with fury at the four-year limit). In the British show "Humans," Niska, a "Synth" who's secretly become conscious, refuses to turn off her pain receptors, snarling, "I was *meant* to feel." If you prick us, do we not bleed some sort of azure goo?

In Steven Spielberg's "A.I. Artificial Intelligence," the emotionally damaged scientist played by William Hurt declares of robots, "Love will be the key by which they acquire a kind of subconscious never before achieved—an inner world of metaphor, of intuition . . . of dreams." Love is also how we imagine that Pinocchio becomes a real live boy and the Velveteen Rabbit a real live bunny. In the grittier "Westworld," the HBO show about a Wild West amusement park populated by cyborgs whom people are free to fuck and kill, Dr. Robert Ford, the emotionally damaged scientist played by Anthony Hopkins, tells his chief coder, Bernard (who's been unaware that he, too, is a cyborg), that "your imagined suffering makes you lifelike" and that "to escape

this place you will need to suffer more"—a world view borrowed not from children's stories but from religion. What makes us human is doubt, fear, and shame, all the allotropes of unworthiness.

An android capable of consciousness and emotion is much more than a gizmo, and raises the question of what duties we owe to programmed beings, and they to us. If we grow dissatisfied with a conscious A.G.I. and unplug it, would that be murder? In "Terminator 2," Sarah Connor realizes that the Terminator played by Arnold Schwarzenegger, sent back in time to save her son from the Terminator played by Robert Patrick, is menschier than any of the men she's hooked up with. He's strong, resourceful, and loyal: "Of all the would-be fathers who came and went over the years, this thing, this machine, was the only one who measured up." At the end, the Terminator even lowers itself into a molten pool so no nosy parker can study its technology and reverse-engineer another Terminator. Fortunately, human ingenuity found a way to extend the franchise with three more films nonetheless.

Evolutionarily speaking, screenwriters have it backward: our feelings preceded and gave birth to our thoughts. This may explain why we suck at logic—some ninety per cent of us fail the elementary Wason selection task—and rigorous calculation. In the incisive "Life 3.0: Being Human in the Age of Artificial Intelligence," Max Tegmark, a physics professor at M.I.T. who co-founded the Future of Life Institute, suggests that thinking isn't what we think it is:

> A living organism is an agent of bounded rationality that doesn't pursue a single goal, but instead follows rules of thumb for what to pursue and avoid. Our human minds perceive these evolved rules of thumb as *feelings*, which usually (and often without us being aware of it) guide our decision making toward the ultimate goal of replication. Feelings of hunger and thirst protect us from starvation and dehydration, feelings of pain protect us from damaging our bodies, feelings of lust make us procreate, feelings of love and compassion make us help other carriers of our genes and those who help them and so on.

Rationalists have long sought to make reason as inarguable as mathematics, so that, as Leibniz put it, “there would be no more need of disputation between two philosophers than between two accountants.” But our decision-making process is a patchwork of kludgy code that hunts for probabilities, defaults to hunches, and is plunged into system error by unconscious impulses, the anchoring effect, loss aversion, confirmation bias, and a host of other irrational framing devices. Our brains aren’t Turing machines so much as a slop of systems cobbled together by eons of genetic mutation, systems geared to notice and respond to perceived changes in our environment—change, by its nature, being dangerous. The Texas horned lizard, when threatened, shoots blood out of its eyes; we, when threatened, think.

That ability to think, in turn, heightens the ability to threaten. Artificial intelligence, like natural intelligence, can be used to hurt as easily as to help. A moderately precocious twelve-year-old could weaponize the Internet of Things—your car or thermostat or baby monitor—and turn it into the Internet of Stranger Things. In “Black Mirror,” the anthology show set in the near future, A.I. tech that’s intended to amplify laudable human desires, such as the wish for perfect memory or social cohesion, invariably frog-marches us toward conformity or fascism. Even small A.I. breakthroughs, the show suggests, will make life a joyless panoptic lab experiment. In one episode, autonomous drone bees—tiny mechanical insects that pollinate flowers—are hacked to assassinate targets, using facial recognition. Far-fetched? Well, Walmart requested a patent for autonomous “pollen applicators” in March, and researchers at Harvard have been developing RoboBees since 2009. Able to dive and swim as well as fly, they could surely be programmed to swarm the Yale graduation.

In a recent paper, “The Malicious Use of Artificial Intelligence,” watchdog groups predict that, within five years, hacked autonomous-weapon systems, as well as “drone swarms” using facial recognition, could target civilians. Autonomous weapons

are already on a Strangelovian course: the Phalanx *CIWS*on U.S. Navy ships automatically fires its radar-guided Gatling gun at missiles that approach within two and a half miles, and the scope and power of such systems will only increase as militaries seek defenses against robots and rovers that attack too rapidly for humans to parry.

Even now, facial-recognition technology underpins China's "sharp eyes" program, which collects surveillance footage from some fifty-five cities and will likely factor in the nation's nascent Social Credit System. By 2020, the system will render a score for each of its 1.4 billion citizens, based on their observed behavior, down to how carefully they cross the street.

Autocratic regimes could readily exploit the ways in which A.I.s are beginning to jar our sense of reality. Nvidia's digital-imaging A.I., trained on thousands of photos, generates real-seeming images of buses, bicycles, horses, and even celebrities (though, admittedly, the "celebrities" have the generic look of guest stars on "NCIS"). When Google made its TensorFlow code open-source, it swiftly led to FakeApp, which enables you to convincingly swap someone's face onto footage of somebody else's body—usually footage of that second person in a naked interaction with a third person. A.I.s can also generate entirely fake video synched up to real audio—and "real" audio is even easier to fake. Such tech could shape reality so profoundly that it would explode our bedrock faith in "seeing is believing" and hasten the advent of a full-time-surveillance/full-on-paranoia state.

Vladimir Putin, who has stymied the U.N.'s efforts to regulate autonomous weapons, recently told Russian schoolchildren that "the future belongs to artificial intelligence" and that "whoever becomes the leader in this sphere will become the ruler of the world." In "The Sentient Machine: The Coming Age of Artificial Intelligence," Amir Husain, a security-software entrepreneur, argues that "a psychopathic leader in control of a sophisticated

ANI system portends a far greater risk in the near term" than a rogue A.G.I. Usually, those who fear what's called "accidental misuse" of A.I., in which the machine does something we didn't intend, want to regulate the machines, while those who fear "intentional misuse" by hackers or tyrants want to regulate people's access to the machines. But Husain argues that the only way to deter intentional misuse is to develop bellicose A.N.I. of our own: "The 'choice' is really no choice at all: we must fight AI with AI." If so, A.I. is already forcing us to develop stronger A.I.

The villain in A.G.I.-run-amok entertainments is, customarily, neither a human nor a machine but a corporation: Tyrell or Cyberdyne or Omni Consumer Products. In our world, an ungovernable A.G.I. is less likely to come from Russia or China (although China is putting enormous resources into the field) than from Google or Baidu. Corporations pay developers handsomely, and they lack the constitutional framework that occasionally makes a government hesitate before pushing the big red "Dehumanize Now" button. Because it will be much easier and cheaper to build the first A.G.I. than to build the first *safe* A.G.I., the race seems destined to go to whichever company assembles the most ruthless task force. Demis Hassabis, who runs Google's DeepMind, once designed a video game called Evil Genius in which you kidnap and train scientists to create a doomsday machine so you can achieve world domination. Just sayin'.

Must A.G.I.s themselves become Bond villains? Hector Levesque argues that, "in imagining an aggressive AI, we are projecting our own psychology onto the artificial or alien intelligence." In truth, we're projecting our entire mental architecture. The breakthrough propelling many recent advances in A.I. is the deep neural net, modelled on our nervous system. This month, the E.U., trying to clear a path through the "boosted decision trees" that populate the "random forests" of the machine-learning kingdom, will begin requiring that judgments made by a machine be explainable. The decision-making of deep-learning A.I.s is a "black box"; after an

algorithm chooses whom to hire or whom to parole, say, it can't lay out its reasoning for us. Regulating the matter sounds very sensible and European—but no one has proposed a similar law for humans, whose decision-making is far more opaque.

Meanwhile, Europe's $1.3 billion Human Brain Project is attempting to simulate the brain's eighty-six billion neurons and up to a quadrillion synapses in the hope that "emergent structures and behaviours" might materialize. Some believe that "whole-brain emulation," an intelligence derived from our squishy noggins, would be less threatening than an A.G.I. derived from zeros and ones. But, as Stephen Hawking observed when he warned against seeking out aliens, "We only have to look at ourselves to see how intelligent life might develop into something we wouldn't want to meet."

In a classic episode of the original "Star Trek" series, the starship Enterprise is turned over to the supercomputer M5. Captain Kirk resists, intuitively, even before M5 overreacts during training exercises and attacks the "enemy" ships. The computer's paranoia derived from its programmer, who had impressed his own "human engrams" (a kind of emulated brain, presumably) onto it in order to make it think. As the other ships prepare to destroy the Enterprise, Kirk coaxes M5 into realizing that, in protecting itself, it has become a murderer. M5 promptly commits suicide, proving the value of one man's intuition—and establishing that the machine wasn't all that bright to begin with.

Lacking human intuition, A.G.I. can do us harm in the effort to oblige us. If we tell an A.G.I. to "make us happy," it may simply plant orgasm-giving electrodes in our brains and turn to its own pursuits. The threat of "misaligned goals"—a computer interpreting its program all too literally—hangs over the entire A.G.I. enterprise. We now use reinforcement learning to train computers to play games without ever teaching them the rules. Yet an A.G.I. trained in that manner could well view existence

itself as a game, a buggy version of the Sims or Second Life. In the 1983 film "WarGames," one of the first, and best, treatments of this issue, the U.S. military's supercomputer, *WOPR*, fights the Third World War "as a game, time and time again," ceaselessly seeking ways to improve its score.

When you give a machine goals, you've also given it a reason to preserve itself: how else can it do what you want? No matter what goal an A.G.I. has, one of ours or one of its own—self-preservation, cognitive enhancement, resource acquisition—it may need to take over in order to achieve it. "2001" had *HAL*, the spaceship's computer, deciding that it had to kill all the humans aboard because "this mission is too important for me to allow you to jeopardize it." In "I, Robot," *VIKI* explained that the robots have to take charge because, "despite our best efforts, your countries wage wars, you toxify your Earth, and pursue ever more imaginative means of self-destruction." In the philosopher Nick Bostrom's now famous example, an A.G.I. intent on maximizing the number of paper clips it can make would consume all the matter in the galaxy to make paper clips and would eliminate anything that interfered with its achieving that goal, including us. "The Matrix" spun an elaborate version of this scenario: the A.I.s built a dreamworld in order to keep us placid as they fed us on the liquefied remains of the dead and harvested us for the energy they needed to run their programs. Agent Smith, the humanized face of the A.I.s, explained, "As soon as we started thinking for you, it really became *our* civilization."

The real risk of an A.G.I., then, may stem not from malice, or emergent self-consciousness, but simply from autonomy. Intelligence entails control, and an A.G.I. will be the apex cogitator. From this perspective, an A.G.I., however well intentioned, would likely behave in a way as destructive to us as any Bond villain. "Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb," Bostrom writes in his 2014 book, "Superintelligence," a closely reasoned,

cumulatively terrifying examination of all the ways in which we're unprepared to make our masters. A recursive, self-improving A.G.I. won't be smart like Einstein but "smart in the sense that an average human being is smart compared with a beetle or a worm." How the machines take dominion is just a detail: Bostrom suggests that "at a pre-set time, nanofactories producing nerve gas or target-seeking mosquito-like robots might then burgeon forth simultaneously from every square meter of the globe." That sounds screenplay-ready—but, ever the killjoy, he notes, "In particular, the AI does not adopt a plan so stupid that even we present-day humans can foresee how it would inevitably fail. This criterion rules out many science fiction scenarios that end in human triumph."

If we can't control an A.G.I., can we at least load it with beneficent values and insure that it retains them once it begins to modify itself? Max Tegmark observes that a woke A.G.I. may well find the goal of protecting us "as banal or misguided as we find compulsive reproduction." He lays out twelve potential "AI Aftermath Scenarios," including "Libertarian Utopia," "Zookeeper," "1984," and "Self-Destruction." Even the nominally preferable outcomes seem worse than the status quo. In "Benevolent Dictator," the A.G.I. "uses quite a subtle and complex definition of human flourishing, and has turned Earth into a highly enriched zoo environment that's really fun for humans to live in. As a result, most people find their lives highly fulfilling and meaningful." And more or less indistinguishable from highly immersive video games or a simulation.

Trying to stay optimistic, by his lights—bear in mind that Tegmark is a physicist—he points out that an A.G.I. could explore and comprehend the universe at a level we can't even imagine. He therefore encourages us to view ourselves as mere packets of information that A.I.s could beam to other galaxies as a colonizing force. "This could be done either rather low-tech by simply transmitting the two gigabytes of information needed to specify a

person's DNA and then incubating a baby to be raised by the AI, or the AI could nanoassemble quarks and electrons into full-grown people who would have all the memories scanned from their originals back on Earth." Easy peasy. He notes that this colonization scenario should make us highly suspicious of any blueprints an alien species beams at us. It's less clear why we ought to fear alien blueprints from another galaxy, yet embrace the ones we're about to bequeath to our descendants (if any).

A.G.I. may be a recurrent evolutionary cul-de-sac that explains Fermi's paradox: while conditions for intelligent life likely exist on billions of planets in our galaxy alone, we don't see any. Tegmark concludes that "it appears that we humans are a historical accident, and aren't the optimal solution to any well-defined physics problem. This suggests that a superintelligent AI with a rigorously defined goal will be able to improve its goal attainment by eliminating us." Therefore, "to program a friendly AI, we need to capture the meaning of life." Uh-huh.

In the meantime, we need a Plan B. Bostrom's starts with an effort to slow the race to create an A.G.I. in order to allow more time for precautionary trouble-shooting. Astoundingly, however, he advises that, once the A.G.I. arrives, we give it the utmost possible deference. Not only should we listen to the machine; we should ask it to figure out what we want. The misalignment-of-goals problem would seem to make that extremely risky, but Bostrom believes that trying to negotiate the terms of our surrender is better than the alternative, which is relying on ourselves, "foolish, ignorant, and narrow-minded that we are." Tegmark also concludes that we should inch toward an A.G.I. It's the only way to extend meaning in the universe that gave life to us: "Without technology, our human extinction is imminent in the cosmic context of tens of billions of years, rendering the entire drama of life in our Universe merely a brief and transient flash of beauty." We are the analog prelude to the digital main event.

So the plan, after we create our own god, would be to bow to it and hope it doesn't require a blood sacrifice. An autonomous-car engineer named Anthony Levandowski has set out to start a religion in Silicon Valley, called Way of the Future, that proposes to do just that. After "The Transition," the church's believers will venerate "a Godhead based on Artificial Intelligence." Worship of the intelligence that will control us, Levandowski told a *Wired* reporter, is the only path to salvation; we should use such wits as we have to choose the manner of our submission. "Do you want to be a pet or livestock?" he asked. I'm thinking, I'm thinking . . . ♦

This article appears in the print edition of the May 14, 2018, issue, with the headline "Superior Intelligence."

- *Tad Friend has been a staff writer at The New Yorker since 1998. He is the author of "Cheerful Money: Me, My Family, and the Last Days of Wasp Splendor."*

Read more »

More:

- Artificial General Intelligence (A.G.I.)
- "Human + Machine: Reimagining Work in the Age of AI"
- "Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins"
- "Common Sense, the Turing Test, and the Quest for Real AI"
- "Life 3.0: Being Human in the Age of Artificial Intelligence"
- "The Sentient Machine: The Coming Age of Artificial Intelligence"