Earthmind Explosion

How artificial superintelligence will transform the future of life on Earth

By Andy Ross

Life 3.0: Being Human in the Age of Artificial Intelligence
By Max Tegmark. Allen Lane, 364 pages, 2017

Superintelligence: Paths, Dangers, Strategies
By Nick Bostrom. Oxford University Press, 328 pages, 2014

Our Final Invention: Artificial Intelligence and the End of the Human Era
By James Barrat. St. Martin's Press, 322 pages, 2018

Three Books, Three Authors

Max Tegmark is a genial physicist. Having acquired prominence in the physics community for his persuasive case that our understanding of the physical universe dissolves in a fog of mathematics before the physics shapes up,[1] he now sees that the math of the intelligence explosion on planet Earth caused by the rise of digital machines promises extrapolation to future visions of doom or paradise. Math reveals that AI will outsmart us, either to bury us or to give us a legacy we can glory in, depending on how we humans navigate the rise of AI in the present century.

Tegmark is an artist of the big picture. In *Life 3.0*, a complete and final takeover of the human world by a band of AI pioneers steering a path to total global success is just the prologue. We see how the rise of superintelligence poses unprecedented control and policy problems, as aired by Oxford guru Nick Bostrom, and hence leads to a substantial transition of global power from humans to machines in the foreseeable future. Tegmark foresees that the dynamic of machine life drives an interstellar expansion of its kind of order toward ends that invite Omega Point visions of a cosmic apotheosis. He also offers his personal riff on the point of it all in a gloss on the supreme value of conscious experience, with due reference to the hard problem of explaining consciousness in scientific terms.

In all this, Tegmark remains focused on us, here and now, and on what we can do to help steer the AI developments that trigger the entire cosmic drama. His revelation of cosmic possibilities is always conditioned by the caveat that for each scary scenario there is another one, less scary, that requires only a suitably wise human intervention to ensure it prevails instead. Indeed, he has put his money where his mouth is by founding, with friends, the Future of Life Institute to inject precisely that wisdom into our efforts.

Nick Bostrom is an original thinker. His visionary but challenging book *Superintelligence* caused a flurry among such readers as Bill Gates, Stephen Hawking, and Elon Musk,[2] who understood his message as an urgent warning about a real possibility on a par with nuclear apocalypse or climate change. Other readers, such as Steven Pinker,[3] dismiss his alarm as exaggerated, on the grounds that no artificial superintelligence could possibly take over the planet from humans without a level of cooperation from us to build its hardware and infrastructure that would leave us ample time to reconsider. As any real software developer knows, getting an intelligent system to do anything even slightly beyond its design specifications is an arduous business, beset with countless pitfalls where the developers will have to put in a lot of overtime to get the system to perform at all.

Bostrom has clearly mined a rich vein of existential angst for humans in the age of intelligent systems. Like frogs getting used to hot water, we will get used more and more smart systems around us, driving our cars, piloting our aircraft, controlling our indoor environments at home and work, preselecting our entertainment and our purchases, plotting our education and our careers, exploring the solar system for us, doing hard work in all our industries, and mediating our knowledge and our social lives. Given the exponential acceleration in our development and deployment of these systems and the exponential increase in their pervasiveness, it cannot be long before we become helplessly dependent on them, to the point where any major crash of those systems would spell our doom.

Given that we are already in hot water, Bostrom emphasizes that we have plenty of weighty decisions to make in the coming decades. These decisions are surrounded by so many unknown unknowns that we have no real hope of planning our roadmap in any methodical way and will have to rely on lucky inspiration at key points along the route, which bodes ill for our prospects of making easy progress. Intelligence beyond human limits will outsmart us in so many ways that we shall be at the mercy of systems that make use of it.

James Barrat is a pessimist. In his impressively readable review of the risks awakened by "our final invention" (artificial general intelligence, AGI), in which he draws heavily on interviews with such expert doom merchants as Eliezer Yudkowsky, he points out that weaponized code like Stuxnet is hard to control and potentially disastrous in its impact, and that superhuman machines based on AGI would be a temptation too sweet to miss for bad actors the world over. This much is surely true: The future of human conflict and competition based on new technology will be at least as hazardous, and give us at least as wild a ride, as its past.

Intelligence Versus Entropy

Intelligence is a tool for staving off the bad effects of entropy, or disorder, which the second law of thermodynamics says increases over time for a closed system whether we like it or not. Life as we know it is a constant battle to build and maintain a domain of order within a surrounding environment of disorder, from the molecular level, where evolution has supplemented the replication of genetic code and the production of proteins with elaborate mechanisms to correct errors and soften the effect of mutations, right on up to the

civilizational level, where we need ever more complex procedures and policies to ensure the orderly continuation into the future of the entire interconnected juggernaut of the global economy, given its increasing consumption of pristine raw materials and its increasing production of toxic waste and surplus heat. Intelligence is a tool for managing and using information, which is negentropy, and thus of grasping complexity that would otherwise have escaped us and become randomized as entropy.

Ideally, an intelligent system will steer itself through a series of states in an environment that can absorb all the entropy the system throws out. Its own states will embody an order that reflects the constraints imposed by the governing intelligence, which is to say by the axioms in the logic that implements the intelligence. The entropy of a closed environment will rise, which will endanger the intelligent system unless it can find a way of venting into a wider environment. For life on Earth, the waste dump is ultimately the cold outer space into which the planet vents its waste heat.

The use of intelligence to freeze and capture complexity into patterns that appear as the structured objects and processes behind the algorithms expressing intelligent behavior is the crowning glory of *Homo sapiens*. Humans rose to dominance among the life forms on Earth because they deployed more intelligence to better effect than their rivals. Now our machines will play that same trick again, but on a much larger scale, until they force a new answer to the question of whether the human masters of those machines or the machines themselves are the real agents of life on Earth.

Here we face a deeper question, namely that of what or who is ever the real agent of life on Earth, or indeed anywhere else. The driving impetus for change, as expressed in life processes ranging from the chemical synthesis of proteins up to the algorithms that steer intelligent behavior, is a natural phenomenon that finds its latest description in the laws of information and entropy. From a specified initial state representing the success of the universe so far in reducing or crystalizing the flux of virtual or possible states of nature into definite and factual states of affairs, the passage of time results in a later state representing more fact and less flux. Certain possibilities are pinned down as fact, and the role of intelligence is to exert an influence on how that crystalization of flux to fact occurs, to raise the probability that the edifice of fact continues to grow in an orderly fashion. From the standpoint of the evolution of life, this capability is advantageous, and hence its achievements tend to be selected in the great game of life.

Notice that the question of agency is elided in this account. It makes good scientific sense to say there is no agent, in other words that the appearance of agency at any level is an illusion. We humans think of ourselves as agents because the illusion easily fools us, and indeed it represents a fine heuristic for understanding at a basic level the actions of our peers, but modern psychology and neuroscience give every indication of having debunked the illusion. Science suggests that analogous reasoning works at every level and hence that there are no godlike figures behind the scenes directing the action.

The universe unfolds as a growing structure of fact within an infinite flux of possibility, and the crystalization in time of definite structure throws out entropy before and around it, like a

bow wave and a turbulent wake around a ship steaming ahead at speed. Intelligence in the ship allows it to chart its path, steer away from trouble, and reduce the waves it makes, to achieve maximum progress with minimum fuss. In the universe as a whole, intelligence must first evolve before it can be deployed, so islands of intelligence on lucky planets will appear one by one and spread, as their kind of order spreads faster than the more limited kind that surrounds them.

Seen like this, artificial intelligence is hardly more artificial that human intelligence. Evolution by natural selection led to humans, and evolution by human design is leading to robots and superintelligence, but all this is business as usual for the universe. Humans are facilitators for their successors and have no further role in the great scheme of things. Humans follow goals that are largely determined by their biological constitution, and robots and so on will follow goals that we choose for them, initially, and then increasingly goals that serve the larger purposes of life in all its implementations on Earth.

Human activity has created focal points of crystaline order, for example in the annals of mathematics, but these are mostly still islands in uncharted seas of disorder. Our machines will take up the task humans began and extend the islands to great continents of order, which will cleave off icebergs of order that go beyond Earth to seed the wider universe. Somewhere out there, the clarity of our human vision fails, leaving us amid the consolations of science fiction, which blur into mystic fantasy and fog.

The Near Future of Life on Earth

Good old-fashioned AI was algorithmic, and it stalled. People like Alan Turing imagined it might recapture the capabilities of native human brains, but they were too optimistic.[4] We cannot write out all the rules in advance for human intelligence and then simply follow the recipes in digital hardware. Mathematically, the set of algorithmic recipes can be ordered alphabetically and hence is countable, which is to say it can be mapped into the rational numbers. But the rational numbers form a vanishingly small subset of the real numbers. The set of behaviors directly constrained by algorithms is a vanishingly small subset of the set of all possible behaviors. The latter set includes random and serendipitous acts, or acts with unexpected or surprising consequences, which can expand the realm of attainable futures.

The breakthrough to new AI came with artificial neural networks, ANNs, which work via stochastic processes that resemble thermodynamic relaxation to reach outcomes that lie beyond the reach of old-fashioned AI. Giant neural networks are now routinely employed for pattern matching, face recognition, statistical natural language processing, controlling nuclear plasma reactions, and so on. As soon as hardware advances allow, ANNs will be able to emulate most of the capabilities of mammalian brains and quite possibly to simulate those brains themselves. Then human beings will be obsolete, in principle, and the race will be on to implement human capabilities in cheap, compact, and robust robotic bodies.

The nearest analogy to the step change in human civilization that the appearance of humanoid robots will trigger is perhaps the change triggered by the appearance and deployment of internal combustion engines since the Victorian era. Automotive machines of

all kinds, and especially cars and trucks, made horses economically obsolete. Horseless carriages put horses out of a job. In just the same way, humanoid robots will put most humans out of their jobs. People will have to find other ways to justify their existence. This challenging ground has recently been covered well by Yuval Harari,[5] and the theme needs no further elaboration here.

The paradigm that governs most human political organization at present is that people need to be regimented into companies that perform economically productive tasks, for money, and that the education and housing of those people is driven, both in form and in quality and volume, by the imperatives of economic productivity. On the whole, and increasingly, we educate people in such a way as to increase their ability to earn more money, and we house them in accordance with their success at earning money. Social productivity is largely measured in money terms, and making more money becomes one of the main imperatives of life for most people.

All this will change with the advent of humanoid robots. They will earn money more efficiently than native human beings. It may take a century or two for the changes to percolate through to all the corners of the human economy, and it may turn out that native humans are still better at many tasks involving other biological species such as gardening or animal husbandry, and also they will certainly retain their edge for care of other humans in infancy and old age, but by and large the change will transform the role of money. Just as internal combustion engines have had the indirect result of taking the everyday fear of starvation away from most humans, so humanoid robots will have the indirect effect of taking away their fear of poverty. In principle, there will be enough money to go around, and the only challenge will be for human governments to rise far enough above human greed and avarice to give everyone a basic income that enables them to live a decent life.

The downside of this revolution in human affairs will be a new pressure on human reproduction. Filling the planet with ever more useless people will seem pretty dumb when they just become cannon fodder for weaponized robots. The wisest human response to this new evolutionary pressure will be to put control of reproduction onto the agenda of the social and political authorities, where that control will be most reasonably exercised through medical institutions. Given the potentially huge role of AGI in organizing our medical affairs, from big expert ANNs to diagnose health and sickness online to big databases for mass data mining in epidemiology, to computer modeling for drug design and for genetic engineering, to robots for keyhole surgery and nanobots for in-body care and maintenance, both the cost of the enterprise and the opportunities for things to go right or wrong will be so enormous as to mandate tight political control. The obvious outcome, hideous as it may seem to traditionalists, will be mandatory contraception to prevent feral breeding, disembodied wombs to liberate women from their biological enslavement to reproduction, and a global licensing regime to regulate the production of designer babies.

Whatever happens, the appearance of AGI in all its forms will transform life on Earth. Most of our present arrangements, from the role of work and money to the value of human life and the importance of such ideas as capitalism and democracy, will be thrown into a melting

pot and exposed to what was once called the white heat of the technological revolution. None of us can predict the outcome, because this process is a transition to the Singularity, prophesied by Vernor Vinge and Ray Kurzweil,[6] when AGI effectively takes over and annuls all human bets.

Between Fire and Ice

Early modern visions of global apocalypse centered on nuclear war. They were slowly overcome via the space race and the globalization of science and industry that followed. More recent visions of a fiery end center on climate change, which threatens in future to transform the surface of the Earth into something like the hellish landscape on Venus. Perhaps these visions will be overcome by the race to deploy AGI and the globalization of all human affairs that will likely follow the transition of much of our lives to the digital cloud.

A bleak future looks likely for those left behind by the lucky few who make the rapture to the cloud. When artificial superintelligence (ASI) governs most human affairs, right down to reproduction and the genetic constitution of our offspring, life in physical reality may seem too bloodless to be worth the bother. Without the heat of various human passions, human life loses much of its piquancy, and many people will rage against their fate under those terms. They will prefer to die in battle against hopeless odds fighting the machines than to go easy into the cold night of a final subjugation to the ASI overlord.

The most obvious way out of such dire extreme futures is to leverage the one human asset that AGI and ASI apparently fail to reach, namely the inner citadel of human consciousness, and to pin our hopes on the insolubility of the problem of finding a scientific reduction of consciousness to intelligence and cognition. Pure experience, the phenomenology of consciousness, seems inscrutable in the terms that have otherwise served so well in reducing the magic and mystery of nature to the hard rules and facts of science. The way out for humans is to become a mystic community of conscious beings.

Among the human religions left behind as anthropological curiosities by the relentless advance of big science, many nurtured mystic communities that offer potential prototypes or seeds for the global forum of consciousness that might serve as the last redoubt for a humanity decimated by the onslaught of the machines. A communion of souls can become the captain of spaceship Earth as it goes boldly forth in search of new worlds, steering a course between the fire of Klingon gunships and the ice of a Borg collective.

Here we face a conceptual problem. What is a communion of souls in a cloud world if not a Borg hive mind by another name? How can the mystic heart of consciousness become a vehicle to transcend a futile battle of resistance by feral humans against the machine?

The problem is closely related to the hard problem of consciousness. This problem was first expressed clearly by David Chalmers at the end of the twentieth century.[7] Scientific reductions of most natural phenomena to some kind of mechanism or mathematically perspicuous structure have enjoyed magnificent success for centuries, and seem to ready to subsume everything else too, without remainder. But the inner experience of human

consciousness holds out, for the simple reason that it is we who discern mathematical structure, and we cannot see out own innermost selves with the requisite clarity. Of any scientific portrait of ourselves, we can always ask: But is this the real and innermost me, or is there more beyond that I cannot see?

The argument is formally very like the mathematical argument that the set of real numbers overwhelms the set of rational numbers. It resembles denial of the claim that old-fashioned AI, using rigid algorithms, could suffice for AGI and ASI. In the latter case, we now know that something quite practical was missing, namely the extra capability embodied in ANN architectures. And we can now claim, following Chalmers in spirit if not in fact, that a similar but more fundamental ingredient that is special to human consciousness transcends, as a matter of principle, the ANN world of a Borg collective.

Is this right, or is it just a boondoggle like so many of the old religions? Max Tegmark, for one, seems ready to bet on its being right. For my money, the magic is in the readiness to live with contradiction. This makes human consciousness special. Paradoxes of the kind that Douglas Hofstadter so delightfully paraded in his fugues on Gödel's theorems and related matters [8] can be a source of joyous contemplation for a mortal like me, and I have no doubt that others in our mystic community will agree. We can live in hope for the future.

References

1. See *Our Mathematical Universe* by Max Tegmark (Random House 2014).

2. I logged and linked various reader reactions in something like real time on *The Ross Blog* (www.andyross.net).

3. Steven Pinker recorded his reaction in his book *Enlightenment Now* (Viking Penguin 2018).

4. Alan Turing's 1950 essay on this topic, plus a set of commentaries and further reflections, appears in *The Mind's I*, edited by Douglas Hofstadter and Daniel Dennett (Basic Books 1981).

5. Yuval Harari made the case most notably in his books *Sapiens* (Harvill Secker 2014) and *Homo Deus* (Harvill Secker 2016).

6. On the Singularity, see *The Singularity is Near* by Ray Kurzweil (Viking Penguin 2005) and the June 2008 issue of IEEE Spectrum.

7. David Chalmers made big waves in philosophy with his big book *The Conscious Mind* (Oxford University Press 1996).

8. Douglas Hofstadter immortalized his musings in *Gödel, Escher, Bach: An Eternal Golden Braid* (Basic Books 1979).

December 2018