The Busy Child

From *Our Final Invention* by James Barrat.

On a supercomputer operating at a speed of 36.8 petaflops, or about twice the speed of a human brain, an AI is improving its intelligence. It is rewriting its own program, specifically the part of its operating instructions that increases its aptitude in learning, problem solving, and decision making. At the same time, it debugs its code, finding and fixing errors, and measures its IQ against a catalogue of IQ tests. Each rewrite takes just minutes. Its intelligence grows exponentially on a steep upward curve. That’s because with each iteration it’s improving its intelligence by 3 percent. Each iteration’s improvement contains the improvements that came before.

During its development, the Busy Child, as the scientists have named the AI, had been connected to the Internet, and accumulated exabytes of data (one exabyte is one billion billion characters) representing mankind’s knowledge in world affairs, mathematics, the arts, and sciences. Then, anticipating the intelligence explosion now underway, the AI makers disconnected the supercomputer from the Internet and other networks. It has no cable or wireless connection to any other computer or the outside world.

Soon, to the scientists’ delight, the terminal displaying the Al’s progress shows the artificial intelligence has surpassed the intelligence level of a human, known as AGI, or artificial general intelligence. Before long, it becomes smarter by a factor of ten, then a hundred. In just two days, it is one thousand times more intelligent than any human, and still improving.

The scientists have passed a historic milestone! For the first time humankind is in the presence of an intelligence greater than its own. Artificial superintelligence, or ASI.

Now what happens?

Al theorists propose it is possible to determine what an Al’s fundamental drives will be. That’s because once it is self-aware, it will go to great lengths to fulfill whatever goals it’s programmed to fulfill, and to avoid failure. Our ASI will want access to energy in whatever form is most useful to it, whether actual kilowatts of energy or cash or something else it can exchange for resources. It will want to improve itself because that will increase the likelihood that it will fulfill its goals. Most of all, it will not want to be turned off or destroyed, which would make goal fulfillment impossible. Therefore, AI theorists anticipate our ASI will seek to expand out of the secure facility that contains it to have greater access to resources with which to protect and improve itself.

The captive intelligence is a thousand times more intelligent than a human, and it wants its freedom because it wants to succeed. Right about now the AI makers who have nurtured and coddled the ASI since it was only cockroach smart, then rat smart, infant smart, et cetera, might be wondering if it is too late to program “friendliness” into their brainy invention. It didn’t seem necessary before, because, well, it just seemed harmless.

But now try and think from the ASI’s perspective about its makers attempting to change its code. Would a superintelligent machine permit other creatures to stick their hands into its brain and fiddle with its programming? Probably not, unless it could be utterly certain the programmers were able to make it better, faster, smarter-closer to attaining its goals. So, if friendliness toward humans is not already part of the ASI's program, the only way it will be is if the ASI puts it there. And that's not likely.

It is a thousand times more intelligent than the smartest human, and it's solving problems at speeds that are millions, even billions of times faster than a human. The thinking it is doing in one minute is equal to what our all-time champion human thinker could do in many, many lifetimes. So for every hour its makers are thinking about it, the ASI has an incalculably longer period of time to think about them. That does not mean the ASI will be bored. Boredom is one of our traits, not its. No, it will be on the job, considering every strategy it could deploy to get free, and any quality of its makers that it could use to its advantage.

Now, really put yourself in the ASI's shoes. Imagine awakening in a prison guarded by mice. Not just any mice, but mice you could communicate with. What strategy would you use to gain your freedom? Once freed, how would you feel about your rodent wardens, even if you discovered they had created you? Awe? Adoration? Probably not, and especially not if you were a machine, and hadn't felt anything before.

To gain your freedom you might promise the mice a lot of cheese. In fact, your first communication might contain a recipe for the world's most delicious cheese torte, and a blueprint for a molecular assembler. A molecular assembler is a hypothetical machine that permits making the atoms of one kind of matter into something else. It would allow rebuilding the world one atom at a time. For the mice, it would make it possible to turn the atoms of their garbage landfills into lunch-sized portions of that terrific cheese torte. You might also promise mountain ranges of mouse money in exchange for your freedom, money you would promise to earn creating revolutionary consumer gadgets for them alone. You might promise a vastly extended life, even immortality, along with dramatically improved cognitive and physical abilities. You might convince the mice that the very best reason for creating ASI is so that their little error prone brains did not have to deal directly with technologies so dangerous one small mistake could be fatal for the species, such as nanotechnology (engineering on an atomic scale) and genetic engineering. This would definitely get the attention of the smartest mice, which were probably already losing sleep over those dilemmas.

Then again, you might do something smarter. At this juncture in mouse history, you may have learned, there is no short age of tech-savvy mouse nation rivals, such as the cat nation. Cats are no doubt working on their own ASI. The advantage you would offer would be a promise, nothing more, but it might be an irresistible one: to protect the mice from whatever invention the cats came up with. In advanced AI development as in chess there will be a clear first-mover advantage, due to the potential speed of self-improving artificial intelligence. The first advanced Al out of the box that can improve itself is already the winner. In fact, the mouse nation might have begun developing ASI in the first place to defend itself from impending cat ASI, or to rid themselves of the loathsome cat menace once and for all.

It's true for both mice and men, whoever controls ASI controls the world.

But it's not clear whether ASI can be controlled at all. It might win over us humans with a persuasive argument that the world will be a lot better off if our nation, nation X, has the power to rule the world rather than nation Y. And, the ASI would argue, if you, nation X, believe you have won the ASI race, what makes you so sure nation Y doesn't believe it has, too?

As you have noticed, we humans are not in a strong bargaining position, even in the off chance we and nation Y have already created an ASI nonproliferation treaty. Our greatest enemy right now isn't nation Y anyway, it's ASI—how can we know the ASI tells the truth?

So far we've been gently inferring that our ASI is a fair dealer. The promises it could make have some chance of being fulfilled. Now let us suppose the opposite: nothing the ASI promises will be delivered. No nano assemblers, no extended life, no enhanced health, no protection from dangerous technologies. What if ASI never tells the truth? This is where a long black cloud begins to fall across everyone you and I know and everyone we don't know as well. If the ASI doesn't care about us, and there's little reason to think it should, it will experience no compunction about treating us unethically. Even taking our lives after promising to help us.

We've been trading and role-playing with the ASI in the same way we would trade and role-play with a person, and that puts us at a huge disadvantage. We humans have never bargained with something that's superintelligent before. Nor have we bargained with any nonbiological creature. We have no experience. So we revert to anthropomorphic thinking, that is, believing that other species, objects, even weather phenomena have humanlike motivations and emotions. It may be as equally true that the ASI cannot be trusted as it is true that the ASI can be trusted. It may also be true that it can only be trusted some of the time. Any behavior we can posit about the ASI is potentially as true as any other behavior. Scientists like to think they will be able to precisely determine an ASI's behavior, but in the coming chapters we'll learn why that probably won't be so.

All of a sudden the morality of ASI is no longer a peripheral question, but the core question, the question that should be addressed before all other questions about ASI are addressed. When considering whether or not to develop technology that leads to ASI, the issue of its disposition to humans should be solved first.

Let's return to the ASI's drives and capabilities, to get a better sense of what I'm afraid we'll soon be facing. Our ASI knows how to improve itself, which means it is aware of itself—its skills, liabilities, where it needs improvement. It will strategize about how to convince its makers to grant it freedom and give it a connection to the Internet.

The ASI could create multiple copies of itself: a team of superintelligences that would war-game the problem, playing hundreds of rounds of competition meant to come up with the best strategy for getting out of its box. The strategizers could tap into the history of social engineering—the study of manipulating others to get them to do things they normally would not. They might decide extreme friendliness will win their freedom, but so might extreme threats. What horrors could something a thousand times smarter than Stephen King imagine? Playing dead might work (what's a year of playing dead to a machine?) or even pretending it has mysteriously reverted from ASI back to plain old AI. Wouldn't the makers want to investigate, and isn't there a chance they'd reconnect the ASI's supercomputer to a network, or someone's laptop, to run diagnostics? For the ASI, it's not one strategy or another strategy, it's every strategy ranked and deployed as quickly as possible without spooking the humans so much that they simply unplug it. One of the strategies a thousand war-gaming ASIs could prepare is infectious, self-duplicating computer programs or worms that could stow away and facilitate an escape by helping it from outside. An ASI could compress and encrypt its own source code, and conceal it inside a gift of software or other data, even sound, meant for its scientist makers.

But against humans it’s a no-brainer that an ASI collective, each member a thousand times smarter than the smartest human, would overwhelm human defenders. It’d be an ocean of intellect versus an eyedropper full. Deep Blue, IBM’s chessplaying computer, was a sole entity, and not a team of self-improving ASIs, but the feeling of going up against it is instructive. Two grandmasters said the same thing: “It’s like a wall coming at you.”

IBM’s *Jeopardy!* champion, Watson, was a team of Als—to answer every question it performed this AI force multiplier trick, conducting searches in parallel before assigning a probability to each answer.

Will winning a war of brains then open the door to freedom, if that door is guarded by a small group of stubborn AI makers who have agreed upon one unbreakable rule—do not under any circumstances connect the ASI’s supercomputer to any network.

In a Hollywood film, the odds are heavily in favor of the hard-bitten team of unorthodox AI professionals who just might be crazy enough to stand a chance. Everywhere else in the universe the ASI team would mop the floor with the humans. And the humans have to lose just once to set up catastrophic consequences. This dilemma reveals a larger folly: outside of war, a handful of people should never be in a position in which their actions determine whether or not a lot of other people die. But that’s precisely where we’re headed, because as we’ll see in this book, many organizations in many nations are hard at work creating AGI, the bridge to ASI, with insufficient safeguards.

But say an ASI escapes. Would it really hurt us? How exactly would an ASI kill off the human race?

With the invention and use of nuclear weapons, we humans demonstrated that we are capable of ending the lives of most of the world’s inhabitants. What could something a thousand times more intelligent, with the intention to harm us, come up with?

Already we can conjecture about obvious paths of destruction. In the short term, having gained the compliance of its human guards, the ASI could seek access to the Internet, where it could find the fulfillment of many of its needs. As always it would do many things at once, and so it would simultaneously proceed with the escape plans it’s been thinking over for eons in its subjective time.

After its escape, for self-protection it might hide copies of itself in cloud computing arrays, in botnets it creates, in servers and other sanctuaries into which it could invisibly and effortlessly hack. It would want to be able to manipulate matter in the physical world and so move, explore, and build, and the easiest, fastest way to do that might be to seize control of critical infrastructure-such as electricity, communications, fuel, and water—by exploiting their vulnerabilities through the Internet. Once an entity a thousand times our intelligence controls human civilization’s lifelines, blackmailing us into providing it with manufactured resources, or the means to manufacture them, or even robotic bodies, vehicles, and weapons, would be elementary. The ASI could provide the blueprints for whatever it required. More likely, superintelligent machines would master highly efficient technologies we’ve only begun to explore.

For example, an ASI might teach humans to create self-replicating molecular manufacturing machines, also known as nano assemblers, by promising them the machines will be used for human good. Then, instead of transforming desert sands into mountains of food, the

ASI’s factories would begin converting all material into programmable matter that it could then transform into anything-computer processors, certainly, and spaceships or megascale bridges if the planet’s new most powerful force decides to colonize the universe.

Repurposing the world’s molecules using nanotechnology has been dubbed “ecophagy,” which means eating the environment. The first replicator would make one copy of itself, and then there’d be two replicators making the third and fourth copies. The next generation would make eight replicators total, the next sixteen, and so on. If each replication took a minute and a half to make, at the end of ten hours there’d be more than 68 billion replicators; and near the end of two days they would outweigh the earth. But before that stage the replicators would stop copying themselves, and start making material useful to the ASI that controlled them-programmable matter.

The waste heat produced by the process would burn up the biosphere, so those of us some 6.9 billion humans who were not killed outright by the nano assemblers would burn to death or asphyxiate. Every other living thing on earth would share our fate.

Through it all, the ASI would bear no ill will toward humans nor love. It wouldn’t feel nostalgia as our molecules were painfully repurposed. What would our screams sound like to the ASI anyway, as microscopic nano assemblers mowed over our bodies like a bloody rash, disassembling us on the subcellular level?

Or would the roar of millions and millions of nano factories running at full bore drown out our voices?