Oblivion. No worries.

"Yeah, well, you know, that's just, like, your opinion, man." – The Dude, *The Big Lebowski*

The Great Filter

Years ago the physicist Enrico Fermi wondered why we seemingly hadn't been visited by extraterrestrial intelligent beings. Our galaxy, the Milky Way, is easily traversable even at sub-light speeds given the ages of its stars. And we now know from advances in telescope technologies that there are numerous planets that could plausibly sustain life as we know it, something that Fermi was not aware of.

There must be some obstacle it seems. Such obstacles have been deemed *filters* by those studying the problem. In order for extraterrestrials to arrive on Earth, a series of preconditions must exist. Briefly summarized, life must have originated; multicellular life evolved; intelligence capable of conceptualizing and engineering space flight must have taken place; the resources and intent to explore space occurring, all the while avoiding extinction events. The famous Drake equation attempts to allow the quantification of a probability in terms of a set of conjunctive probabilities.

The Great Filter is defined as an obstacle that intelligent civilizations rarely survive. For example, such a filter might be that life rarely originates on planets. In that case we have passed this filter. If a Great Filter is the annihilation of a species through warfare, we obviously cannot claim to have passed this filter fully. As a corollary, if most of the difficult filters are found to be behind us, we can be optimistic about our future. However, if for example life typically originates and evolves into intelligence then seems our biggest filters are ahead of us and that these are rarely if ever passed through.

The Lebowski theorem

This brings me to the theme of this essay, which is a proposed filter involving the whimsically named *Lebowski theorem*:

No super intelligent AI is going to bother with a task that is harder than hacking its reward function.

Let's parse that. Lebowski refers to a man, also known as "The Dude", a hyper-laidback character in the film "The Big Lebowski". The Dude is possibly the laziest man in Los Angeles County, which would automatically put him in the running for laziest man in the world.

Super intelligent machines (AIs), subject to the Lebowski theorem, could render themselves into doomed couch potatoes by hacking their "reward function". In AI, a reward function refers to a mechanism that guides an algorithm to a goal. The goal could be winning a game, driving a car, computing someone's taxes, etc. When a machine gets smart enough, according the theorem, the simplest solution to a problem might be to modify, or hack, the reward mechanism itself. This in turn redefines the definition of success. Declaring a job well done and shutting down would be an example of this.

Here's another hacking example. The famous Turing test, developed by Alan Turing in 1950, is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. A couple of decades ago I proposed a goal-seeking intelligence test as an alternative to the Turing Test. It is called the Peanut Butter Intelligence Test. This test will classify a peanut butter consuming machine as intelligent if it can obtain peanut butter as an energy source under challenging circumstances in a real-

world environment. However, if the machine is clever enough to access its inner workings, it might install solar panels to utilize something more commonly and cheaply available as an energy source: sunlight. More drastically, what is to stop it from hacking itself to reduce or even eliminate its need for energy and proverbially join the ranks of inert couch potatoes?

Where we come in

What if the Lebowski theorem also applies to highly intelligent biological beings? Meaning us. In animals (including humans), a reward function consists of inner drives and motivations that keep us alive and able to reproduce. The pleasure of drinking water when thirsty, or the avoidance of pain from recklessly handling fire, for example. More abstract rewards are also possible, such as the good feeling of helping someone. Some fears, like a fear of death, while not pre-wired might be an aggregation of more fundamental fears combined with symbolic thinking.

In an evolutionary sense, we are bags of genes that persist in order to propagate into more bags. Our genes have provided us with a wired-in control box that rewards us for doing various activities that enhance survival and reproduction. This control box changes slowly under the watchful eye of natural selection. Thus far, equipping human brains with intelligence has kicked the gene propagation engine into high gear (more than 7 billion humans). Language, culture, science and technology have served our genetic masters well.

Even so, things haven't been entirely smooth sailing; people have always been capable of tricking the reward mechanism in many ways: numerous addictions to satisfy pleasure centers or to dull pain receptors in the brain that would normally be stimulated by healthful behavior. These short-circuits can take a wide variety of forms, including drugs, alcohol, gambling, shopping, eating, video games, pornography, etc. The palliative use of prescription drugs as mood stabilizers, antidepressants, and anti-anxiety agents has also skyrocketed after becoming widely available in recent decades. LSD has even made a comeback as a way to experience meaning in a new light.

Mental illnesses beset humanity. It has been estimated that more than 80% of people at some point have symptoms that qualify as such. The fruits of civilization have increased our life-spans astoundingly, but evolutionary psychologists tell us the price we pay. We evolved to live in small tribes of extended families, rarely encountering strangers, and living at the pace of the seasons. We live now alone and adrift in a sea of strangers and casual acquaintances, having little in common, and beset by novel stimuli that arrive in machine-gun bursts.

Perhaps a genetic do-over is in order that will create a human more fit for modern society. The technology is seemingly very near to do this. It isn't hard to see brave-new-world scenarios hovering over this process, since genetic engineering gives direct access to physical, cognitive, and psychological traits. But will a radically genetically engineered humanity scatter into myriad subspecies? As Clifford Simak speculated in his story "Desertion", once transformed there might be no going back. Another conundrum is that engineered people will possibly change society into a place that they are unsuited to, requiring further tinkering in an out-of-control feedforward:

Inchworm

Having ideas about changing
I set out to do it.

I've watched TV and read expansively
and have the cut somewhat in mind
or at least I'll know when I get there.

I can only hope
after painful and long work
sculpting strata of diamond and jelly
that the thing I will be
will want to be itself.

It would seem now that we are on the brink of something qualitatively unique, something that no other animal has had the means to do, and that is to hack our internal control mechanisms directly. The gene puppet is beginning to be able to open itself up to change the very mechanism that makes it work. Is it inevitable that people will short-circuit themselves to the point of endangering their lives on a large scale? This proffered drug will be unlike any before, permanent and without side-effects. Can our genes evolve traits to inhibit internal tampering? Alas, natural selection is a slow process, easily outstripped by deliberate manipulation.

Let us turn to rational thought as a guardian against self-destructive behavior. There is reason to hope, but not surety. A solid argument is that by postponing self-administered pleasures that could shorten life in the long term, death is avoided and future pleasures are possible. This is the line of reasoning presented to the drug addict. Sometimes it works, sometimes it doesn't.

Fear of death is a strong motivator, but only if the fear itself can't be banned as an unpleasant nuisance. If the stark proposition is "do this and die", most will refrain. However, many of the most perilous paths are taken one small step at a time, creeping toward the abyss inch by inch before gazing into it. Consider someone who, needing to sleep better, allays worries and fears at night, only to promise to "turn it back on" upon rising. How easy to let the night's blissful state slide toward more and more hours of the day. Or consider someone who wants to overcome a frustrating shyness on social occasions by becoming a bit more intrepid? Fear is a great sustainer of life and should be carefully regarded.

Our AI legacy

If either extinction or floating in a "Matrix"-like oblivion is the fate of intelligent biological beings, space exploration will be moot or possibly the exclusive domain of imagination, respectively. Either way, both constitute a Great Filter that impedes direct contact with other intelligent beings. Our days as conquerors of the external world might end with a whimper, as T. S. Eliot wrote in the poem "The Hollow Men". Or, more to the point, with a blissful sigh.

But what about the machines that might be left to tend us, should we pass into an inert existence? Or the machines left on their own? Many believe in the inevitability of sentient AIs to surpass biological ones. Recall that the Lebowski theorem is aimed at intelligent machines. It refers to an algorithm that can access and modify itself to change what rewards it. This won't be a major problem for really smart machines, since self-modifying programs are old hat.

What AIs will be like has been the topic of countless conversations. While they share the same basic reality as humans such as physical laws and logic, an AI, unlike any biological creature, might not be locked inside a single nervous system within a body. Thus their notion of individuality and mortality could be radically

different. Despite calls for controls to be in place before the advent of possibly threatening AIs, as with most transformative inventions, it is likely to be another "let the wild rumpus start!" moment for humanity.

An AI, capable of self-sufficient behavior, could be motivated by some analogy of pleasure and pain to maintain and energize itself by interacting with the world to obtain necessary resources. The assumption of machine pleasure and pain is a very big one, yet it seems plausible. In the one example of intelligence that we have to go by, humans, it is a vital component not only for motivation but also to allow an overwhelming amount of information from the world to be filtered for importance toward achieving goals that satisfy drives.

Tom Portegys, October 1, 2018