Long-termism

by GWYNNE DYER

Which would be worse: a global nuclear war with all buttons pressed, or real, self-conscious artificial intelligence that goes rogue? You know, the central theme of the ?Terminator' movies.

An AI called Skynet wakes up and immediately realises that humanity could simply switch it off again, so it triggers a nuclear war that destroys most of mankind. The few survivors end up waging a losing war against the machines and extinction. But this fantasy has too many moving parts. Let's try again.

Which would be worse: a nuclear ?war orgasm' (Herman Kahn's description of the Pentagon's nuclear war strategy circa 1960) or a designer plague created in some secret biowar lab? The plague, obviously, because it could theoretically wipe out the human race, whereas all-out nuclear war probably can't.

The distinction between a 99% wipe-out and a 100% wipe-out is insignificant if you happen to be one of the victims, but Oxford University philosopher Derek Parfit thought that it actually made a huge difference.

If only one percent of the human race survived, they would repopulate the world in only a few centuries. If the human race had learned something from its mistake, it might then continue for, let us say, a million years, the average length of time a mammalian species survives before going extinct.

Even if the human population is limited to one billion next time round, that's a trillion lives in the balance, and most of them would probably be worth living. (By the way, the climate change problem goes away instantly if you reduce the human population by 99%.) Whereas if 100% of the population dies now, all those potential future lives are also lost.

As Parfit wrote: ?Civilisation only began a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilised human history.? This perspective is sometimes called ?long-termism', and few people can manage to hold onto it for very long.

That's hardly surprising, because there has been little in our evolutionary history that really rewarded long-term thinking. We didn't even know about big threats to our survival like giant asteroid strikes, and if we had known there was nothing we could have done about them anyway.

Now we do know about them, and they have multiplied because of our own inventions, but it took another Oxford philosopher, Toby Ord, to list and rank them. It turns out that the most dangerous threats are not human hardware. They're software.

?I put the existential risk this century at around one in six: Russian roulette,? Ord says in his book ?The Precipice'.? But ?existential' actually means a threat to the existence of (in this case) the human race, and we're quite hard to kill off.

Nuclear war is not likely to do it. Even if it caused a full-scale ?nuclear winter' lasting for years and starving the overwhelming majority of the human race, a few ?breeding pairs' (in Jim Lovelock's words) would almost certainly survive.

A hothouse world or an extreme glaciation wouldn't do the trick either. The planet's climate has been through all sorts of extremes in its long history, and life survived them all. On a big planet like ours there's always some places where it's warm enough or cool enough to hang on through the extreme times.

The truly existential threats are the ones we might create ourselves, like AI that gets out of hand, or an ethno-specific engineered killer virus that mutates just a little bit. But that's software (or ?wetware'), and few people take it seriously.

As Ord points out, ?The international body responsible for the continued prohibition of bioweapons (The Biological Weapons Convention) has an annual budget of just $1.4 million ? less than the average McDonald's restaurant.? And only a few tens of millions is spent on research into AI safety, compared to many billions in general AI research.

So if we keep rolling the dice, some time in the next few centuries we're bound to get the apocalypse in one way or other. But Ord's prediction, even if it is accurate, is based on the assumption that we carry on heedlessly, and never develop the long-term perspective that would enable us to reduce the risks.

In fact, many human beings are already starting to think long-term and act accordingly ? not all of us, and much too slowly, but it is happening.

We are trying to change our entire economy to avert catastrophic climate change. We are even experimenting with ways to divert asteroids on a collision course with Earth. It's not nearly enough, but it's not bad when you consider that 500 years ago most people didn't even know that the Earth is round.