Is Existential Risk an Authentic Challenge or the Higher Moral Evasion?

Steve Fuller

An *existential risk* is a potential disaster - envisaged as occurring in the foreseeable future - that would eliminate humanity altogether.

In 2011, former UK Royal Astronomer and Royal Society head Martin Rees (author of the doom-laden *Our Final Century*) joined forces with the largesse of Skype founder Jan Tallinn to establish a Centre for the Study of Existential Risk at Cambridge University. It has attracted some of today's brightest minds, who were recently characterised in the *Guardian*, the UK's left-leaning paper of record, as "Guardians of the Galaxy."

The flagship philosopher of this venture is Nick Bostrom, who works in another privately funded Oxbridge institute, Oxford's Future of Humanity Institute. Here too the founder was an information technology entrepreneur, the late James Martin. That computer-generated wealth should be propelling the focus on existential risk is not accidental. The usual scenarios of existential risk involve advanced computers acquiring an autonomy that defy the will of their original human programmers.

In his new book *Superintelligence* , Bostrom attempts to turn this prospect into a morally neutral engineering problem by envisaging a very well-designed computer that might regard humans as disposable in its programmed quest to collect all the paper clips in the world. Our extinction turns out to be a simple by-product of the computer doing what it is supposed to do, executed entirely without malice.

Versions of Bostrom's monomaniacal "superintelligent" computer, albeit freighted with moral ballast, are familiar from science fiction, and it would be foolish to deny that our species may be seriously damaged in the future by such cases of what the American political scientist Langdon Winner has called "autonomous technology." However, Bostrom's paper clip-collecting computer makes two contestable assumptions, even granting the feats of heroic abstraction normally tolerated in philosophical thought experiments:

- that a strong "us *versus* them" distinction can be drawn between humans and machines;
- that the threat of existential risk hangs over all humans equally.

These two assumptions, no doubt designed to exorcise sentimentality from how we think about handling existential risk, amount to a serious moral evasion that makes one long for a return to the Cold War's "thinking about the unthinkable" - in Herman Kahn's resonant phrase, now more than a half-century old. Here the paradigm case of existential risk was "limited nuclear war" (between the United States and the USSR), which would entail "mutually assured destruction" - but not complete annihilation.

Although Bostrom and his fellow doomsayers present themselves as clear-headed and

analytic thinkers, they differ from Kahn and his Cold War colleagues in their apparent insensitivity to the problem of what Jean-Paul Sartre popularised as "dirty hands."

Existential risk and the problem of "dirty hands"

"Dirty hands" can be understood in a couple of ways, both of which are ultimately indebted to theodicy, the theological discipline concerned with justifying evil in the world, given the existence of a omnibenevolent deity. On the one hand, it may be seen as a "no pain, no gain" principle, whereby an ultimately good end justifies the widest possible range of means deployed in its pursuit, which may well include the destruction or radical transformation of aspects of our world that we currently value. On the other hand, it may refer to, so to speak, the "suspenseful" nature of history - namely, that on the most important value issues, we don't normally know who is right or wrong until the consequences of decisions taken have played themselves out over time.

We might think of the first interpretation of dirty hands as operating from God's point of view, outside the stream of history, and the second from humans embedded in that stream. But in both cases it is clear that no act or event is morally "neutral" in the sense of affecting all morally relevant beings equally. Thus, in contrast to Bostrom's superintelligent computer, limited nuclear war typically involved a scenario in which at least a quarter of the human population is eliminated, as well as much of our infrastructure. The serious ethical questions then revolved around whose security should we try to ensure now and how should these survivors go about picking up the pieces of civilization in the war's aftermath.

The nearest analogue to this Cold War conceptualisation of existential risk today is provided by global climate change, which interestingly is only a secondary consideration for our "Guardians of the Galaxy." Yet it is here that we face a Cold War-style moral quandary of just how many members of *Homo sapiens* need to survive for our humanity to remain intact, given that we are embarked quite knowingly on a trajectory that is reasonably likely to lead to one or more global catastrophes, all committed in the name of a certain ways of being - be it called "liberalism" or "socialism" - that themselves have universal aspirations.

In his day, Herman Kahn - the model for Stanley Kubrick's *Dr Strangelove* - was notorious for joking that one possible benefit of limited nuclear war is that the radioactive contamination of the survivors would finally impose a tolerance of physical differences ("mutations") that two hundred years of liberal democratic rhetoric had failed to achieve. In a similar spirit, Kahn 2.0 might opine that all of the endless agonising and posturing about "saving the poor" would disappear, if they are simply unable to adapt to massive climate change. Meanwhile the survivors would find their relative advantage sufficiently diminished that they are then forced to work more diligently toward a just and equitable global order.

While it would be easy to dismiss Kahn's "unthinkable" thoughts as mere provocation, nevertheless his dark sense of irony recalls Jean-Jacques Rousseau's response to the Enlightenment's own real-world example of existential risk, the great Lisbon earthquake of 1755. Unlike Voltaire who held that the disaster demonstrated divine indifference to the human condition, and equally unlike Voltaire's theological opponents who reasserted the inscrutability of divine agency, Rousseau believed that the earthquake might finally teach humans some respect for nature - the sort of thing that we might expect an environmental activist to say today.

Unlike Bostrom's superintelligent menace, the prospect of either limited nuclear war or global climate change forces clarity about the kind of world in which we would wish to live. For example, would we necessarily value all humans above all creatures, both living and artificial? On the one hand, a more ecologically sustainable post-apocalyptic world might stress symbiotic relationships among a diverse range of species rather than the overriding dominance of humanity. On the other hand, if we wish to preserve a semblance of human civilization before the apocalypse, we might preferentially allocate scarce energy resources to the maintenance of certain complex technologies, not least ones which have become integral to our identities (smartphones?), reflecting our unwitting metamorphosis into cyborgs.

Thus, *contra* Bostrom, it would be a mistake to suppose that a consensus could be easily reached among humans to pull the plug on even a very destructive yet still very useful machine.

Should existential risks be avoided?

Bostrom defines successful strategies for dealing with existential risks in terms of a "maxipok" principle - that is, one should maximize the probability of things turning out OK, even in the face of unintended consequences. There are many possible concrete applications of the principle, all of which are broadly "precautionary." For example, before making a scientific or technological innovation generally available, you must first establish its sustainability, perhaps through experimental trials or a pilot study in the "real world." The aim is to avoid irreversible damage to what we already value, however attractive future prospects may appear.

This general approach makes two countervailing assumptions about our current values: they are largely the right ones, yet their continued existence is precarious. The two assumptions are logically independent of each other but taken together they conjure an image of a world that is good enough as it already is, such that we lose it at our peril. The world's goodness is effectively tied to the specific ways in which individuals live their lives and relate to each other and the physical world more generally. It follows that disruption to these patterns of being-in-the-world places our overall value system at serious risk.

There is much to question about this line of reasoning. Nevertheless, it is worth observing that the *maxipok principle* could be embraced by both Catholics and Darwinists, who by rather different chains of causal reasoning reach largely the same conclusion - namely, that our survival as a species is dependent on our recognising and following a normative order implicitly given by nature. The difference is that for Catholics this order is permanent and underwritten by God, whereas for Darwinists it is transient and purposeless. The concept of existential risk would not appear so rhetorically compelling, were it not for this hidden convergence of world-views.

My opposing view is that even if we grant that our current values are worth promoting indefinitely, it does not follow that our default patterns of living provide the best means to pursue them. More to the point, it does not follow that a radical disruption of the human condition would necessarily undermine the realization of those values. Put in theological terms, we might take seriously that normally we exist in a fallen state: we "talk the talk" of values like equality, liberty, productivity and so on, but we only half-heartedly "walk the walk" in terms of creating a world which realizes those values. Indeed, the inclination to sustain

and ameliorate our current institutional arrangements may function as a poor proxy for a proper realization of our values.

In short, the precautionary policy of maxipok makes us too beholden to our biological and cultural inheritance, as an updated version of what Christians call "Original Sin."

Seen in this light, anything that breaks these default lines of thought and behaviour - "awakens us from our dogmatic slumbers," as Kant might say - is potentially welcomed. In *The Proactionary Imperative* , I discuss this distinctive orientation toward our fallen state - which provides an incentive to regard the *status quo* with suspicion - as emblematic of progressive "leftist" politics in the modern era. Here I mean a broad range of projects, from the great revolutionary movements to more incremental measures of intergenerational reform such in education and healthcare that were central to the foundation of the welfare state.

The bottom line is that the taking of risks - even ones that might be reasonably called "existential risks" - is not something to be avoided but embraced as potentially opening up opportunities that had been previously closed, precisely due to the previous success of the *status quo*.

In the nineteenth and twentieth centuries, the bargain was struck in terms of "the costs of progress." Indeed, the displacement and destruction of nature and people that we nowadays associate with the Industrial Revolution - still regarded as overall positive development in human history - may be understood as simply the first phase of a process whose second phase may be marked by the sorts of displacement and destruction that are now anticipated with the onset of global climate change.

Toward a more authentic sense of existential risk

Taken together, the Industrial Revolution and today's global climate change constitute what ecologists increasingly call the "anthropocene," the period when our species became the prime mover of environmental transformation. However, the two phases appear to differ in moral standing in today's world. The destructiveness of the Industrial Revolution is often conceded as a necessary price to pay for a globalized modernity, whereas global climate change is often presented as something that we should do our utmost to mitigate, if not outright prevent, because we cannot foresee the benefits that would justify the costs.

In other words, the perceived difference between the two phases lies less in the actual damage they will turn out to have inflicted - in both cases enormous - than in our capacity to construct a balance sheet that provides some agreed account of the costs and benefits. Here it is worth recalling that it was only in the 1880s that the idea of an "Industrial Revolution" began to be presented in unequivocally positive terms.

However, it would be a mistake to reduce the matter simply to our lack of 20/20 historical vision. A striking feature of today's debates over global climate change is the relative absence of serious utopian proposals comparable to the ones - including Marxist ones - that justified the undeniable costs of mass industrialisation in the nineteenth and twentieth centuries, which in turn encouraged a sense of perseverance in the face of adversity. These utopias constituted the basis for the modern imagination in science, art and politics.

To be sure, they were consistently challenged by various doomsday scenarios of

environmental degradation and human exploitation that aimed to halt and maybe even reverse industrialisation. Indeed, many - if not most - of these scenarios did come to pass and their consequences are very much with us today. Nevertheless, they do not seem as bad now as when originally presented because their significance has been offset by the benefits inspired by the more utopian sense of the future, which over time has served to reshape humanity's value orientation in its favour.

In a nutshell, then, the problem with the conception of existential risk as presented by Bostrom and other would-be "Guardians of the Galaxy" is its failure to recognize the *positive* side of risk, which is the realization that a radical improvement in the human condition may require a leap into the unknown, the short term consequences of which may be quite harmful but which in the long term issue in greater benefits that in turn serve to justify the risky undertaking.

However, this oversight may reflect a larger sense of fatalism in the general culture, one that finds it difficult to achieve the necessary distance from events to make overarching evaluations of harms and benefits. In theological terms, one may regard this fatalism as symptomatic of an incapacity for faith, which is precisely about adopting a positive attitude toward the unknown, and equally an unwillingness to entertain the divine point-of-view.

In the end is there a problem of existential risk worthy of sustained attention? The answer is most certainly *yes*, and it centres on how we might unwittingly undermine our own values in the course of their pursuit. So, while it is true that a radical change to the human condition - like the Industrial Revolution - may enable our values to be pursued more effectively, it is equally true that we may end up with false proxies for those values that we rationalize as better simply because they are part of the world that we are now stuck with. Social psychologists speak of this process as "adaptive preference formation," whereby we come to aspire to what we are likely to get.

The resulting state of mind is sometimes called "sweet lemons" - the flipside of "sour grapes. When the Existentialists struggled with the problem of "authenticity," being true to oneself, they were approaching this problem.

Contra Bostrom, it is not the problem that humanity might be annihilated by machines, but that we might become machines in the name of becoming human: the destruction of "humanity" as a concept more than the destruction of "humanity" as a population.

Steve Fuller is Auguste Comte Professor of Social Epistemology at the University of Warwick. He is the author of more than twenty books, most recently (with Veronika Lipinska) The Proactionary Imperative: A Foundation for Transhumanism. You can hear him in conversation with Joe Gelonesi on this weekend's The Philosopher's Zone on Radio National.