FONDAZIONE BRUNO KESSLER
MAGAZINE

The dark side of scientific progress: the vulnerable world hypothesis

May 23, 2023

Technological development has made us and will make us more and more powerful. Are we wise enough to use this power to make the world a better place? Or are we accelerating toward our own downfall?

The lives of human beings have been radically transformed, for the better, by scientific and technological development in recent centuries. Infant mortality has plummeted, as has extreme poverty; global average life expectancy has more than doubled since 1800; sanitation, electricity, and the Internet are accessible to more and more people.

Of course, scientific progress has also brought new problems, such as global warming and weapons of mass destruction, but all in all, its effect has been extremely positive. However, we might ask: **will it continue to be positive in the future?**

Imagine technological development as a drawing from an urn filled with balls. Each ball corresponds to an idea or invention, and its color indicates the impact of that discovery on the world. White balls are beneficial, such as the oral rehydration solution that saved millions of lives from diarrheal diseases; while gray balls are moderately harmful or double-edged weapons, such as nuclear fission, which gave us clean energy but also the Hiroshima and Nagasaki bombs.

Finally, there is a third type of ball. We are not certain of its existence, since it has never been drawn in the entire history of mankind. But if it did exist, it would be **the worst risk we take when playing the lottery of progress**: black balls. They represent technologies that, if invented, would inevitably destroy our civilization.

This metaphor was proposed by Nick Bostrom – professor of philosophy at Oxford, director of the *Future of Humanity Institute*, and author of the best-selling *Superintelligence* – in his article *The Vulnerable World Hypothesis*. According to this hypothesis, there is a **level of technology** that – if reached in the absence of extraordinary capabilities for preventive surveillance and global governance – **would make the devastation of our civilization virtually certain**. In other words, a vulnerable world is one in which the urn of possible inventions contains at least one black ball.

Bostrom distinguishes four types of “civilizational vulnerabilities,” all caused by black balls. The first involves **technology that allows small groups or individuals to cause serious damage with ease**. For example, if assembling an atomic bomb were much easier and cheaper than it (fortunately) really is, our civilization would be terribly unstable. The percentage of people who would decide, for whatever reason, to build and use such a weapon would likely be minimal; but, there being eight billion people on Earth, it would still be enough to cause a catastrophe.

Could there be such technologies in our future? **Synthetic biology** offers a plausible candidate: the costs for sequencing and synthesizing DNA is getting lower and lower, which is extremely good for research, but it could make the development of dangerous viruses easily accessible to criminals.

The second and third types of vulnerabilities, on the other hand, involve **technologies that incentivize actors, even if well-intentioned, to cause damage**. In one case, these are a few actors powerful enough to cause a catastrophe unilaterally, such as nuclear-armed countries: if some technology had allowed the U.S. or the USSR to launch a nuclear first strike safely, thereby undermining the fragile stability provided by mutual assured destruction, the Cold War probably would not have been so cold.

In the other case, it is instead many individuals whose actions have only marginally negative, but catastrophic consequences when combined. Global warming is the most glaring example: the effect of a single car trip is negligible, but the sum of all greenhouse gas emissions has and will have a terrible impact. Severe as it is, however, it is unlikely that global warming will directly end our civilization, but it is hard to rule out that some future technology could create a similar and worse situation.

Finally, the fourth type of vulnerability involves a **technology that poses an unknown or not well understood risk**, and thus leads to catastrophe even in the absence of malicious intent or bad incentives. A new technology can thus be risky if it creates unprecedented conditions. For example, the goal of several companies – including OpenAI, creator of the famous ChatGPT – is to develop an **artificial general intelligence** (AGI): an AI that can do everything a human can do, and perhaps do it better. Such technology could, for the first time in human history, **undermine our position as the most intelligent beings on the planet**, with potentially disastrous consequences for all of us. This example is flawed, however: given the economic incentives to develop an AGI before competitors and the strategic incentives to develop it before other world powers, companies and countries could decide to relax security measures and release unsafe AIs, bringing us back to the second type of vulnerability.

Bostrom then analyzes various ways we can mitigate risk and survive a black ball. Halting technological development – stop pulling balls out of the urn – is neither realistic nor desirable. We could, however, try to **slow the development of dangerous technologies and accelerate the development of protective technologies**: the order in which the balls are drawn might make a difference, if the urn also contains balls that can protect us from black ones.

But this and other methods do not seem sufficient, and potentially more effective methods are unappealing: in order to stabilize as many vulnerabilities as possible, a **combination of**

preventive surveillance and global governance may indeed be needed. The first, plausibly achieved through unprecedented mass surveillance, would prevent disastrous criminal acts. The second, perhaps in the form of a single world government, would easily solve coordination problems among countries. Bostrom does not endorse any of such measures: survival to black balls must be balanced against all the problems, even Orwellian ones, that such a regime would entail. Moreover, there may be other, as yet unexplored, ways to safeguard our future.

We do not know for sure whether the urn of human creativity contains black balls: the truth of the vulnerable world hypothesis is an open question. For Bostrom, however, **the available evidence makes it unreasonable to be sure that this hypothesis be false**. And given what is at stake, we should perhaps begin to take it seriously.

If we have so far never drawn a black ball, it is not because we have been particularly careful or wise: we have been lucky. As long as scientific and technological research continues, we will keep drawing balls from the urn, and **relying solely on Lady Luck could be fatal**. There is an urgent need for a conversation about the future of humanity, the existential risks posed by emerging technologies, and how we can mitigate those risks, since the current strategy – hoping that there are no black balls, or totally ignoring the possibility that there might be some – is irresponsible.

PERMALINK

https://magazine.fbk.eu/en/news/the-dark-side-of-scientific-progress-the-vulnerable-world-hypothesis/

TAGS

- #vulcano

AUTHORS

- Stocco Luca