The New York Times https://nyti.ms/2MpvEoJ

How to Stop Superhuman A.I. Before It Stops Us

The answer is to design artificial intelligence that's beneficial, not just smart.

By Stuart Russell
Dr. Russell is a professor of computer science at the University of California, Berkeley.

Oct. 8, 2019

The arrival of superhuman machine intelligence will be the biggest event in human history. The world's great powers are finally waking up to this fact, and the world's largest corporations have known it for some time. But what they may not fully understand is that *how* A.I. evolves will determine whether this event is also our last.

The problem is not the science-fiction plot that preoccupies Hollywood and the media — the humanoid robot that spontaneously becomes conscious and decides to hate humans. Rather, it is the creation of machines that can draw on more information and look further into the future than humans can, exceeding our capacity for decision making in the real world.

To understand how and why this could lead to serious problems, we must first go back to the basic building blocks of most A.I. systems. The "standard model" in A.I., borrowed from philosophical and economic notions of rational behavior, looks like this:

*"Machines are **intelligent** to the extent that their actions can be expected to achieve **their** objectives."*

Because machines, unlike humans, have no objectives of their own, we give them objectives to achieve. In other words, we build machines, feed objectives into them, and off they go. The more intelligent the machine, the more likely it is to complete that objective.

This model recurs throughout society, not just A.I. Control engineers design autopilots to minimize deviations from level flight; statisticians design algorithms that reduce prediction errors; retailers choose store locations that will maximize shareholder return; and governments make policy choices to accelerate G.D.P. growth.

Unfortunately, this standard model is a mistake. It makes no sense to design machines that are beneficial to us *only if* we write down our objectives completely and correctly,

because if we insert the wrong objective into the machine and it is more intelligent than us, we lose.

Until recently, we avoided the potentially serious consequences of poorly designed objectives only because our A.I. technology was not especially smart and it was mostly confined to the lab. Now, however, even the relatively simple learning algorithms for social media, which optimize clicks by manipulating human preferences, have been disastrous for democratic systems because they are so pervasive in the real world.

The effects of a superintelligent algorithm operating on a global scale could be far more severe. What if a superintelligent climate control system, given the job of restoring carbon dioxide concentrations to preindustrial levels, believes the solution is to reduce the human population to zero?

Some A.I. researchers like to claim that "we can always just switch them off" — but this makes no more sense than arguing that we can always just play better moves than the superhuman chess or Go program we're facing. The machine will anticipate all the ways in which a human might interfere and take pre-emptive steps to prevent this from happening.

The solution, then, is to change the way we think about A.I. Instead of building machines that exist to achieve *their* objectives, we want a model that looks like this:

*"Machines are **beneficial** to the extent that their actions can be expected to achieve **our** objectives."*

This fix might seem small, but it is crucial. Machines that have *our* objectives as their only guiding principle will be necessarily *uncertain* about what these objectives are, because they are in us — all eight billion of us, in all our glorious variety, and in generations yet unborn — not in the machines.

Uncertainty about objectives might sound counterproductive, but it is actually an essential feature of safe intelligent systems. It implies that no matter how intelligent they become, machines will always defer to humans. They will ask permission when appropriate, they will accept correction, and, most important, they will *allow* themselves to be switched off — precisely because they want to avoid doing whatever it is that would give humans a reason to switch them off.

Once the focus shifts from building machines that are "intelligent" to ones that are "beneficial," controlling them will become a far easier feat. Consider it the difference between nuclear power and nuclear explosions: a nuclear explosion is nuclear power in an uncontrolled form, and we greatly prefer the controlled form.

Of course, actually putting a model like this into practice requires a great deal of research. We need "minimally invasive" algorithms for decision making that prevent machines from messing with parts of the world whose value they are unsure about, as well as machines that learn more about our true, underlying preferences for how the future should unfold. Such machines will then face an age-old problem of moral philosophy: how to apportion benefits and costs among different individuals with conflicting desires.

All this could take a decade to complete — and even then, regulations will be required to ensure provably safe systems are adopted while those that don't conform are retired. This won't be easy. But it's clear that this model must be in place *before* the abilities of A.I. systems exceed those of humans in the areas that matter.

If we manage to do that, the result will be a new relationship between humans and machines, one that I hope will enable us to navigate the next few decades successfully.

If we fail, we may face a difficult choice: curtail A.I. research and forgo the enormous benefits that will flow from it, or risk losing control of our own future.

Some skeptics within the A.I. community believe they see a third option: continue with business as usual, because superintelligent machines will never arrive. But that's as if a bus driver, with all of humanity as passengers, said, "Yes, I'm driving as fast as I can toward a cliff, but trust me, we'll run out of gas before we get there!" I'd rather not take the risk.

Stuart Russell is the author of "Human Compatible: Artificial Intelligence and the Problem of Control."

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow The New York Times Opinion section on Facebook, Twitter (@NYTopinion) and Instagram.