When Machines Outsmart Humans

By: Nick Bostrom
September 10, 2014
CNN.com

Nick Bostrom

***Editor's note:** Nick Bostrom is professor and director of the Future of Humanity Institute at the Oxford Martin School at Oxford University. He is the author of "Superintelligence: Paths, Dangers, Strategies" (OUP). The opinions expressed in this commentary are solely those of the author.*

(CNN) -- Machines have surpassed humans in physical strength, speed and stamina. What would happen if machines surpassed human intellect as well? The question is not just hypothetical; we need to start taking this possibility seriously.

Most people might scoff at the prospect of machines outsmarting humanity. After all, even though today's artificial intelligence can beat humans within narrow domains (such as chess or trivia games), machine brains are still extremely rudimentary in general intelligence. Machines currently lack the flexible learning and reasoning ability that enables an average human to master any of thousands of different occupations, as well as all the tasks of daily life. In particular, while computers are useful accessories to scientists, they are very, very far from doing the interesting parts of the research themselves.

But this could change. We know that evolutionary processes can produce human-level general intelligence, because they have already done so at least once in Earth's history. How quickly engineers achieve a similar feat is still an open question. By 2050 we may, according to a recent survey of leading artificial intelligence researchers, have a 50/50 chance of achieving human-level machine intelligence (defined here as "one that can carry out most human professions at least as well as a typical human").

Even a cursory glance at technological development reveals multiple paths that could lead to human-level machine intelligence in this century. One likely path would be to continue studying the general properties of the human brain to decipher the computational structures it uses to generate intelligent behavior. Another path would be the more mathematical "top-down" approach. And if somehow all the other approaches don't work, scientists might simply brute-force the evolutionary process on computers.

Regardless of when and how we get there, the consequences of reaching human-level machine intelligence are profound, because human-level machine intelligence is not the final destination. Machine intelligence would reach a recursive tipping point after which the design and improvement of such intelligence would no longer be in human hands. The next stop from human level intelligence, just a short distance farther along the tracks, is machine superintelligence. The train might not even decelerate at Humanville Station: It is likely instead to swoosh right past.

This brings us to what I think may well be the most important task of our time. If there will eventually be an "intelligence explosion," how exactly can we set up the initial conditions so as to achieve an outcome that is survivable and beneficial to existing persons? In "Superintelligence: Paths, Dangers, Strategies," I focus on the dynamics of an intelligence explosion; what will happen if and when we gain the ability to create machine superintelligence? This topic is largely ignored and poorly funded. But we must keep at it:
How could we engineer a controlled detonation that would protect human values from being overwritten by the arbitrary values of a misbegotten artificial superintelligence?

The picture that emerges from this work is fascinating and disconcerting. It looks like there are major existential risks associated with the creation of entities of greater-than-human intelligence. A superintelligence wouldn't even need to start with a physical embodiment to be catastrophically dangerous. Major engineering projects and financial transactions on Earth are mediated by digital communication networks that would be at the mercy of an artificial superintelligence.

Placing an online order for an innocent-looking set of advanced blueprints or fooling its creators into thinking it is benign could be an initial step, followed by the possibility of permanently altering the global biosphere to pursue its preferences.

The control problem—how to engineer a superintelligence to be safe and human-friendly—appears to be very difficult. It should be solvable in principle, but in practice it may not be solved in time for when the solution is needed. The difficulty is compounded by the need to get it right on the first try. An unfriendly superintelligence would not permit a mulligan. Remember HAL from "2001: A Space Odyssey"? Let's try to avoid that.

If we could solve the technical problem of constructing a motivation system that we can load with some terminal goal of our choosing, a further question remains: Which goal would we give the superintelligent A.I.? Much would hinge on that choice. In some scenarios, the first superintelligence becomes extremely powerful and shapes the entire future according to its preferences.

We want an A.I. that is safe, beneficial and ethical, but we don't know exactly what that entails. Some may think we have already arrived upon full moral enlightenment, but is is far more likely that we still have blind spots. Our predecessors certainly had plenty -- in the practice of slavery and human sacrifice, or the condoning of manifold forms of brutality and oppression that would outrage the modern conscience. It would be a grave mistake to think we have reached our moral apogee, and thus lock our present-day ethics into such powerful machines.

In this sense, we have philosophy with a deadline. Our wisdom must precede our technology, and that which we value in life must be carefully articulated—or rather, it must be pointed to with the right mathematics—if it is to be the seed from which our intelligent creations grow.