The social impact of artificial intelligence

Technology, Artificial Intelligence

Shame Leg proved that a machine (such as your brain) cannot predict (and hush cannot control) a machine of greater algorithmic complexity, which Is a bound on a formal measure of intelligence. Informally, we cannot tell what a smarter machine will do, because if we could, we would already be that smart. As a consequence, AY is an evolutionary process: each generation experimentally creates modified versions of Itself without knowing which versions will be smarter. Friendliness is hard to define.

Do you want a smart gun that makes moral decisions about its target, or a gun that fires when you pull the trigger? Likelier S. Woodcocks reposed coherent extrapolated volition as a model of friendliness: a machine should predict what we would want If we were smarter. But this Is only a deflation. It does not say how we can program a machine to actually have the goal of granting our (projected) wishes, or how this goal can be reliably propagated through generations of recursive self improvement in the face of evolutionary pressure favoring only rapid reproduction and acquisition of computing resources.

An analogy Is helpful. Your dog does not want to get a vaccination, but It does not want to get rabies either. How does your dog know if you are acting in its best interests? Our problem is even harder. It is like asking the dog to choose an owner whose descendents will act in its best interest. But In my view, the problem Is even more fundamental. Retaining control over a superintendence would be the worst thing we could do. Modern humans do not have a lower suicide rate than humans living in medieval squalor, or even over lower animals.

In a utopian world where machines might serve our every need, answer all our questions, cure all disease, protect us from hazards, end aging ND death, and make us smarter by upgrading our brains with more computing power, would we be happier? If the brain is a machine that can be simulated on a computer, then it could also be reprogrammed. When a rat can electrically stimulate certain parts of its brain by pressing a lever, it will forgo food, water, and sleep until it dies. Do you really want a future where you can have everything you want, Including the ability to change what you want? Plodding If you could not control a godlike intelligence, then could you become one? My view is that the outcome is the same either way. Suppose a computer simulated your brain, with all the same memories and goals, so that everyone who talked to It, Including you, was convinced that It was you. Would such a machine be conscious? Would It be you? If you died, would your consciousness live on through this machine? If you 1 OFF atom, and then the original you was killed, would the copy be you, or would it be a philosophical zombie?

If the two copies are identical, does it matter which one dies? Chalmers' fading quail argument asks if the neurons in your brain were replaced one by one with artificial but functionally equivalent devices, at what point do you come a zombie? These questions are hard to answer because it exposes a conflict between logic, which says that there is no physical basis for consciousness, and your brain's hardwired belief that consciousness exists, that there is a " you" inside your brain that experiences your perceptions, feels pleasure and pain, and controls your thoughts and actions.

But like the P-zombie, there is no test to tell if a human, animal, or machine actually feels pleasure and pain, or only behaves as if it does. The autobahns model of a logic gate that responds to reinforcement learning is clearly not unconscious, but does not differ fundamentally from the way that humans respond to pleasure and pain. Any good optimization process with the goal of maximizing an accumulated reward signal will trade off short term exploitation against long term exploration, a characteristic that gives the appearance of quail and free will.

Animals that behaved differently were eliminated by natural selection. The Turing Test for AY sidesteps the question of whether machines can think. By this widely accepted test, a machine is intelligent if a human communicating with it via teletype cannot distinguish it from another human. I propose a modified Turing test for uploading: if the Judges are people that know you well (friends, relatives, co-workers, etc) and they cannot distinguish between you and the machine, then that machine passes the upload test.

For the upload test, you are allowed to collaborate with the judges in advance, for example, to agree on a secret password. Clearly, to pass the upload test, a machine must both be intelligent (pass the Turing test) and know everything that you know. One possible outcome of an intelligence explosion is that the entire Earth is turned into a computer (perhaps a Tyson sphere of grey go) with en bit of memory in each of its 1051 atoms. Suppose that this computer contained atomic-level maps of the brains of all of the 1011 humans who ever lived, at 1027 atoms each.

This would represent 10-13 of its total memory. This is larger than the 109 intelligence gap between the human brain (101 5 synapses) and the e. Coli genome (106 base pairs). To say that this machine would be you would be like saying that you are the self replicating RNA molecules from which you evolved over the last 3 billion years. But the most troubling aspect is this: If your memories were changed, how would you know? How do you know that you were not born one second ago, with your lifetime of memories Just now written into your brain?

Even if your memories made up a significant fraction of the godlike AY, would it matter if they were replaced with something else? Logically I know that consciousness does not exist, but I cannot refute the belief encoded in my DNA. I am aware that my beliefs are inconsistent, and leave it at that. When Will the Singularity Occur? Alan Turing predicted in 1950 that a computer with 109 bits of memory, but no faster than current technology, would achieve AY in 2000. Vernon Vine predicted in 1993 that the singularity will occur in about 30 years, but most certainly between 2005 and 2030.

Ray Skuzzier studied dozens of trends in technology advancement, and forecasts about the middle of this century. Roger Penrose believes that AY is not human consciousness (although his views are not widely accepted). If we knew how much computing power was required for AY, we could forecast it using Moor's law. Unfortunately there is a gap of 106 between cognitive and neurological models. Cognitive models believed to be sufficient to pass the Turing test are based on tests f long term memory recall, which Lander [1] estimates to be 109 bits.

Also, humans read and hear about 109 bits of language (as compressed text) in a lifetime. The human brain has about 1011 neurons and 101 5 synapses. Most artificial neural networks model learning using Web's rule, where the relevant signal is the rate of firing (O to 300 Hz's), not the individual pulses (with some exceptions, such as signaling phase information for stereoscopic sound perception). In the Hayfield model [2], each synapse stores 0. 15 bits of information, although a computer model may require overall bits per synapse.

The Blue Brain project in early 2007 simulated half of a mouse cortex with 8, 000, 000 neurons (1112, 500 of a human brain) and 6300 synapses per neuron at 1 ms resolution, running at 1/10 real time speed using 4096 processors and 1 TAB memory on the Bludgeon/L, the world's fastest supercomputer as of Novo. 2007. My work in text compression, which is equivalent to language learning, attempts to narrow this gap by comparing the computational requirements of many models.

This work is far from complete, but early results suggest that 2 KGB of memory s far from adequate for even a low level semantic language model, and totally inadequate for learning grammar. If AY requires 1015 bits of memory and 1016 operations per second, and Moor's law continues to double computing power every 1. 5 or 2 years (as it has since the sass's), then we should expect the Internet to exceed the computing power of the human population in the sass's. Has a Singularity Already Occurred? If your brain were placed in a bottle and hooked to a computer that simulated the outside world, how would you know?

You could argue that technology has not yet advanced to the point where this is possible. That is true in the universe you observe, but we know nothing about the universe where the simulation is taking place. For all we know, space, time, and matter may be meaningless abstract concepts in this universe. Marcus Hutted showed that the optimal behavior of an interactive reward- seeking agent is to guess at each step that the environment is simulated by the shortest program consistent with past observation. Hatter's XIX model is a formal statement of Scam's Razor: the simplest solution is best.

The fact that Scam's Razor works in real life suggests (but does not prove) that the universe is a simulation. Another fact that suggests that the universe could be simulated is that it has finite entropy, or information content. According to quantum mechanics, a closed system with energy E (or mass m = E/co) and diameter d has on the order of expo(De/ hoc) possible quantum states, where h is Plank's constant and c is the speed of light. For the universe under the Big Bang model it would take about 10122 (and growing) bits of information to describe this state, which would be an exact model of the inverse. By a curious coincidence, if you divided the universe into this many regions, each one would be about the size of a proton or neutron, even though the number does not depend on the properties of any particles). Although we cannot observe the universe which simulates ours, XIX tells us which models are more likely. Therefore from least to most likely. Model Computation steps required Complexity (bits) Probability Single brain: you are the only conscious being in the universe. Other people are

simulated. 1025 (neural model x 109 seconds) 109 (cognitive model) Least likely

Single universe: simulated Big Bang at the level of quantum physics. 10122 (entropy of the universe) 102 (complexity of the free parameters in string theory) Multiversity: enumeration of Turing machines until a universe supporting life is found. 10200 to 10300 (trying simpler universes first) 100 (complexity of N by Wolfram's principle of computational equivalence) Most likely What Should We Do? That is the wrong question. The question is what will we do? AY could potentially automate all human labor, valued at US $66 trillion Just in 2006. Our biologically programmed fear of death will also push us to upload.

We will proceed. AY will be the third paradigm shift in evolution. The first was RNA/DNA based life 3 billion years ago. The second was the rise of language and culture about 10, 000 years ago. I do not believe the Singularity will be an apocalypse. It will be invisible; a barrier you cannot look beyond from either side. A godlike intelligence could no more make its presence known to you than you could make your presence known to the bacteria in your gut. Asking what we should do would be like bacteria asking how they can evolve into humans who won't use antibiotics.