How Long Before Superintelligence?

Nick Bostrom
Oxford University
www.nickbostrom.com

(1997) Copyright. Revised 25 October, 1998, and a postscript added. Second postscript added 28 August, 2000. Third postscript added 30 October, 2005. Fourth postscript added 12 March, 2008. Originally published in *Int. Jour. of Futures Studies*, 1998, vol. 2. Reprinted in *Linguistic and Philosophical Investigations*, 2006, Vol. 5, No. 1, pp. 11-30.

Formats: HTML, PDF

Abstract

This paper outlines the case for believing that we will have superhuman artificial intelligence within the first third of the next century. It looks at different estimates of the processing power of the human brain; how long it will take until computer hardware achieves a similar performance; ways of creating the software through bottom-up approaches like the one used by biological brains; how difficult it will be for neuroscience to figure out enough about how brains work to make this approach work; and how fast we can expect superintelligence to be developed once there is human-level artificial intelligence.

Definition of "superintelligence"

By a "superintelligence" we mean an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills. This definition leaves open how the superintelligence is implemented: it could be a digital computer, an ensemble of networked computers, cultured cortical tissue or what have you. It also leaves open whether the superintelligence is conscious and has subjective experiences.

Entities such as companies or the scientific community are *not* superintelligences according to this definition. Although they can perform a number of tasks of which no individual human is capable, they are not intellects and there are many

fields in which they perform much worse than a human brain – for example, you can't have a real-time conversation with "the scientific community".

Moore's law and present supercomputers

Moore's law [A] states that processor speed doubles every eighteen months. The doubling time used to be two years, but that changed about fifteen years ago. The most recent data points indicate a doubling time as short as twelve months. This would mean that there will be a thousand-fold increase in computational power in ten years. Moore's law is what chip manufacturers rely on when they decide what sort of chip to develop in order to remain competitive.

If we estimate the computational capacity of the human brain, and allow ourselves to extrapolate available processor speed according to Moore's law (whether doing so is permissible will be discussed shortly), we can calculate how long it will take before computers have sufficient raw power to match a human intellect.

The fastest supercomputer today (December 1997) is 1.5 teraops, 1.5×10^{12} ops. There is a project that aims to extract 10 teraops from the Internet by having a hundred thousand volunteers install a screen saver on their computers that would allow a central computer to delegate some computational tasks to them. This (so-called *metacomputing*) approach works best for tasks that are very easy to parallelize, such as doing an exhaustive journey through search space in attempting to break a code. With better bandwidth connections in the future (e.g. optical fibers), large-scale metacomputing will work even better than today. Brain simulations should by their nature be relatively easy to parallelize, so maybe huge brain simulations distributed over the Internet could be a feasible alternative in the future. We shall however disregard this possibility for present purposes and regard the 1.5 Tops machine as the best we can do today. The potential of metacomputing can be factored into our prognosis by viewing it as an additional reason to believe that available computing power will continue to grow as Moore's law predicts.

Even without any technology improvement we can do somewhat better, for example by doubling the number of chips that we put in the box. A 3 Tops computer has been ordered by the US government to be used in testing and developing the nation's stockpile of nuclear weapons. However, considering that the cost of this machine is $94,000,000, it is clear that even massive extra funding would only yield a very modest increase in computing power in the short term.

How good grounds are there to believe that Moore's law will continue to hold in the future? It is clear that sooner or later it must fail. There are physical limitations on the density with which matter can store and process information. The Bekenstein bound gives an upper limit on the amount of information that can be contained within any given volume using a given amount of energy. Since space colonization would allow at most a polynomial ($\sim t^3$) expansion rate

(assuming expansion rate is bounded by the speed of light), the exponential increase of available computational power cannot be continued indefinitely, unless new physics is forthcoming.

In my opinion, Moore's law loses its credibility long before we reach absolute physical limits. It probably hasn't got much predictive power beyond, say, the next fifteen years. That is not to say that processor speed will not continue to double every twelve or eighteen months after 2012; only that we cannot use Moore's law to argue that it will. Instead, if we want to make predictions beyond that date, we will have to look directly at what is physically feasible. That will presumably also mean that we have to content ourselves with a greater uncertainty interval along the time axis. Physical feasibility studies tell us, at best, what will happen given that people want it to happen; but even if we assume that the demand is there, it will still not tell us *when* it will happen.

In about the year 2007 we will have reached the physical limit of present silicon technology. Moore's law, however, has survived several technological phase transitions before, from relays to vacuum tubes to transistors to integrated circuits to Very Large Scale Integrated circuits (VLSI). There is no reason why present VLSI designs on two-dimensional silicon wafers should be the last word in chip technology. Several ways to overcome the limits of the present technology have been proposed and are being developed.

In the near future, it might for example be possible to use phase shift masks to push the minimum circuit-line width on a microchip down to as little as 0.13 micrometer, even while remaining in the optical range with the lithographic irradiation. Leaving the optical range, we could use x-rays or at least extreme ultraviolet ("EUV", also called "soft x-rays") to attain still finer precision. Failing this, it should be feasible to use electron beam writing, although this production method would be slow and hence expensive. A compromise would be to write some of the gates with an electron beam, especially at bottlenecks where speed is absolutely crucial, and use optical or EUV to write the other elements of the chip.

We can also increase the power of a chip by using more layers, a technique that has only recently been mastered, and by making bigger wafers (up to 300 mm should not be a problem). Drastically bigger chips could be manufactured if there were some error tolerance. Tolerance to error could be obtained by using evolvable hardware (de Garis 1997).

It is also possible to push the physical limits on how small the transistors can be made by switching to new materials, such as gallium arsenide. Quantum transistors are presently being developed, promising a major step forward for circuitry where high switching speed or low energy consumption is essential.

Because of the highly parallel nature of brain-like computations, it should also be possible to use a highly parallel architecture, in which case it will suffice to produce a great number of moderately fast processors, and then have them connected. You could either put them in the same box which would give you a

bus-based multiprocessor (which are quite popular today) or you could link them up to a high-bandwidth local-area network (an option that will be increasingly attractive as the performance of standard networking technology improves)[B].

These are all things that are being developed today. Massive funding is pumped into these technologies[C]. Although the difficulties can appear staggering to a person working in the field, who is constantly focused on the immediate problems, it is fair to say that there is widespread optimism among the experts that the prospects are good that computers will continue to grow more powerful for the foreseeable future.

Notes

A) It is not clear what, exactly, Moore's law says. The law derives its name from Gordon Moore, co-founder of Intel Corp., who back in 1965 noted that microchips were doubling in circuit density every year. In 1975 he made the prediction that from then on, the doubling time would be two years. The actual doubling time has fluctuated a bit, starting at one year, going up to two years, and is now back to approximately one year again. So one ambiguity in citing Moore's law is that it is unclear whether the time constant is supposed to be one year, two years, or whether it is supposed to be whatever the most recent data points indicate. A second ambiguity resides in the fact that the initial statement was phrased in terms of the number of transistors that could be fitted into an area unit, rather than in terms of the speed of the resulting chip. Until now, this distinction hasn't mattered much, because circuitry density and speed have been highly correlated. When we look to the future, however, it is possible that we will achieve increased computing power by other means than by making transistors smaller. It therefore makes sense to reformulate Moore's law into a statement asserting an exponential growth in *computing power* (per inflation-adjusted dollar) rather than chip density. It is better to apply the label "Moore's law" to this slightly modified hypothesis than to invent a new term for what is basically the same idea.

B) In the longer term, we also have to consider molecular nanotechnology and maybe quantum computing.

C) It nowadays takes about 400 engineers to produce a new chip. A modern chip factory may cost over $2 billion. About $20 to $30 billion is spent on microchip R&D every year. These figures have grown over the years, so it should be pointed out that one factor that could slow the pace of development would be if funding begins to level out, as it sooner or later will.

Hardware requirements

The human brain contains about 10^{11} neurons. Each neuron has about 5×10^3 synapses, and signals are transmitted along these synapses at an average

frequency of about 10^2 Hz. Each signal contains, say, 5 bits. This equals 10^{17} ops.[A]

The true value cannot be much higher than this, but it might be much lower. There seems to be great redundancy in the brain; synchronous firing of large pools of neurons is often required if the signal is not to drown in the general noise. An alternative way of calculating the total capacity is to consider some part of the cortex that performs a function that we know how to replicate on digital computers. We calculate the average computer-equivalent processing capacity of a single neuron in that cortical area, and multiply this value with the number of neurons in the brain. Hans Moravec has done this calculation using data about the human retina (Moravec 1997) and compared it with known computational demands of edge extraction in robot vision. He got the value 10^{14} ops for the human brain as a whole. That is three orders of magnitude less than the upper bound calculated by assuming that there is no redundancy.

It is hard to see any reason to suppose that the redundancy in the retina should be greater than in the cortex. If anything, one would rather expect it to be the other way around, since edge extraction is a more low-level task than higher cognitive processes and therefore presumably more optimized (by evolution and individual learning).

If we need 100 Tops to simulate the human brain then the required computational power will be reached sometime between 2004 and 2008, depending on whether we assume a doubling time of 12 or 18 months. This would be the best experimental supercomputers in the world, not necessarily the computers available to AI researchers. Depending on how much funding is forthcoming, it might take up to an additional decade before researchers experimenting with general artificial intelligence have access to machines with this capacity.

This is if we take the retina simulation as a model. At the present, however, not enough is known about the neocortex to allow us to simulate it in such an optimized way. But the knowledge might be available by 2004 to 2008 (as we shall see in the next section). What is required, if we are to get human-level AI with hardware power at this lower bound, is the ability to simulate 1000-neuron aggregates in a highly efficient way.

The extreme alternative, which is what we assumed in the derivation of the upper bound, is to simulate each neuron individually. The number of clock cycles that neuroscientists can expend simulating the processes of a single neuron knows of no limits, but that is because their aim is to model the detailed chemical and electrodynamic processes in the nerve cell rather than to just do the minimal amount of computation necessary to replicate those features of its response function which are relevant for the total performance of the neural net. It is not known how much of the detail that is contingent and inessential and how much needs to be preserved in order for the simulation to replicate the performance of the whole. It seems like a good bet though, at least to the author, that the nodes

could be strongly simplified and replaced with simple standardized elements. It appears perfectly feasible to have an intelligent neural network with any of a large variety of neuronal output functions and time delays.

It does look plausible, however, that by the time when we know how to simulate an idealized neuron and know enough about the brain's synaptic structure that we can put the artificial neurons together in a way that functionally mirrors how it is done in the brain, then we will also be able to replace whole 1000-neuron modules with something that requires less computational power to simulate than it does to simulate all the neurons in the module individually. We might well get all the way down to a mere 1000 instructions per neuron and second, as is implied by Moravec's estimate (10^{14} ops / 10^{11} neurons = 1000 operations per second and neuron). But unless we can build these modules without first building a whole brain then this optimization will only be possible *after* we have already developed human-equivalent artificial intelligence.

If we assume the upper bound on the computational power needed to simulate the human brain, i.e. if we assume enough power to simulate each neuron individually (10^{17} ops), then Moore's law says that we will have to wait until about 2015 or 2024 (for doubling times of 12 and 18 months, respectively) before supercomputers with the requisite performance are at hand. But if by then we know how to do the simulation on the level of individual neurons, we will presumably also have figured out how to make at least some optimizations, so we could probably adjust these upper bounds a bit downwards.

So far I have been talking only of processor speed, but computers need a great deal of memory too if they are to replicate the brain's performance. Throughout the history of computers, the ratio between memory and speed has remained more or less constant at about 1 byte/ops. Since a signal is transmitted along a synapse, on average, with a frequency of about 100 Hz and since its memory capacity is probably less than 100 bytes (1 byte looks like a more reasonable estimate), it seems that speed rather than memory would be the bottleneck in brain simulations on the neuronal level. (If we instead assume that we can achieve a thousand-fold leverage in our simulation speed as assumed in Moravec's estimate, then that would bring the requirement of speed down, perhaps, one order of magnitude below the memory requirement. But if we can optimize away three orders of magnitude on speed by simulating 1000-neuron aggregates, we will probably be able to cut away at least one order of magnitude of the memory requirement. Thus the difficulty of building enough memory may be significantly smaller, and is almost certainly not significantly greater, than the difficulty of building a processor that is fast enough. We can therefore focus on speed as the critical parameter on the hardware front.)

This paper does not discuss the possibility that quantum phenomena are irreducibly involved in human cognition. Hameroff and Penrose and others have suggested that coherent quantum states may exist in the microtubules, and that the brain utilizes these phenomena to perform high-level cognitive feats. The author's opinion is that this is implausible. The controversy

surrounding this issue won't be entered into here; it will simply be assumed, throughout this paper, that quantum phenomena are not functionally relevant to high-level brain modelling.

In conclusion we can say that the hardware capacity for human-equivalent artificial intelligence will likely exist before the end of the first quarter of the next century, and may be reached as early as 2004. A corresponding capacity should be available to leading AI labs within ten years thereafter (or sooner if the potential of human-level AI and superintelligence is by then better appreciated by funding agencies).

Notes

A) It is possible to nit-pick on this estimate. For example, there is some evidence that some limited amount of communication between nerve cells is possible without synaptic transmission. And we have the regulatory mechanisms consisting of neurotransmitters and their sources, receptors and re-uptake channels. While neurotransmitter balances are crucially important for the proper functioning of the human brain, they have an insignificant information content compared to the synaptic structure. Perhaps a more serious point is that neurons often have rather complex time-integration properties (Koch 1997). Whether a specific set of synaptic inputs result in the firing of a neuron depends on their exact timing. The author's opinion is that except possibly for a small number of special applications such as auditory stereo perception, the temporal properties of the neurons can easily be accommodated with a time resolution of the simulation on the order of 1 ms. In an unoptimized simulation this would add an order of magnitude to the estimate given above, where we assumed a temporal resolution of 10 ms, corresponding to an average firing rate of 100 Hz. However, the other values on which the estimate was based appear to be too high rather than too low, so we should not change the estimate much to allow for possible fine-grained time-integration effects in a neuron's dendritic tree. (Note that even if we were to adjust our estimate upward by an order of magnitude, this would merely add three to five years to the predicted upper bound on when human-equivalent hardware arrives. The lower bound, which is based on Moravec's estimate, would remain unchanged.)

Software via the bottom-up approach

Superintelligence requires software as well as hardware. There are several approaches to the software problem, varying in the amount of top-down direction they require. At the one extreme we have systems like CYC which is a very large encyclopedia-like knowledge-base and inference-engine. It has been spoon-fed facts, rules of thumb and heuristics for over a decade by a team of human knowledge enterers. While systems like CYC might be good for certain practical tasks, this hardly seems like an approach that will convince

AI-skeptics that superintelligence might well happen in the foreseeable future. We have to look at paradigms that require less human input, ones that make more use of bottom-up methods.

Given sufficient hardware and the right sort of programming, we could make the machines learn in the same way a child does, i.e. by interacting with human adults and other objects in the environment. The learning mechanisms used by the brain are currently not completely understood. Artificial neural networks in real-world applications today are usually trained through some variant of the Backpropagation algorithm (which is known to be biologically unrealistic). The Backpropagation algorithm works fine for smallish networks (of up to a few thousand neurons) but it doesn't scale well. The time it takes to train a network tends to increase dramatically with the number of neurons it contains. Another limitation of backpropagation is that it is a form of supervised learning, requiring that signed error terms for each output neuron are specified during learning. It's not clear how such detailed performance feedback on the level of individual neurons could be provided in real-world situations except for certain well-defined specialized tasks.

A biologically more realistic learning mode is the Hebbian algorithm. Hebbian learning is unsupervised and it might also have better scaling properties than Backpropagation. However, it has yet to be explained how Hebbian learning by itself could produce all the forms of learning and adaptation of which the human brain is capable (such as the storage of structured representation in long-term memory – Bostrom 1996). Presumably, Hebb's rule would at least need to be supplemented with reward-induced learning (Morillo 1992) and maybe with other learning modes that are yet to be discovered. It does seem plausible, though, to assume that only a very limited set of different learning rules (maybe as few as two or three) are operating in the human brain. And we are not very far from knowing what these rules are.

Creating superintelligence through imitating the functioning of the human brain requires two more things in addition to appropriate learning rules (and sufficiently powerful hardware): it requires having an adequate initial architecture and providing a rich flux of sensory input.

The latter prerequisite is easily provided even with present technology. Using video cameras, microphones and tactile sensors, it is possible to ensure a steady flow of real-world information to the artificial neural network. An interactive element could be arranged by connecting the system to robot limbs and a speaker.

Developing an adequate initial network structure is a more serious problem. It might turn out to be necessary to do a considerable amount of hand-coding in order to get the cortical architecture right. In biological organisms, the brain does not start out at birth as a homogenous *tabula rasa*; it has an initial structure that is coded genetically. Neuroscience cannot, at its present stage, say exactly what this structure is or how much of it needs to be preserved in a

simulation that is eventually to match the cognitive competencies of a human adult. One way for it to be unexpectedly difficult to achieve human-level AI through the neural network approach would be if it turned out that the human brain relies on a colossal amount of genetic hardwiring, so that each cognitive function depends on a unique and hopelessly complicated inborn architecture, acquired over aeons in the evolutionary learning process of our species.

Is this the case? A number of considerations suggest otherwise. We have to content ourselves with a very brief review here. For a more comprehensive discussion, the reader may consult Phillips & Singer (1997).

Quartz & Sejnowski (1997) argue from recent neurobiological data that the developing human cortex is largely free of domain-specific structures. The representational properties of the specialized circuits that we find in the mature cortex are not generally genetically prespecified. Rather, they are developed through interaction with the problem domains on which the circuits operate. There are genetically coded *tendencies* for certain brain areas to specialize on certain tasks (for example primary visual processing is usually performed in the primary visual cortex) but this does not mean that other cortical areas couldn't have learnt to perform the same function. In fact, the human neocortex seems to start out as a fairly flexible and general-purpose mechanism; specific modules arise later through self-organizing and through interacting with the environment.

Strongly supporting this view is the fact that cortical lesions, even sizeable ones, can often be compensated for if they occur at an early age. Other cortical areas take over the functions that would normally have been developed in the destroyed region. In one study, sensitivity to visual features was developed in the auditory cortex of neonatal ferrets, after that region's normal auditory input channel had been replaced by visual projections (Sur et al. 1988). Similarly, it has been shown that the visual cortex can take over functions normally performed by the somatosensory cortex (Schlaggar & O'Leary 1991). A recent experiment (Cohen et al. 1997) showed that people who have been blind from an early age can use their visual cortex to process tactile stimulation when reading Braille.

There are some more primitive regions of the brain whose functions cannot be taken over by any other area. For example, people who have their hippocampus removed, lose their ability to learn new episodic or semantic facts. But the neocortex tends to be highly plastic and that is where most of the high-level processing is executed that makes us intellectually superior to other animals. (It would be interesting to examine in more detail to what extent this holds true for all of neocortex. Are there small neocortical regions such that, if excised at birth, the subject will never obtain certain high-level competencies, not even to a limited degree?)

Another consideration that seems to indicate that innate architectural differentiation plays a relatively small part in accounting for the performance

of the mature brain is that the neocortical architecture, especially in infants, is remarkably homogeneous over different cortical regions and even over different species:

> Laminations and vertical connections between lamina are hallmarks of all cortical systems, the morphological and physiological characteristics of cortical neurons are equivalent in different species, as are the kinds of synaptic interactions involving cortical neurons. This similarity in the organization of the cerebral cortex extends even to the specific details of cortical circuitry. (White 1989, p. 179).

One might object at this point that cetaceans have much bigger cortices than humans and yet they don't have human-level abstract understanding and language[A]. A large cortex, apparently, is not sufficient for human intelligence. However, one can easily imagine that some very simple difference between human and cetacean brains can account for why we have abstract language and understanding that they lack. It could be something as trivial as that our cortex is provided with a low-level "drive" to learn about abstract relationships whereas dolphins and whales are programmed not to care about or pay much attention to such things (which might be totally irrelevant to them in their natural environment). More likely, there are some structural developments in the human cortex that other animals lack and that are necessary for advanced abstract thinking. But these uniquely human developments may well be the result of relatively simple changes in just a few basic parameters. They do not require a large amount of genetic hardwiring. Indeed, given that brain evolution that allowed *Homo sapiens* to intellectually outclass other animals took place over a relatively brief period of time, evolution cannot have embedded very much content-specific information in these additional cortical structures that give us our intellectual edge over our humanoid or ape-like ancestors.

These considerations (especially the one of cortical plasticity) suggest that the amount of neuroscientific information needed for the bottom-up approach to succeed may be very limited. (Notice that they do not argue against the modularization of adult human brains. They only indicate that the greatest part of the information that goes into the modularization results from self-organization and perceptual input rather than from an immensely complicated genetic look-up table.)

Further advances in neuroscience are probably needed before we can construct a human-level (or even higher animal-level) artificial intelligence by means of this radically bottom-up approach. While it is true that neuroscience has advanced very rapidly in recent years, it is difficult to estimate how long it will take before enough is known about the brain's neuronal architecture and its learning algorithms to make it possible to replicate these in a computer of sufficient computational power. A wild guess: something like fifteen years. This is not a prediction about how far we are from a complete understanding of all important phenomena in the brain. The estimate refers to the time when we might be

expected to know enough about the basic principles of how the brain works to be able to implement these computational paradigms on a computer, without necessarily modelling the brain in any biologically realistic way.

The estimate might seem to some to underestimate the difficulties, and perhaps it does. But consider how much has happened in the past fifteen years. The discipline of computational neuroscience did hardly even exist back in 1982. And future progress will occur not only because research with today's instrumentation will continue to produce illuminating findings, but also because new experimental tools and techniques become available. Large-scale multi-electrode recordings should be feasible within the near future. Neuro/chip interfaces are in development. More powerful hardware is being made available to neuroscientists to do computation-intensive simulations. Neuropharmacologists design drugs with higher specificity, allowing researchers to selectively target given receptor subtypes. Present scanning techniques are improved and new ones are under development. The list could be continued. All these innovations will give neuroscientists very powerful new tools that will facilitate their research.

This section has discussed the software problem. It was argued that it can be solved through a bottom-up approach by using present equipment to supply the input and output channels, and by continuing to study the human brain in order to find out about what learning algorithm it uses and about the initial neuronal structure in new-born infants. Considering how large strides computational neuroscience has taken in the last decade, and the new experimental instrumentation that is under development, it seems reasonable to suppose that the required neuroscientific knowledge might be obtained in perhaps fifteen years from now, i.e. by year 2012.

Notes

A) That dolphins don't have abstract language was recently established in a very elegant experiment. A pool is divided into two halves by a net. Dolphin A is released into one end of the pool where there is a mechanism. After a while, the dolphin figures out how to operate the mechanism which causes dead fish to be released into both ends of the pool. Then A is transferred to the other end of the pool and a dolphin B is released into the end of the pool that has the mechanism. The idea is that if the dolphins had a language, then A would tell B to operate the mechanism. However, it was found that the average time for B to operate the mechanism was the same as for A.

Why the past failure of AI is no argument against its future success

In the seventies and eighties the AI field suffered some stagnation as the exaggerated expectations from the early heydays failed to materialize and

progress nearly ground to a halt. The lesson to draw from this episode is not that strong AI is dead and that superintelligent machines will never be built. It shows that AI is more difficult than some of the early pioneers might have thought, but it goes no way towards showing that AI will forever remain unfeasible.

In retrospect we know that the AI project couldn't possibly have succeeded at that stage. The hardware was simply not powerful enough. It seems that at least about 100 Tops is required for human-like performance, and possibly as much as 10^{17} ops is needed. The computers in the seventies had a computing power comparable to that of insects. They also achieved approximately insect-level intelligence. Now, on the other hand, we can foresee the arrival of human-equivalent hardware, so the cause of AI's past failure will then no longer be present.

There is also an explanation for the relative absence even of noticeable progress during this period. As Hans Moravec points out:

> [F]or several decades the computing power found in advanced Artificial Intelligence and Robotics systems has been stuck at insect brain power of 1 MIPS. While computer power per dollar fell [should be: rose] rapidly during this period, the money available fell just as fast. The earliest days of AI, in the mid 1960s, were fuelled by lavish post-Sputnik defence funding, which gave access to $10,000,000 supercomputers of the time. In the post Vietnam war days of the 1970s, funding declined and only $1,000,000 machines were available. By the early 1980s, AI research had to settle for $100,000 minicomputers. In the late 1980s, the available machines were $10,000 workstations. By the 1990s, much work was done on personal computers costing only a few thousand dollars. Since then AI and robot brain power has risen with improvements in computer efficiency. By 1993 personal computers provided 10 MIPS, by 1995 it was 30 MIPS, and in 1997 it is over 100 MIPS. Suddenly machines are reading text, recognizing speech, and robots are driving themselves cross country. (Moravec 1997)

In general, there seems to be a new-found sense of optimism and excitement among people working in AI, especially among those taking a bottom-up approach, such as researchers in genetic algorithms, neuromorphic engineering and in hardware implementations of neural networks. Many experts who have been around, though, are wary not again to underestimate the difficulties ahead.

Once there is human-level AI there will soon be superintelligence

Once artificial intelligence reaches human level, there will be a positive feedback loop that will give the development a further boost. AIs would help constructing better AIs, which in turn would help building better AIs, and so forth.

Even if no further software development took place and the AIs did not accumulate new skills through self-learning, the AIs would still get smarter if processor speed continued to increase. If after 18 months the hardware were upgraded to double the speed, we would have an AI that could think twice as fast as its original implementation. After a few more doublings this would directly lead to what has been called "weak superintelligence", i.e. an intellect that has about the same abilities as a human brain but is much faster.

Also, the marginal utility of improvements in AI when AI reaches human-level would also seem to skyrocket, causing funding to increase. We can therefore make the prediction that once there is human-level artificial intelligence then it will not be long before superintelligence is technologically feasible.

A further point can be made in support of this prediction. In contrast to what's possible for biological intellects, it might be possible to copy skills or cognitive modules from one artificial intellect to another. If one AI has achieved eminence in some field, then subsequent AIs can upload the pioneer's program or synaptic weight-matrix and immediately achieve the same level of performance. It would not be necessary to again go through the training process. Whether it will also be possible to copy the best parts of several AIs and combine them into one will depend on details of implementation and the degree to which the AIs are modularized in a standardized fashion. But as a general rule, the intellectual achievements of artificial intellects are additive in a way that human achievements are not, or only to a much lesser degree.

The demand for superintelligence

Given that superintelligence will one day be technologically feasible, will people choose to develop it? This question can pretty confidently be answered in the affirmative. Associated with every step along the road to superintelligence are enormous economic payoffs. The computer industry invests huge sums in the next generation of hardware and software, and it will continue doing so as long as there is a competitive pressure and profits to be made. People want better computers and smarter software, and they want the benefits these machines can help produce. Better medical drugs; relief for humans from the need to perform boring or dangerous jobs; entertainment – there is no end to the list of consumer benefits. There is also a strong military motive to develop artificial intelligence. And nowhere on the path is there any natural stopping point where technophobes could plausibly argue "hither but not further".

It therefore seems that up to human equivalence, the driving forces behind improvements in AI will easily overpower whatever resistance might be present. When the question is about human-level or greater intelligence then it is conceivable that there might be strong political forces opposing further development. Superintelligence might be seen to pose a threat to the supremacy, and even to the survival, of the human species. Whether by suitable programming we can arrange the motivation systems of the superintelligences

in such a way as to guarantee perpetual obedience and subservience, or at least non-harmfulness, to humans is a contentious topic. If future policy-makers can be sure that AIs would not endanger human interests then the development of artificial intelligence will continue. If they can't be sure that there would be no danger, then the development might well continue anyway, either because people don't regard the gradual displacement of biological humans with machines as necessarily a bad outcome, or because such strong forces (motivated by short-term profit, curiosity, ideology, or desire for the capabilities that superintelligences might bring to their creators) are active that a collective decision to ban new research in this field cannot be reached and successfully implemented.

Conclusion

Depending on degree of optimization assumed, human-level intelligence probably requires between 10^{14} and 10^{17} ops. It seems quite possible that very advanced optimization could reduce this figure further, but the *entrance level* would probably not be less than about 10^{14} ops. If Moore's law continues to hold then the lower bound will be reached sometime between 2004 and 2008, and the upper bound between 2015 and 2024. The past success of Moore's law gives some inductive reason to believe that it will hold another ten, fifteen years or so; and this prediction is supported by the fact that there are many promising new technologies currently under development which hold great potential to increase procurable computing power. There is no direct reason to suppose that Moore's law will not hold longer than 15 years. It thus seems likely that the requisite hardware for human-level artificial intelligence will be assembled in the first quarter of the next century, possibly within the first few years.

There are several approaches to developing the software. One is to emulate the basic principles of biological brains. It is not implausible to suppose that these principles will be well enough known within 15 years for this approach to succeed, given adequate hardware.

The stagnation of AI during the seventies and eighties does not have much bearing on the likelihood of AI to succeed in the future since we know that the cause responsible for the stagnation (namely, that the hardware available to AI researchers was stuck at about 10^6 ops) is no longer present.

There will be a strong and increasing pressure to improve AI up to human level. If there is a way of guaranteeing that superior artificial intellects will never harm human beings then such intellects will be created. If there is no way to have such a guarantee then they will probably be created nevertheless.

Postscript I

(25 October, 1998)

The U.S. Department of Energy has ordered a new supercomputer from IBM, to be installed in the Lawrence Livermore National Laboratory in the year 2000. It will cost $85 million and will perform 10 Tops. This development is in accordance with Moore's law, or possibly slightly more rapid than an extrapolation would have predicted.

Many steps forward have been taken during the past year. An especially nifty one is the new chip-making techniques being developed at Irvine Sensors Corporation (ISC). They have found a way to stack chips directly on top of each other in a way that will not only save space but, more importantly, allow a larger number of interconnections between neighboring chips. Since the number of interconnections has been a bottleneck in hardware implementations of neural networks, this breakthrough could prove very important. In principle, it should allow you to have an arbitrarily large cube of neural network modules with high local connectivity and moderate non-local connectivity.

Postscript II

(28 August, 2000)

Is progress still on schedule? – In fact, things seem to be moving somewhat faster than expected, at least on the hardware front. (Software progress is more difficult to quantify.) IBM is currently working on a next-generation supercomputer, Blue Gene, which will perform over 10^{15} ops. This computer, which is designed to tackle the protein folding problem, is expected to be ready around 2005. It will achieve its enormous power through massive parallelism rather than through dramatically faster processors. Considering the increasing emphasis on parallel computing, and the steadily increasing Internet bandwidth, it becomes important to interpret Moore's law as a statement about how much computing power can be bought for a given sum of (inflation adjusted) money. This measure has historically been growing at the same pace as processor speed or chip density, but the measures may come apart in the future. It is how much computing power that can be bought for, say, 100 million dollars that is relevant when we are trying to guess when superintelligence will be developed, rather than how fast individual processors are.

Postscript III

(30 October, 2005)

The fastest supercomputer today is IBM's Blue Gene/L, which has attained 260 Tops (2.6×10^{14} ops). The Moravec estimate of the human brain's processing power (10^{14} ops) has thus now been exceeded.

The "Blue Brain" project was launched by the Brain Mind Institute, EPFL, Switzerland and IBM, USA in May, 2005. It aims to build an accurate software replica of the neocortical column within 2-3 years. The column will consist of 10,000 morphologically complex neurons with active ionic channels. The

neurons will be interconnected in a 3-dimensional space with 10^7–10^8 dynamic synapses. This project will thus use a level of simulation that attempts to capture the functionality of individual neurons at a very detailed level. The simulation is intended to run in real time on a computer performing 22.8×10^{12} flops. Simulating the entire brain in real time at this level of detail (which the researchers indicate as a goal for later stages of the project) would correspond to circa 2×10^{19} ops, five orders of magnitude above the current supercomputer record. This is two orders of magnitude greater than the estimate of neural-level simulation given in the original paper above, which assumes a cruder level of simulation of neurons. If the "Blue Brain" project succeeds, it will give us hard evidence of an upper bound on the computing power needed to achieve human intelligence.

Functional replication of the functionality of early auditory processing (which is quite well understood) has yielded an estimate that agrees with Moravec's assessment based on signal processing in the retina (i.e. 10^{14} ops for whole-brain equivalent replication).

No dramatic breakthrough in general artificial intelligence seems to have occurred in recent years. Neuroscience and neuromorphic engineering are proceeding at a rapid clip, however. Much of the paper could now be rewritten and updated to take into account information that has become available in the past 8 years.

Molecular nanotechnology, a technology that in its mature form could enable mind uploading (an extreme version of the bottom-up method, in which a detailed 3-dimensional map is constructed of a particular human brain and then emulated in a computer), has begun to pick up steam, receiving increasing funding and attention. An upload running on a fast computer would be weakly superintelligent – it would initially be functionally identical to the original organic brain, but it could run at a much higher speed. Once such an upload existed, it might be possible to enhance its architecture to create strong superintelligence that was not only faster but functionally superior to human intelligence.

Postscript IV

(12 March, 2008)

I should clarify what I meant when in the abstract I said I would "outline the case for believing that we will have superhuman artificial intelligence within the first third of the next [i.e. this] century". I chose the word "case" deliberately: In particular, by outlining "the case for", I did not mean to deny that one could also outline a case against. In fact, I would all-things-considered assign less than a 50% probability to superintelligence being developed by 2033. I do think there is great uncertainty about whether and when it might happen, and that one should take seriously the possibility that it might happen by then, because of the kinds of consideration outlined in this paper.

There seems to be somewhat more interest now in artificial general intelligence (AGI) research than there was a few years ago. However, it appears that as yet no major breakthrough has occurred.

Acknowledgements

The author would like to thank all those who have contributed comments on earlier versions of this paper. The very helpful suggestions by Hal Finney, Robin Hanson, Carl Feynman, Anders Sandberg, and Peter McCluskey were especially appreciated.

References

- Bostrom N. 1996. "Cortical Integration: Possible Solutions to the Binding and Linking Problems in Perception, Reasoning and Long Term Memory". Tech. report, available from https://nickbostrom.com/old/cortical.
- Cohen L., G. et al. 1997. "Functional relevance of cross-modal plasticity in blind humans". *Nature* 389: 180-83.
- de Garis, H. 1997. Home page. http://www.hip.atr.co.jp:80/~degaris/
- Hameroff & Penrose http://psyche.cs.monash.edu.au/psyche-index-v2.html
- Koch, C. 1997. "Computation and the single neuron". *Nature* 385: 207-10.
- Moravec, H. 1998. "When will computer hardware match the human brain?" *Journal of Transhumanism*, vol. 1. At https://www.jetpress.org/volume1/moravec.pdf
- Moravec, H. 1997. https://www.frc.ri.cmu.edu/~hpm/book97/ch3/retina.comment.html
- Morillo, C., R. 1992. "Reward event systems: reconceptualizing the explanatory roles of motivation, desire and pleasure". *Phil. Psych.* Vol. 5, No. 1, pp. 7-32.
- Phillips W. A. & Singer W. 1997. "In Search of Common Foundations for Cortical Computations". *Behavioural and Brain Sciences*, 20, 657-722.
- Quartz S. R. & Sejnowski T. J. 1997. "The neural basis of cognitive development: A constructivist manifesto". *Behavioural and Brain Sciences*, 20, 537-596.
- Schlaggar, B. L. & O'Leary, D. D. M. 1991. "Potential of visual cortex to develop an array of functional units unique to somatosensory cortex". *Science* 252: 1556-60.
- Sur, M. et al. 1988. "Experimentally induced visual projections into auditory thalamus and cortex". *Science* 242: 1437-41

- White, E. L. 1989. *Cortical Circuits: Synaptic Organization of the Cerebral Cortex. Structure, Function and Theory.*