

Artificial Consciousness: Utopia or Real Possibility?

Giorgio Buttazzo
University of Pavia

Can a machine ever become self-aware?
Does consciousness depend on the material neurons are made of or can humans use hardware to replicate it?

Since the beginnings of computer technology, researchers have speculated about the possibility of building smart machines that could compete with human intelligence. Given the current pace of advances in artificial intelligence and neural computing, such an evolution seems to be a more concrete possibility. Many people now believe that artificial consciousness is possible and that, in the future, it will emerge in complex computing machines.

However, a discussion of artificial consciousness gives rise to several philosophical issues:

- Can computers think or do they just calculate?
- Is consciousness a human prerogative?
- Does consciousness depend on the material that comprises the human brain, or can computer hardware replicate consciousness?

Answering these questions is difficult because it requires combining information from many disciplines including computer science, neurophysiology, philosophy, and religion. Further, we must consider the influence of science fiction—especially science fiction films—when addressing artificial consciousness. As a product of the human imagination, such works express human desires and fears about future technologies and may influence the course of progress. At a societal level, science fiction simulates future scenarios that can help prepare us for crucial transitions by predicting the consequences of significant technological advances.

ROBOTS IN SCIENCE FICTION

Since the early 1950s, science fiction movies have depicted robots as sophisticated human-crafted machines that perform complex operations, work with us on safety-critical missions in hostile environments or, more often, pilot and control spaceships in galactic travels.¹ At the same time, however, the film industry has portrayed *intelligent* robots as dangerous entities capable of working against humanity in their pursuit of self-serving agendas.

Robotic rampages

The most significant example of this archetype, HAL 9000, is the main character in Stanley Kubrick's 1968 epic, *2001: A Space Odyssey*. Although HAL

controls the entire spaceship, talks amiably with the astronauts, plays chess, renders aesthetic judgments, and recognizes the crew's emotions, it also murders four of the five astronauts in pursuit of a plan elaborated from flaws in its programming.²

More recent films, such as James Cameron's *Terminator* and the Wachowski brothers' *The Matrix*, present an even more catastrophic view of the future in which robots become self-aware and dominate the human race. For example, Cameron's 1991 film, *Terminator 2: Judgment Day*, begins with a scene depicting a horrendous war between humans and robots, which onscreen text forecasts will occur in Los Angeles, 2029 AD. According to Cameron's script, the fictional corporation Cyberdyne becomes the US military's largest computer systems supplier. It then morphs into Skynet, a powerful neural processing network built to execute strategic defense decisions. The network, however, becomes self-aware; when human engineers try to deactivate it, the network retaliates by unleashing the US's nuclear arsenal on its creators.

Robotic rescues

Few movies depict robots as reliable assistants that serve humans rather than conspire against them. In Robert Wise's 1951 film, *The Day the Earth Stood Still*, Gort is perhaps the first robot—albeit an extraterrestrial one—that supports a humanitarian agenda by helping its alien owner deliver an offer of peaceful coexistence to Earth. Likewise, Cameron's 1986 film, *Aliens*, shows a synthetic android that acts on behalf of its human owners despite their suspicion of it.

When SF and reality collide

Strongly influenced by theories on connectionism and artificial neural networks, which seek to replicate processing mechanisms typical of the human brain,³ Cameron's *Terminator* represents the prototypical imaginary robot. The robot can walk, talk, perceive, and behave like a human being. Its power cell can supply energy for 120 years, and an alternate power circuit provides fault tolerance in case of damage. More importantly, *Terminator* can learn. Essentially, a neural processor—a computer that modifies the robot's behavior based on past experience—controls it. Intriguingly, the neural processor is so complex, it learns at an exponential rate, eventually becoming self-aware. Thus, the film raises important philosophical questions about artificial consciousness. Can a machine become self-aware? If so, how can we verify that an intelligent being is self-conscious?

TURING TEST

The work of computer science pioneer Alan Turing may help answer these questions. In 1950, Turing

tackled a similar problem focused on intelligence. To establish whether we can consider a machine as intelligent as a human, he proposed the now-famous Turing test. The test uses two keyboards—one connected to a computer, the other positioned in front of a human operator. Both computer and operator are hidden from view, with only a monitor visible to display their output.

An examiner inputs questions on any topic that comes to mind. Both the computer and the human respond to each question. If the examiner cannot with confidence distinguish between the computer and the operator based on the nature of their answers, we must conclude that the machine has passed the Turing test.

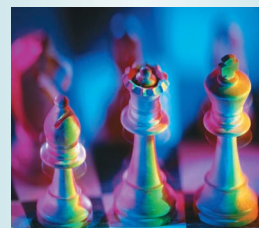
In 1990, the Turing test received its first formal acknowledgment. Hugh Loebner, a New York philanthropist, and the Cambridge Center for Behavioral Studies in Massachusetts established the Loebner Prize Competition in Artificial Intelligence (<http://www.loebner.net/Prize/loebner-prize.html>). Loebner pledged to award a \$100,000 prize for the first computer whose responses could not be distinguished from a human's. The first competition took place at the Computer Museum of Boston in November 1991.

Although the contest was constrained to a single narrow topic for some years, since 1998 the questioning's scope has been unlimited. After the conversation, each judge scores the interlocutor on a scale of 1 to 10, in which 1 means human and 10 indicates computer.⁴ So far, no computer has given responses totally indistinguishable from a human's, but every year the computer's scores edge closer to an average of 5. Nevertheless, current computers can pass the Turing test only if we restrict the interaction to highly specific topics—like chess.

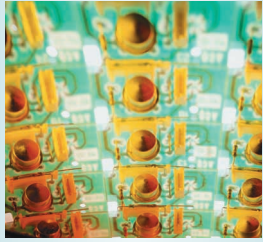
Deep Blue conquers chess

In 1997, for the first time in history, a computer beat reigning world chess champion Garry Kasparov. Like all computers, however, the victor—IBM's Deep Blue—does not understand chess; it simply applies rules to find a move that leads to a better position, according to an evaluation criterion programmed by chess experts.

Claude Shannon estimated that the search space in a chess game includes about 10^{120} possible positions. Deep Blue could analyze 200 million positions per second (<http://www.research.ibm.com/deepblue/meet/html/d.3.html>). Exploring the entire search space for Deep Blue would therefore take about 10^{95} billion years. Nevertheless, Deep Blue's victory can be attrib-



A combination of speed and a smart search algorithm gave Deep Blue positional and material advantages.



Even if machines become as skilled as humans in many disciplines, we cannot assume they have become self-aware.

uted to a combination of speed and a smart search algorithm, which gave the computer positional and material advantages.

Although mathematics clearly shows that Kasparov succumbed to brute-force computation rather than sophisticated machine intelligence, in interviews during and after the match he expressed doubts that his opponent was a computer and reported that, at times, he felt as if he were playing against a human. Kasparov also remarked on the beauty of the computer's moves, ascribing aesthetic motivations to what were, essentially, the results of raw computation. If we accept Turing's view, we can say that Deep Blue plays chess intelligently, but we must also admit that the computer no more understands the

meaning of its moves than a television understands the meaning of the images it displays.

Computers challenge humans in other domains

In addition to chess, computers have begun approaching human ability in an increasing number of other domains. In music, for example, many commercial programs can create melodic lines or entire songs according to specific styles, ranging from Bach to jazz. Other programs can also generate solos atop a given chord sequence, emulating jazz masters like Charlie Parker and Miles Davis much better than an average human musician could.

In 1997, Steve Larson, a music professor at the University of Oregon, proposed a musical variation of the Turing test. He asked an audience to listen to pairs of classical compositions and determine for each pair which one was written by a computer and which was the authentic composition. The audience classified many computer pieces as authentic compositions and vice versa. A loose interpretation of these results could indicate that, with regard to musical composition, the computer passed the Turing test.⁵

Computers now approach human levels of understanding in continuous speech, electrocardiogram diagnostics, theorem proving, and aircraft guidance. In the future, we can expect similar levels of computer performance in areas that include complex tasks such as driving, real-time language translation, house cleaning, surgery, surveillance, and law enforcement.

However, even if machines become as skilled as humans in many disciplines, such that we cannot distinguish between their performance and that of humans, we cannot assume that they have become self-aware. At the same time, we cannot assume that such machines are *not* self-aware. In fact, while intelligence is an expression of an external behavior that

we can measure with specific tests, self-consciousness is a property of an internal brain state, which we cannot measure. Hence, to resolve this question, we must turn to philosophy.

PHILOSOPHICAL VIEWS OF SELF-AWARENESS

From a purely philosophical perspective, we cannot verify the presence of consciousness in another brain, either human or artificial, because only the possessor itself can verify this property. Because we cannot enter another being's mind, we cannot be sure about its consciousness. Douglas Hofstadter and Daniel Dennett⁶ discuss this problem at length in their book, *The Mind's I*.

Nevertheless, we can develop theories regarding the nature of self-awareness based on different philosophical approaches.

Pragmatic

We could follow Turing's approach and say that we can consider a being self-conscious if it can convince us by passing specific tests. We base our belief that humans are self-conscious on our inherent similarities: Because we have the same organs and similar brains, it is reasonable to conclude that if each of us is self-conscious, so is everyone else. If, however, the creature in front of us, although behaving like a human, were comprised of synthetic tissues, mechatronic organs, and neural processors, we might respond differently.

The most common objection to granting electronic-circuit-driven computers self-conscious status is the perception that, working in a fully automated mode, they cannot exhibit creativity, emotions, or free will. A computer, like a washing machine, is a slave operated by its components.

Logic demands, however, that we must apply this reasoning to machines' biological counterparts. At a neural level, the same electrochemical reactions present in machinery operate in the human brain. Each neuron automatically responds to its inputs according to fixed laws. However, these mechanisms do not prevent us from experiencing happiness, love, or irrational behaviors.

With the emergence of artificial neural networks, the problem of artificial consciousness becomes even more intriguing because neural networks replicate the brain's basic electrical behavior and provide the proper support for realizing a processing mechanism similar to the one adopted by the brain. In *Impossible Minds: My Neurons, My Consciousness*, Igor Aleksander⁷ addresses this topic in depth and with scientific rigor.

Religious

If we remove the structural diversity between biological and artificial brains, artificial consciousness

becomes a religious issue. If we believe that divine intervention determines human consciousness, no artificial system can ever become self-aware. If, instead, we believe that human consciousness is a natural electrical property developed by complex brains, realizing an artificial self-aware being remains an open possibility.

Dualistic

Many believe that consciousness is not a product of brain activity, but rather a separate immaterial entity, often identified with the soul. Seventeenth-century philosopher René Descartes developed such a dualistic theory about the brain and mind. However, this theory raised several unanswerable questions:

- If a mind is separate from its brain, how can it physically interact with the body and activate a neural circuit?
- If the mind operates outside the brain, how does it move atoms to create electrical pulses?
- Do mysterious forces activate neural cells?
- Does a mind interact with a brain by violating the fundamental laws of physics?
- If a conscious mind can exist outside the brain, why do we have a brain?
- If emotions and thoughts come from outside, why does a brain stimulated with electrodes and drugs respond by generating thoughts?
- Why does the patient experience severely affected conscious behavior when a portion of a diseased brain is surgically removed?

These and other concerns caused this theory to fade from popularity in the philosophical community.

Reductionism and idealism

To resolve dualism's inconsistencies, researchers developed alternative theories. At one extreme, reductionism did not recognize the existence of mind as a subjective, private sense-data construct and considered all mental activities as specific neural states of the brain. At the other extreme, idealism tried to reject the physical world by considering all events as a product of mental constructions.

With the progress of computer science and artificial intelligence, scientists and philosophers developed a new approach that considers the mind a form of computation emerging at a higher level of abstraction with respect to neural activity.

The major weakness of the reductionist approach to comprehending the mind is that it attempts to recursively decompose a complex system into simpler subsystems, until at some stage the units can be fully analyzed and described. This method works perfectly for linear systems, in which any output can be seen as

a sum of simpler components. However, a complex system is often nonlinear. Thus, analyzing a system's basic components offers insufficient understanding of its global behavior. Complex systems contain holistic features that cannot be seen at a smaller detail level, appearing instead only when we consider the structure and interactions among the system's components.

Paul Davies, in *God and the New Physics*,⁸ explains this concept by observing that a digital picture of a face consists of many colored dots—pixels. But the shape takes form only when we observe the picture at a distance that allows us to see all the pixels. Thus, the face is not a property of the individual pixels but of the set of pixels.

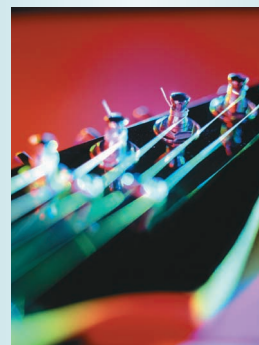
The brain as ant colony

In *Gödel, Escher, Bach*,⁹ Douglas Hofstadter explains this concept by describing a large ant colony's behavior. Ants have a complex and highly organized social structure based on work distribution and collective responsibility. Although each ant has minimal intelligence and limited capabilities, the whole ant colony exhibits a highly complex behavior. In fact, building an ant colony requires a large and intricate design, but clearly no individual ant can conceptualize the entire project. Nevertheless, a scheme and a finalized behavior emerge at the colony level. In some sense, we can consider the whole colony a living being.

A brain resembles a large ant colony in many respects. For instance, a brain consists of billions of neurons that cooperate to achieve a common objective. The interaction among neurons is tighter than among ants, but the underlying principles are similar—work subdivision and collective responsibility. Consciousness is not a property of individual neurons, which automatically operate as switches that respond to input signals. Rather, consciousness is a holistic property that emerges and flourishes from neural cooperation when the system reaches a sufficiently organized complexity.

Consciousness and matter

Although most people can accept the existence of holistic features, others still believe that consciousness cannot emerge from a silicon substratum because it is an intrinsic property of biological materials such as neural cells. Thus, we can reasonably ask: Does consciousness depend on the material that comprises neurons? In *Beyond Humanity: CyberEvolution and Future Minds*,¹⁰ Gregory S. Paul and Earl Cox write:



Complex systems contain holistic features that appear only when we consider the structure and interaction among the system's components.

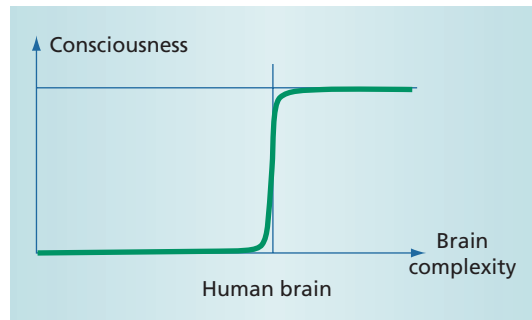


Figure 1. The self-awareness threshold. If consciousness is a function of brain complexity, the human brain marks the complexity threshold required for conscious thought.

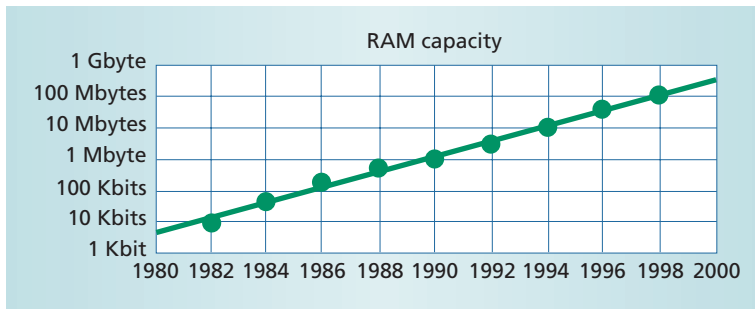


Figure 2. Typical random-access memory configurations installed in personal computers in the past 20 years.

It would be astonishing that the most powerful information-processing tool can be derived only from organic cells and chemistry. Aircraft [are] made out of different materials from birds, bats, and bugs; solar panels are made out of different materials from leaves. There is usually more than one way to build a given type of machine.... This just happens to be what genetics was able to work with.... Other elements and combinations of elements may be able to do a better job processing information in a self-aware manner.

If we support the hypothesis of consciousness as a physical property of the brain, the question becomes: When will a computer become self-aware?

A PREDICTION

Attempting to provide an answer to this question is hazardous. Nevertheless, we can determine at least a necessary precondition without which a machine cannot develop self-awareness. This precondition derives from the assertion that, to develop self-awareness, a neural network must be at least as complex as the human brain.

Why this assertion? Because it appears that less-complex brains cannot produce conscious thought.

Consciousness seems to represent a step function of brain complexity and the human brain provides the threshold, as Figure 1 shows.

How much memory would a computer require to replicate the human brain's complexity? The human brain has about 10^{12} neurons, and each neuron makes about 10^3 synaptic connections with other neurons, on average, for a total of 10^{15} synapses. Artificial neural networks can simulate synapses using a floating-point number that requires 4 bytes of memory to be represented in a computer. As a consequence, simulating 10^{15} synapses requires a total of 4 million Gbytes. Simulating the human brain requires 5 million Gbytes, including the auxiliary variables for storing neuron outputs and other internal brain states.

When will such a memory be available in a computer? During the past 20 years, random-access memory capacity increased exponentially by a factor of 10 every four years. The plot in Figure 2 shows the typical memory configuration installed on personal computers since 1980.

By interpolation, we can derive the following equation, which gives RAM size as a function of the year:

$$\text{bytes} = 10^{\left(\frac{\text{year} - 1966}{4}\right)}$$

For example, from this equation we can derive that in 1990, personal computers typically had 1 Mbyte of RAM, whereas in 1998, a typical configuration had 100 Mbytes of RAM. Assuming that RAM will continue to grow at the same rate, we can invert this relationship to predict the year in which computers will have a given amount of memory:

$$\text{year} = 1966 + 4\log_{10}(\text{bytes})$$

To calculate the year in which computers will have 5 million Gbytes of RAM, we substitute that number in the equation above. This gives us the year 2029. Ray Kurzweil,⁵ Gregory S. Paul and Earl Cox,¹⁰ and Hans Moravec¹¹ derived similar predictions.

To understand the calculated prediction, we must take into account several important considerations. First, the computed date refers only to a necessary but not sufficient condition for the development of an artificial consciousness. The existence of a powerful computer equipped with millions of gigabytes of RAM is not in itself sufficient to guarantee that the machine will become self-aware.

Other important factors influence this process, such as the progress of theories on artificial neural networks and the basic biological mechanisms of mind, for which it is impossible to attempt precise estimates. Further, some could argue that the presented computation was done on personal computers, which do not

represent top-of-the-line technology. Others could object that the same amount of RAM could be available using a computer network or virtual-memory management mechanisms to exploit hard-disk space. In any case, even if we adopt different numbers, the computation's basic principle remains the same, and we could advance the date by only a few years.

How about Moore's law?

Some may object that the 2029 prediction relies on mindless extrapolation of current trends, without considering events that could alter that trend. Gordon Moore, one of Intel's founders, noted the exponential growth of computing power and memory in 1973. He predicted that the number of transistors on integrated circuits would continue to double every 18 months until reaching fundamental physical limits. That prediction proved so accurate that it is called Moore's law.

But how much longer will this law hold true? Chip companies estimated that Moore's law will be valid for another 15 to 20 years. However, when transistors reach the size of a few atoms, the conventional approach will not work, and this paradigm will break down. What's next? Will the microprocessor evolution come to an end around 2020?

Computing power's exponential growth

Some people, including Ray Kurzweil⁵ and Hans Moravec,¹¹ noticed that computers were growing exponentially in power long before the integrated circuit's invention in 1958, despite the hardware used. So, Moore's law was not the first but actually the fifth paradigm to track computing's exponential growth. Each new paradigm came along precisely when needed, suggesting that exponential growth will not stop when Moore's law is no longer valid. For example, scientists are investigating new technologies such as three-dimensional chip design, optical computing, and quantum computing¹² that promise to extend computing power's exponential growth for many years to come.

RELATED ISSUES

Other important aspects of artificial consciousness include the possibility of achieving self-awareness in sequential machines and the notion of time for a fast-thinking mind.

Sequential machines

Could a sequential machine develop self-awareness? If consciousness is a product of a highly organized information-processing system, that property does not depend on the hardware substratum that performs the computation and thus could also emerge in a sequential machine. Indeed, sequential computers simulate

most artificial neural networks today because such computers are more flexible than a hardwired network.

Someone could argue that a simulated process differs from the process itself. Clearly, this is true when simulating a physical phenomenon, like a thunderstorm or a planetary system. However, for a neural network, a simulation and a physical network produce the same result because both are information-processing systems. Similarly, the software calculator applet that runs on personal computers is functionally equivalent to its hardware counterpart and can perform the same operations.

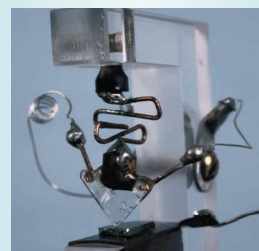
The notion of time

Does consciousness depend on the response time of the computing elements? It is hard to say because we cannot change our brain's speed. Experiments on real-time control systems suggest that processing speed is important to meet the timing constraints the physical world imposes on our actions, but it should not affect the results of a computation. In other words, a different neural speed would certainly give us a different perception of time but perhaps would not prevent us from being self-aware.

If we could ideally slow down the neurons uniformly in our entire brain, we would perhaps perceive the world as a fast-motion movie in which events occur faster than our reactive capabilities. This sounds reasonable because our brain evolved and adapted in a world in which the important events for reproduction and survival fall within a scale that spans a few tenths of a second. If we could speed up the events in the environment or slow down our neurons, we would not be able to operate in real time in such a world and probably would not survive.

Conversely, how would we feel with a faster brain? Today a logic port is six orders of magnitude faster than a neuron. Biological neurons respond within a few milliseconds, but electronic circuits can respond within a few nanoseconds. This observation leads to an interesting question: If consciousness emerges in an artificial machine, what will time perception be like to a simulated brain that thinks millions of times faster than a human brain?

It is possible that, for conscious machines, the world around them will seem to move slower. Perhaps insects view the world this way, so that, to a fly, a swatting human hand looks like it is moving in slow motion, thus giving the fly plenty of time to glide leisurely out of the way.



When transistors reach the size of a few atoms, the conventional approach will not work, and Moore's law will break down.

Gregory S. Paul and Earl Cox¹⁰ address this issue in *Beyond Humanity: CyberEvolution and Future Minds*:

... a cyberbeing will be able to learn and think at hyperspeed. They will observe a speeding bullet fired from a moderate distance away, calculate its trajectory, and dodge it if necessary.... Imagine being a robot near a window. To your fast thinking mind, a bird takes what seems like hours to cross the field of view, and a day lasts seemingly forever.

Why should we build a self-aware machine? Except for ethical issues that could significantly influence progress in this field, the strongest motivation for constructing a self-aware machine is the innate human desire to discover new horizons and enlarge the frontiers of science.

Further, developing an artificial brain based on the same principles as in the biological brain would provide a means for transferring the human mind to faster and more robust support, opening the door to immortality. Freed from a fragile and degradable body, a

human being with synthetic organs, including an artificial brain, could represent humanity's next evolutionary step.

Such a new species—a natural result of human technological progress—could quickly colonize the universe, search for alien civilizations, survive to the death of the solar system, control the energy of black holes, and move at the speed of light by transmitting the information necessary for replication to other planets. As has proven the case with all important human discoveries—from nuclear energy to the atomic bomb, from genetic engineering to human cloning—the real problem will be keeping technology under control. Should self-aware computers become possible, we must ensure that we use them for human progress and not for catastrophic aims. *

References

1. G. Buttazzo, "Can a Machine Ever Become Self-Aware?," *Artificial Humans*, R. Aurich, W. Jacobsen, and G. Jatho, eds., Goethe Institute, Los Angeles, 2000, pp. 45-49.
2. D.G. Stork, ed., *HAL's Legacy: 2001's Computer as Dream and Reality*, MIT Press, Cambridge, Mass., 1997.
3. I. Asimov, *I, Robot*, Grafton Books, London, 1968.
4. M. Krol, "Have We Witnessed a Real-Life Turing Test?," *Computer*, Mar. 1999, pp. 27-30.
5. R. Kurzweil, *The Age of Spiritual Machines*, Viking Press, New York, 1999.
6. D.R. Hofstadter and D.C. Dennett, *The Mind's I*, Harvester/Basic Books, New York, 1981.
7. I. Aleksander, *Impossible Minds: My Neurons, My Consciousness*, World Scientific, River Edge, N.J., 1997.
8. P. Davies, *God and the New Physics*, Simon & Schuster, New York, 1984.
9. D. Hofstadter, *Gödel, Escher, Bach*, Basic Books, New York, 1979.
10. G.S. Paul and E. Cox, *Beyond Humanity: CyberEvolution and Future Minds*, Charles River Media, Rockland, Mass., 1996.
11. H. Moravec, *Robot: Mere Machine to Transcendent Mind*, Oxford University Press, Oxford, UK, 1999.
12. L. Geppert, "Quantum Transistors: Toward Nanoelectronics," *IEEE Spectrum*, Sept. 2000, pp. 46-51.

Giorgio Buttazzo is a professor of computer engineering at the University of Pavia, Pavia, Italy. His research interests include real-time systems, advanced robotics, and neural networks. Buttazzo received a PhD in computer engineering from the Scuola Superiore Sant'Anna of Pisa. He is a member of the IEEE. Contact him at buttazzo@unipv.it.

**Help Shape
the IEEE
Computer
Society of
tomorrow.**

Vote for 2002

Computer Society officers.

Polls open 11 August

<http://computer.org/election/>

