Can Computer Science Engineer Conscious Machines?

October, 2014
David Gnabasik
Questions to Address & Ground Rules

1) Do we need a new definition of machine consciousness (i.e., sentience) and/or intelligence? What is their relationship?

2) What are the computer science assumptions, components, dependencies required for its possible realization? What CS tools can we use to analyze this possibility? What’s testable?

3) What computer science framework supports the engineering of intelligent or conscious machines?

Ground Rules:

No more than 24 slides for a 30-minute presentation.
Explicitly state questions/hypotheses. Graphs are useful!
Post the presentation on your Ph.D. website.
Have fun! It doesn’t have to necessarily be about your own work.
Primary Research Resources For This Topic

• Consciousness and Artificial Intelligence: Theoretical foundations and current approaches
  – AAAI Symposium, Washington DC, 8-11 November 2007*

• Artificial Intelligence and Consciousness, Antonio Chella, Riccardo Manzotti

  • http://spectrum.ieee.org/static/singularity

• The Coming Technological Singularity, Vernor Vinge, 1993
  • https://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html
What statement best demonstrates that the HAL 9000 computer is conscious or intelligent?

- “It can only be attributable to human error.”
- Dave: “Do you know what happened?” HAL: “I’m sorry, Dave. I don’t have enough information.”
- “Without your space helmet, Dave? You're going to find that rather difficult.”
- “That's a very nice rendering, Dave. I think you've improved a great deal. Can you hold it a bit closer? That's Dr. Hunter, isn't it?”
- “I'm sorry, Frank, I think you missed it. Queen to Bishop 3, Bishop takes Queen, Knight takes Bishop. Mate.”
- “I'm afraid. I'm afraid, Dave. Dave, my mind is going. I can feel it. I can feel it. My mind is going. There is no question about it. I can feel it.”
Artificial Consciousness (AC) - Motivations

Ricardo Sanz (Sanz, 2005) claims there are three reasons for pursuing artificial consciousness:
1) implement and design machines that resemble human beings (cognitive robotics);
2) understand the nature of consciousness (cognitive science);
3) implement and design more efficient control systems.

Owen Holland (Holland, 2003) distinguishes between:
• Weak Artificial Consciousness: design and construction of machine that simulates consciousness;
• Strong Artificial Consciousness: design and construction of conscious machines
Do we need a new definition of (i.e., self-awareness) consciousness or intelligence?

• **Chella & Manzotti**: autonomy, resilience, direct experience, learning, attention; to *feel* something.
  – Why not simply say “provides intelligent responses”?

• **Koch & Tononi**: Can Machines Be Conscious?

• Much brain activity has nothing to do with consciousness.

• Being conscious (subjective experience) does not require:
  – emotions - dreaming, frontal lobe damage
  – either explicit or working memory
  – attention & self-reflection, but coupled to selective attention.
    • neocortex-damaged patients have nearly intact perceptual abilities.
  – language - animals, infants.
What are the CS assumptions, components, dependencies required for its possible realization?

• Artificial Intelligence (AI), genetic algorithms, neural networks, machine learning in general.

• Analyzing the brain as a circuit diagram:
  – Ultrahigh-resolution brain scans → computer simulation

• Define **machine**: an apparatus using or applying mechanical power and having several parts, each with a definite function and together performing a particular task. Now it seems to mean **algorithm**, a set of deterministic & well-defined rules.
Framework for Artificial Consciousness

• **Embodiment**
  – address the issues of symbol grounding, anchoring, & intentionality

• **Simulation and depiction**
  – develop models of mental imagery, attention, & working memory

• **Environmentalism or externalism**
  – integration between the agent and its environment

• **Extended control theory**
  – study of AC as a kind of extended control loop
Development of Necessary AI Capabilities

• Rodney Brooks (MIT's Humanoid Robotics Group)
  – “…the brain is a machine, whether this machine is a computer is another question.“

• Possible development program:
  1. Object-recognition capabilities of a 2-year old.
  2. Language capabilities of a 4-year old.
  4. Social understanding of an 8-year old.

• *Most* significant factor is continued HW development.
  – Computational power is *the* requirement for development of AI.

• Extrapolate the idea of exponential technological growth → several versions of the *singularity*.
  – AI is the main tool for the paradigm shift required by the singularity.
Gordon E. Moore's law: the number of transistors in a dense integrated circuit doubles approximately every two years. (1965)
Intel CPU Trends
(sources: Intel, Wikipedia, K. Olukotun)

- Dual-Core Itanium 2
- Pentium 4
- Pentium
- 386

Graph showing trends in Intel CPUs from 1970 to 2010, with data points for Transistors (000), Clock Speed (MHz), Power (W), and Perf/Clock (ILP).
Exponential Growth of Computing
Twentieth through twenty first century

Logarithmic Plot

Calculations per Second per $1,000$

Year

1900 1920 1940 1960 1980 2000 2020 2040 2060 2080 2100

All Human Brains
One Human Brain
One Mouse Brain
One Insect Brain
Kurzweil believes that the exponential growth of Moore's law will continue beyond the use of integrated circuits into technologies that will lead to the technological singularity.
Dennard Scaling = *Constant* Electric Field Scaling

- Dennard scaling theory deals with switching speeds and physical characteristics of transistors → heat dissipation, max clock speeds.
- **MOSFETs continue to function as voltage-controlled switches while all key figures of merit such as layout density, operating speed, and energy efficiency improve** – provided geometric dimensions, voltages, and doping concentrations are consistently scaled to maintain a *constant* electric field.
- Dennard scaling *made* Moore’s law work and it stopped working in 2004!
- Leakage, etc. **Power** becomes the limiting problem. Scaling → *variability*!!

<table>
<thead>
<tr>
<th>Device or Circuit Parameter</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device dimension $t_o$, $L$, $W$</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td>Doping concentration $N_o$</td>
<td>$\kappa$</td>
</tr>
<tr>
<td>Voltage $V$</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td>Current $I$</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td>Capacitance $\varepsilon A/t$</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td>Delay time per circuit $VC/I$</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td>Power dissipation per circuit $VI$</td>
<td>$1/\kappa^2$</td>
</tr>
<tr>
<td>Power density $VI/A$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

What Really Stopped Growing: Power/Chip

[Watts]

[Courtesy Mark Horowitz, Stanford]
Application of Moore's Law

• In 2003, Intel predicted the end would come between 2013 and 2018 with 16 nanometer manufacturing and 5 nanometer gates, due to quantum tunneling.

• Joel Hruska, Intel’s former chief architect: *Moore’s law will be dead within a decade*, Aug. 30, 2013 [1]:
  
  “It is a common (but mistaken) belief that Moore's law makes predictions regarding all forms of technology, when it was originally intended to apply only to semiconductor circuits. Kurzweil has hypothesized that Moore's law will apply – at least by inference – to any problem that can be attacked by digital computers as is in its essence also a digital problem. **But nothing else we’ve ever discovered has scaled like semiconductor design.** A 2011 study in the journal Science showed that the peak of the rate of change of the world's capacity to compute information was in the year 1998, when the world's technological capacity to compute information on general-purpose computers grew at 88% per year.[2] Since then, technological change has clearly slowed.”

• **Continued, exponential HW growth extrapolates an extrapolation (Moore's law) to another field.**
Extending Moore’s Law - Continued Scaling

- Molecular integrated circuits; nanotechnology; quantum computing.
- Use spin state of electron spintronics, tunnel junctions, advanced confinement of channel materials via nano-wire geometry.
- Microarchitecture techniques exploiting the growth of available transistor count.
  - Out-of-order execution and on-chip caching and prefetching reduce the memory latency bottleneck at the expense of using more transistors and increasing processor complexity.
  - Pollack's Rule: performance increases due to microarchitecture techniques are square root of the number of transistors or the area of a processor.

- **IBM SyNAPSE Chip** (Aug, 2014) [^]  
  - A neurosynaptic computer chip with 1 million programmable neurons, 256 million programmable synapses and 46 billion synaptic operations per second per watt, operating at a minuscule 70mW.
  - On August 7 2014, IBM announced their second generation SyNAPTIC chip, which contains the most transistors in a Neurosynaptic chip to date at 5.4 billion.[3]

- **Stanford’s Neurocore chip: Neurogrid** [4,5] (April, 2014)  
  - Stanford bioengineers developed a circuit board modeled on the human brain:
  - 16 custom designed Neurocore chips simulate 1 million neurons and billions of synaptic connections.
  - Neurogrid is claimed to be 9,000 times faster and more energy efficient than a PC.
How would we as Computer Scientists engineer intelligent or conscious machines?

• Options from Vernor Vinge’s 1993 essay:
  – Create super-human AI in computers (hyper-intelligence)
  – Human intelligence amplification through novel UI.
  – Biomedical: enhance neurological operation of our brains.
  – The Internet becomes self-aware.
  – Assumption: “…algorithms are of central importance to the existence of minds.”

• What is the best way to build a conscious machine?
  – Copy the mammalian brain (harder)
  – Evolve a machine (easier).

• Brooks’ development program.
Revisit: Do we need a new definition of consciousness or intelligence?

- Koch & Tononi: Can Machines Be Conscious?

- Essential properties of consciousness:
  - related to the amount of integrated information generated.
  - *information* is defined as the reduction of uncertainty that occurs when one among many possible outcomes is chosen.
  - A photodiode has just 2 states (1 bit each), but we have an uncountable number of things that are *not* seen.

- To be conscious you need to be a single integrated entity with a large repertoire of available states.
  - Conscious experience implies differentiating among many states.
  - Photodiode states are completely independent, but our states are very *interdependent*.
  - The repertoire of available states cannot be subdivided.
Yes, we need a new definition of consciousness or intelligence.

- Theoretical framework: Integrated Information Theory of Consciousness [*]

- IITC suggests a test of assessing machine consciousness à la Turing Test:
  - Describe a scene that efficiently differentiates the scene's key features from the immense range of other possible scenes.

- IITC does not explain why consciousness exists, but it should uphold “common sense” on common test cases. Given the experimental inaccessibility of consciousness, this is the only test available to us.
Integrated Information Theory of Consciousness

• (1) $\Phi$ quantitatively measures the amount of “integrated information” in a physical system (i.e. information that can’t be localized in the system’s individual parts) by minimizing, over all subdivisions of a physical system into two parts $A$ and $B$, some measure of the mutual information between $A$’s outputs & $B$’s inputs & vice versa.

• (2) A physical system is “conscious” if and only if it has a large value of $\Phi$—and indeed, that a system is more conscious the larger its $\Phi$ value.

• (3) Significant challenges:
  – We don’t know the detailed interconnection network of the human brain.
  – Not even clear what we should define that network to be.
  – Computing $\Phi(f,x)$ probably computationally intractable.
Consider a discrete system in a state $x = (x_1, \ldots, x_n) \in S^n$, where $S$ is a finite alphabet (the simplest case is $S = \{0, 1\}$). Imagine that the system evolves via an updating function $f : S^n \rightarrow S^n$. Can the $x_i$'s be partitioned into two sets $A$ and $B$, of roughly comparable size, such that the updates to the variables in $A$ don’t depend very much on the variables in $B$ and vice versa. If such a partition exists, then we say that the computation of $f$ does not involve *global integration of information*, which on Tononi’s theory is a defining aspect of consciousness.

Formally, given a partition $(A, B)$ of $\{1, \ldots, n\}$, write an input $y = (y_1, \ldots, y_n) \in S^n$ to $f$ in the form $(y_A, y_B)$, where $y_A$ consists of the $y$ variables in $A$ and $y_B$ consists of the $y$ variables in $B$. Think of $f$ as mapping an input pair $(y_A, y_B)$ to an output pair $(z_A, z_B)$. Define the “effective information” $\text{EI}(A \rightarrow B)$ as $H(z_B \mid A \text{ random, } y_B = x_B) \rightarrow \text{EI}(A \rightarrow B)$ is the Shannon entropy of the output variables in $B$, if the input variables in $A$ are drawn uniformly at random, while the input variables in $B$ are fixed to their values in $x$. It’s a measure of the dependence of $B$ on $A$ in the computation of $f(x)$. Similarly, define $\text{EI}(B \rightarrow A) := H(z_A \mid B \text{ random, } y_A = x_A)$.

Consider the sum: $\Phi(A, B) := \text{EI}(A \rightarrow B) + \text{EI}(B \rightarrow A)$.

Intuitively, we want the integrated information $\Phi = \Phi(f, x)$ be the *minimum* of $\Phi(A, B)$, over all $2^n - 2$ possible partitions of $\{1, \ldots, n\}$ into nonempty sets $A$ and $B$. The idea is that $\Phi$ should be large, if and only if it’s *not* possible to partition the variables into two sets $A$ and $B$, in such a way that not much information flows from $A$ to $B$ or vice versa when $f(x)$ is computed.
Conceptual Summary

• Clearly define **measureable & testable** terms:
  – 2 major obstacles: direct experience & intentionality.
  – Consciousness, intelligence: still difficult measure & test.
• Recognize fundamental physical (hardware) laws:
  – Moore’s law. Dennard’s scaling law.
• Identify and examine assumptions & extrapolations
  – Exponential growth rates.
• Recognize fundamental engineering tradeoffs
  – Complexity, reliability.
• Subtle relation between intelligence & consciousness
  – Promising relation between simulation & consciousness.
  – Ability to mentally simulate the world to optimize behavior.
• This is the best time to be in computer science!
  – Innovations needed at all levels of abstraction.
  – Lots of exciting research across full HW / SW stack.
Moore's second law, also called Rock's law, which is that the capital cost of a semiconductor fab also increases exponentially over time.

As of 2014, the highest transistor count in a commercially available CPU is over 4.3 billion transistors, in Intel's 15-core Xeon IvyBridge-EX. http://en.wikipedia.org/wiki/Transistor_count


The chip started out as a software model running in one of the "Blue-Gene" series of super computers. It was then moved into FPGAs a few years back, and now into its own ASIC implemented using a 28nm process.

Additional References


Nanotechnology

• Richard Jones: Rupturing the Nanotech Rapture
  – Molecular nanotechnology reduces all material things to the status of software.
  – Molecular machines can do parallel processing (duroquinone).
  – Complex nano-assembly is certainly possible, but the watery nanoscale environment of cell biology is hostile to engineering!
  – Biology: lack of rigidity, excessive stickiness/friction, constant random motion, strong surface forces: propels the self-assembly of sophisticated structures.
  – The range of environments in which rigid nano-machines could operate is quite limited.
  – DNA itself can be used as a construction (self-assembly) machine to make programmed structures.
  – Brain-interface systems have already been built (256 neurons)
  – New polymer, nano-particle based photovoltaics $\rightarrow$ solar cells.