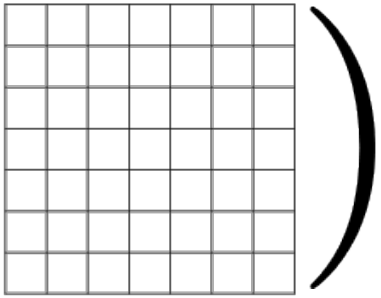
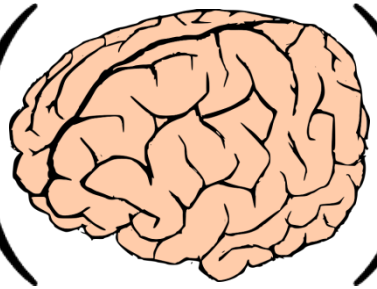
The Search for Physical Correlates of Consciousness: Lessons from the Failure of Integrated Information Theory

Φ() $\gg$ Φ()

Scott Aaronson (University of Texas at Austin)

FQXi, Tuscany, July 24, 2019

“The Hard Problem of Consciousness” (Chalmers)

Central Difficulty: Given any proposed description of the world, can “zombify” with no effect on observable phenomena

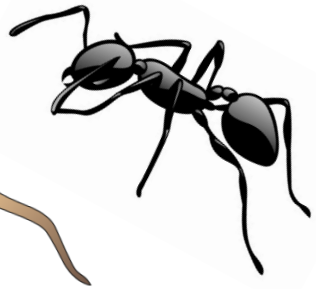
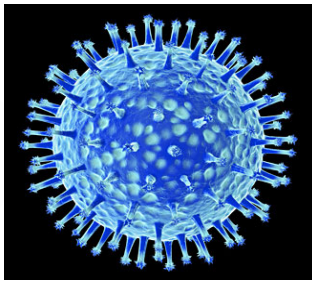
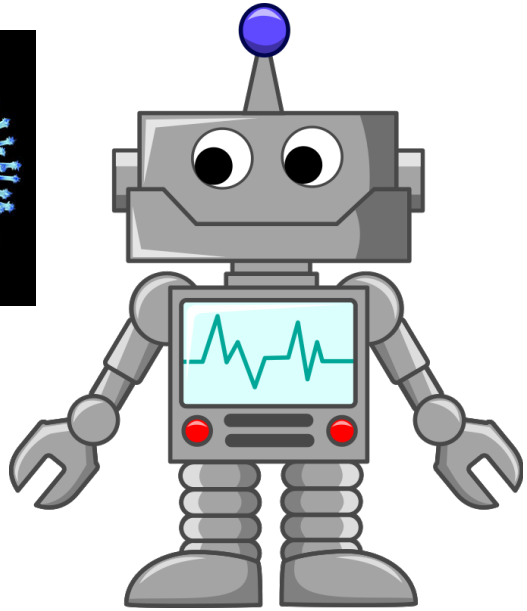
(except that now there are no “observers”…)

“The Pretty-Hard Problem of Consciousness” (A. 2014)

Give a criterion for which physical systems are and aren’t associated with consciousness

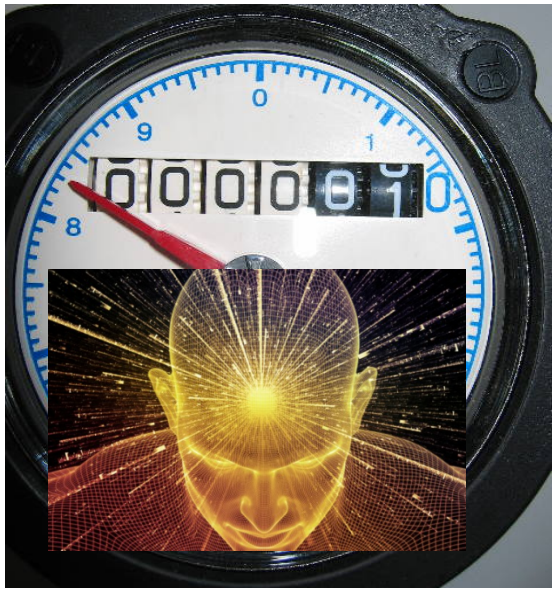
Better yet: what kind of consciousness, how much, etc.



Obvious Dilemma: How Could We Ever Test A Proposed Solution?

No independent consciousness-meter with which to calibrate predictions!



Seem to need some combination of: logic, simplicity, agreement with how people used the word “consciousness” in relatively clear-cut cases

IIT: Integrated Information Theory
(Tononi 2004)

Proposed solution to the Pretty-Hard Problem

Lists 5 "axioms of experience" (intrinsic existence, composition, information, integration, exclusion)

Then gives a **numerical measure Φ of "information integration,"** to quantify how conscious a system is

The definition of Φ is claimed to follow from the axioms. **Nothing resembling a derivation is ever given** (in any case, the definition has often changed)

But let's set that aside and see one definition…

One Definition of Φ

Given a finite set S (say {0,1}), we consider a system with an **initial state** $x=(x_1,\ldots,x_n)\in S^n$, and an **updating function** $f:S^n \rightarrow S^n$

Φ measures how well x can be partitioned into two roughly equal-sized pieces, x_A and x_B, such that calculating $(y_A,y_B)=f(x_A,x_B)$ induces little "cross-dependence" between the A and B parts

$$\mathrm{EI}(A \rightarrow B) := H(y_B \mid x_A \text{ random}, x_B)$$

$$\Phi(A,B) := \mathrm{EI}(A \rightarrow B) + \mathrm{EI}(B \rightarrow A)$$

$$\Phi := \Phi(A,B) \text{ s.t. } \frac{\Phi(A,B)}{\min\{|A|,|B|\}} \text{ minimized}$$

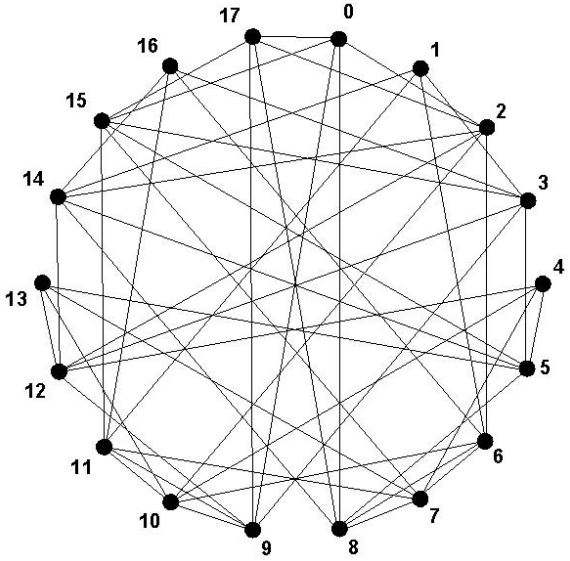
Observations

Not clear what to do if multiple (A,B)'s achieve the minimum—but we won't worry about it

$$0 \leq \Phi \leq n \log |S|$$

Φ will be close to 0 if f splits $x=(x_1,\ldots,x_n)$ into two weakly-interacting "hemispheres"

For Φ to be large: the graph of dependencies among the x_i's should be what computer scientists call an "expander"



My "Counterexample"

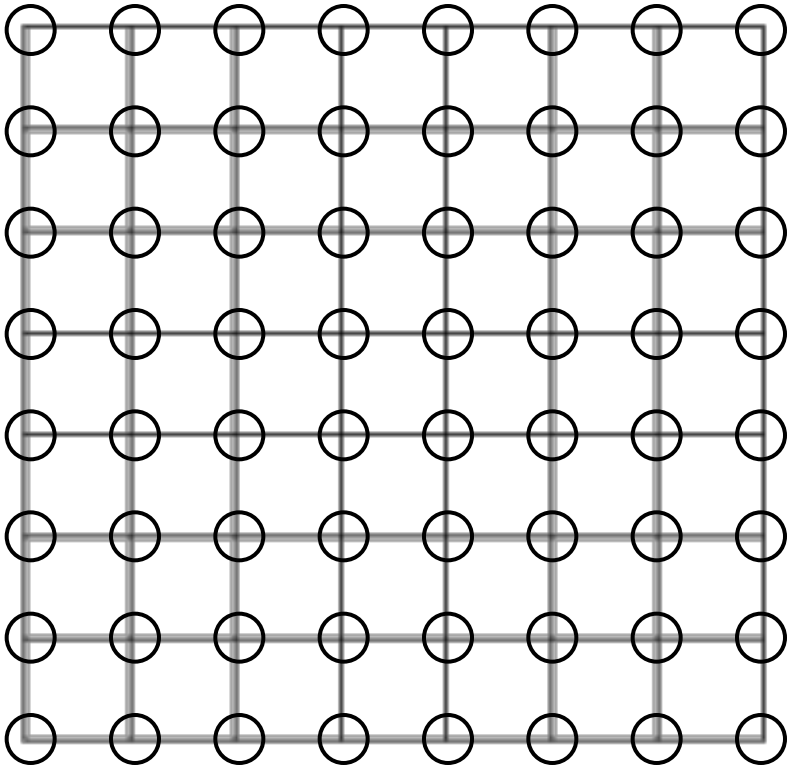
Expander graphs appear constantly in CS, for reasons having nothing obviously to do with intelligence or consciousness (e.g., error-correcting codes)!

For concreteness, consider the Vandermonde transformation over the finite field with p elements:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1^0 & \cdots & 1^{n-1} \\ \vdots & \ddots & \vdots \\ n^0 & \cdots & n^{n-1} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

For a slight variant, can prove $\Phi \geq \frac{n}{2} \log p$

Even Simpler Counterexample



2D grid of XOR gates

$$\Phi \approx c\sqrt{n}$$

By letting be n large enough, could easily make this

$$>> \Phi_{\text{HumanBrain}}$$

So, is this XOR-grid to humans as humans are to bacteria??

Tononi's Response

"Yes, the XOR-grid **does** have superhuman consciousness! Who are you to say otherwise? You're privileging your personal intuitions over our best scientific theory, IIT!"

The Problem

In testing a proposed solution to the Pretty-Hard Problem, **what do we have to go on**, besides intuitions like "humans are more conscious than blank walls"?

If a solution gets those cases wrong, then what's left for it to get right?

Crucially, even Tononi seems to agree—e.g., when he uses Φ(cerebrum) >> Φ(cerebellum) as evidence in **favor** of IIT!!

Lessons

Any theory of the form “sufficient complicatedness / interconnection / etc. ⇒ consciousness” is **doomed to failure**

Not just some technical problem with the details of IIT

Any proposed solution to the Pretty-Hard Problem must **constantly** be checked against reductio ad absurdums

What Are Other Possible Necessary Conditions for Consciousness?

Intelligent behavior (passing some sort of Turing Test?)

Unpredictability to outside observers, ability to surprise

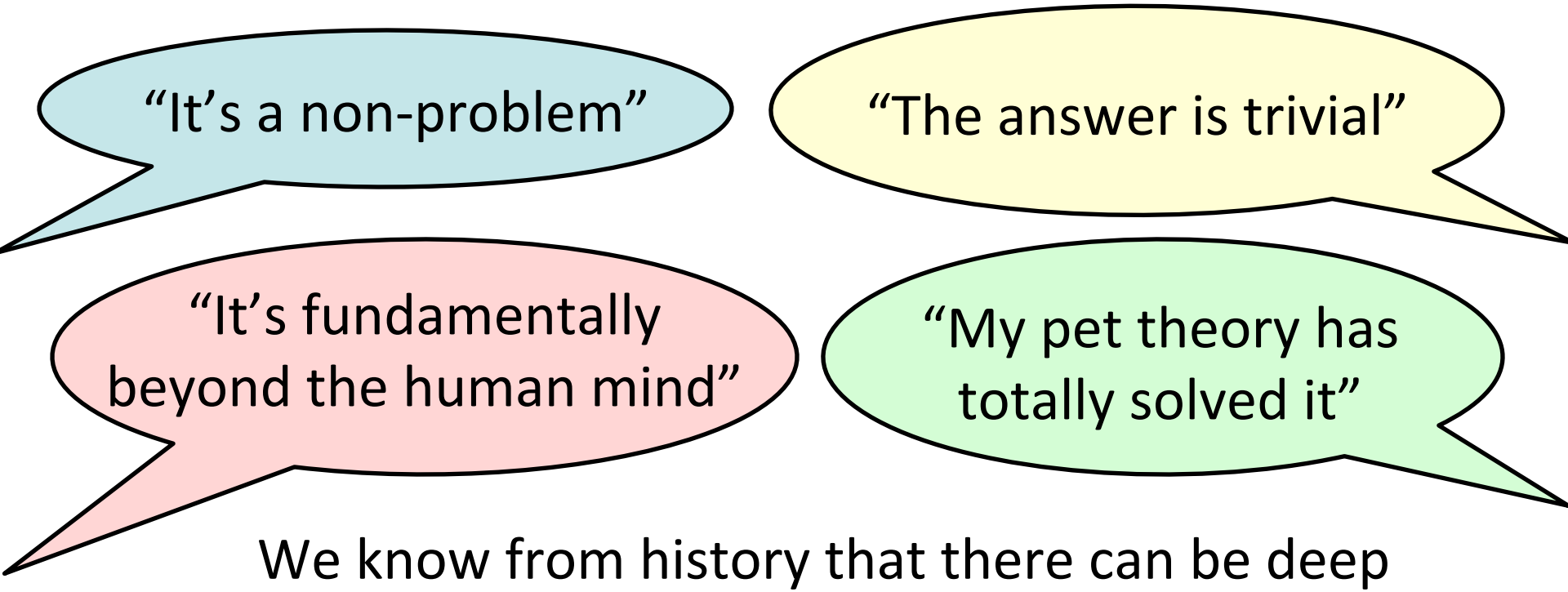
"Not being a giant lookup table or Boltzmann brain"

"Full participation in the thermodynamic Arrow of Time"
(Constantly amplifying microscopic degrees of freedom into permanent records)

Cf. my "Ghost in the Quantum Turing Machine" essay: arXiv:1306.0159

Concluding Thought

If **P** vs. **NP** or quantum gravity were treated like the Pretty-Hard Problem of Consciousness…



We know from history that there can be deep things to say, even when it will be centuries before anyone thinks to say them…