Design Ceilings on Enhancing the Mind

Susan Schneider

NASA/Blumberg Chair in Astrobiology, NASA and Library of Congress

Director, AI, Mind and Society ("AIMS") Group
The University of Connecticut

SchneiderWebsite.com

MIND DESIGN

The evolution of the brain was constrained by environmental and metabolic demands.

AI-based brain enhancement technologies have opened up a vast design space, offering novel ways to explore the space at a rate much faster than biological evolution.

Mind design -- a type of intelligent design. But we humans, not God, are the designers.

Businesses like *Facebook* and *Google* will try to engage in mind design… What could go wrong?

Today: the idea of "mind-machine merger."

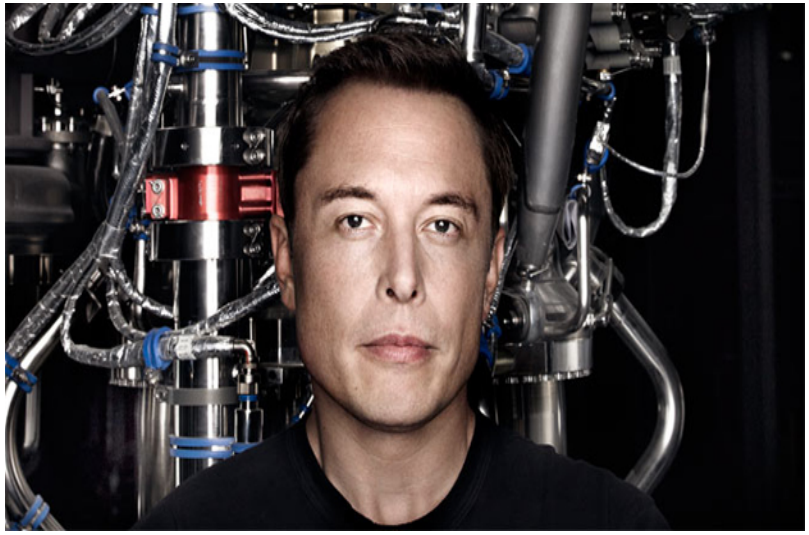
Mind-Machine Merger

Elon Musk: humans can escape being outmoded by AI and keep up with superintelligence by "having some sort of merger of biological intelligence and machine intelligence."

Neuralink: eventual implantable chip allowing data from your brain to travel wirelessly to one's digital devices.

Neural prosthetics: DARPA, Ted Berger lab, big pharma, Facebook, etc.

Today: Is attempting to "merge with AI" a sensible path toward human flourishing?



"Merging with AI"

Transhumanist trajectory:

21st century unenhanced human → significant "upgrading" with cognitive and other physical enhancements→ posthuman status → "superintelligent AI" (uploaded on the cloud).

N*ote: transhumanists (Bostrom, Kurzweil, Musk) tend to say these enhancements involve replacing parts of the brain with AI components.*

"Fusion Optimism"

Seems futuristic, but important:

- **Future of humanity**
- **Nature of the mind—do consciousness and mindedness transcend the brain?**
- **AI regulations/cyberpunk dystopias**

Today: Two Design Ceilings on Human Intelligence Augmentation

- The first design ceiling arises if microchips fail to underlie conscious experience — let's call this the "consciousness ceiling." (AIs wouldn't have this limitation. They can outpace us.)
- This second ceiling, in contrast, involves the survival of the self. This "self ceiling" is a point beyond which the person who attempts to enhance is no longer the same individual as before, for the procedure causes that individual who sought enhancement to cease to exist.
- There's more potential ceilings…

Potential Consciousness Ceiling

***Artificial You*: Wait and see approach to machine consciousness**

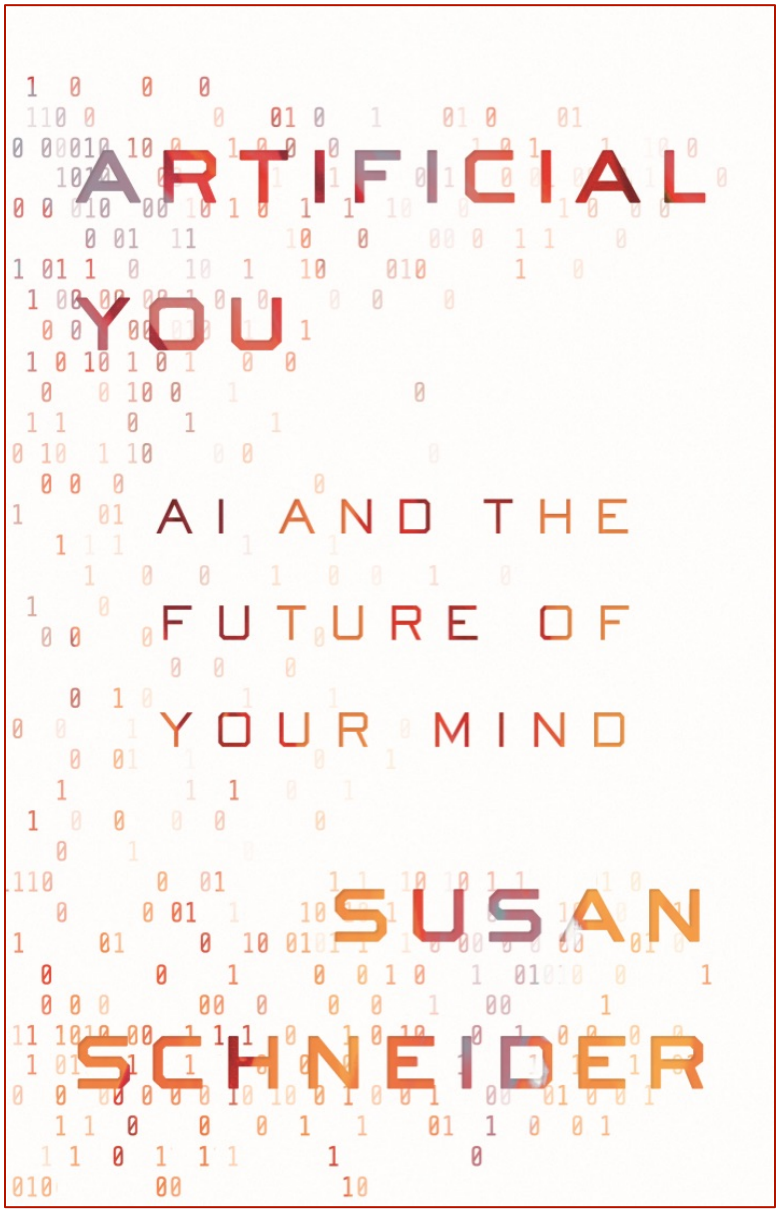
If brain chips can't underlie consciousness, merging with machines would be a BAD IDEA. Zombification.

Reply: just enhance the brain's conscious computations then.

Problem: Working memory and attention —attentional bottlenecks, etc.

We will learn whether there's a consciousness ceiling as we begin to replace parts of brain responsible for consciousness with neural prosthetics.

See my NYT op-eds.



The Self Ceiling: What are You?

Suppose you go to a Center for Mind Design…

In order to understand whether you should enhance yourself, you must first understand what you are to begin with.

What is a person?

Would you continue to exist or would you have been replaced by someone or something else?

Longstanding and controversial issue in the field of metaphysics…

Metaphysics of everyday objects—espresso machine

Essential properties

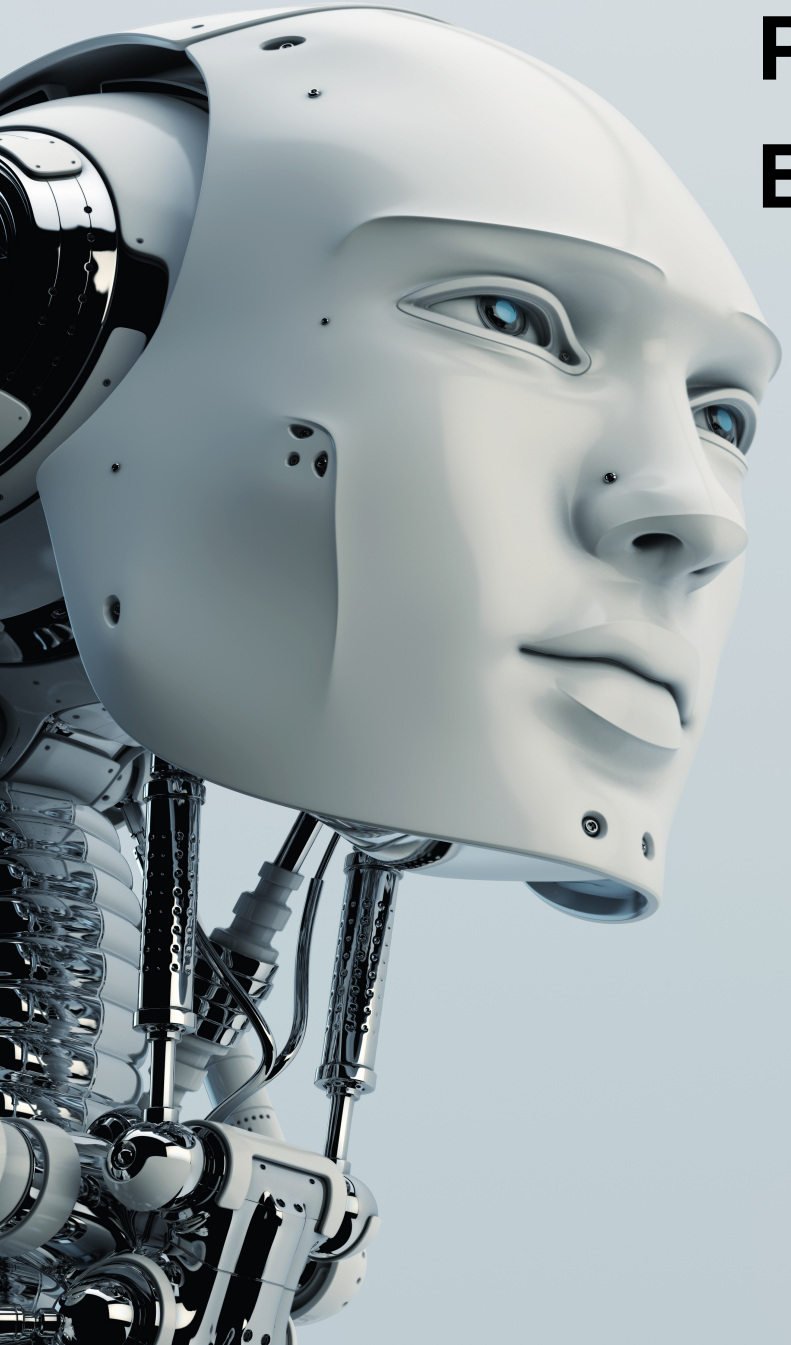
Even if "enhancement" brings such goodies as superhuman intelligence and radical life extension, it must not involve the elimination of any of your essential properties.

The "merger" wouldn't work. The sharper mind would be experienced by someone else!

Perverse realization of AI technology, even if tech unemployment and control problem are solved.

Some Theories of the Nature of the Person/"Personal Identity"

- **Brain-based materialism**: you are essentially your brain (and perhaps elements of your body).
- **The psychological continuity theory**: you are essentially your memories and ability to reflect on yourself (Locke) and, in its most general form, you are your overall psychological configuration, what Kurzweil refers to as your "pattern."



Patternism (Kurzweil, Bostrom)

What is essential to you is your computational configuration: the sensory systems/subsystems your brain has (e.g. early vision), the association areas that integrate these basic sensory subsystems, the neural circuitry making up your domain-general reasoning, your attentional system, your memories, and so on. Together these form the algorithm that your brain computes.

Some Theories, Cont.

- **The soul theory:** your essence is your soul or mind, understood as a nonphysical entity distinct from your body.
- **The no-self view:** the self is an illusion. There are bundles of impressions but there is no underlying self (Hume). There is no survival because there is no person (Buddha, Parfit).

Should you Merge with AI?

Each of these views of personal identity has its own implications about whether to attempt to merge with AI.

Brain-based materialism: you can't change substrates. You die if you replace too much of biological brain.

Neurotechnology should develop biological brain enhancements and minimal AI-enhancements that do not *replace or damage* key parts of brain.

- Soul theory: your decision to enhance would seem to depend on whether you have justification for believing that the enhanced body would retain your soul or immaterial mind.

The No Self View

- No Self View: you don’t merge with AI because there is no ”you.”
- Still, you may strive to enhance. For instance, you might find intrinsic value in adding more superintelligence to the universe—you might value life forms with higher forms of consciousness and wish that your “successor” be such a creature.

Psychological Continuity View/Patternism

The reduplication problem: you can make many copies of an informational pattern. Which one is you?

Your pattern is not *sufficient* for identity over time.

Add additional requirement?

- Spacetime worm suggestion
- Rules out uploading

Patternist Reply

- One could still merge with AI through a series of gradual, but cumulatively significant enhancements that added AI-based components inside the head, slowly replacing neural tissue. This wouldn't be uploading because one's thinking would still be inside the head, but the series still amounts to an attempt to transfer one's mental life to another substrate.
- So humans **can** merge with AI.

More Problems

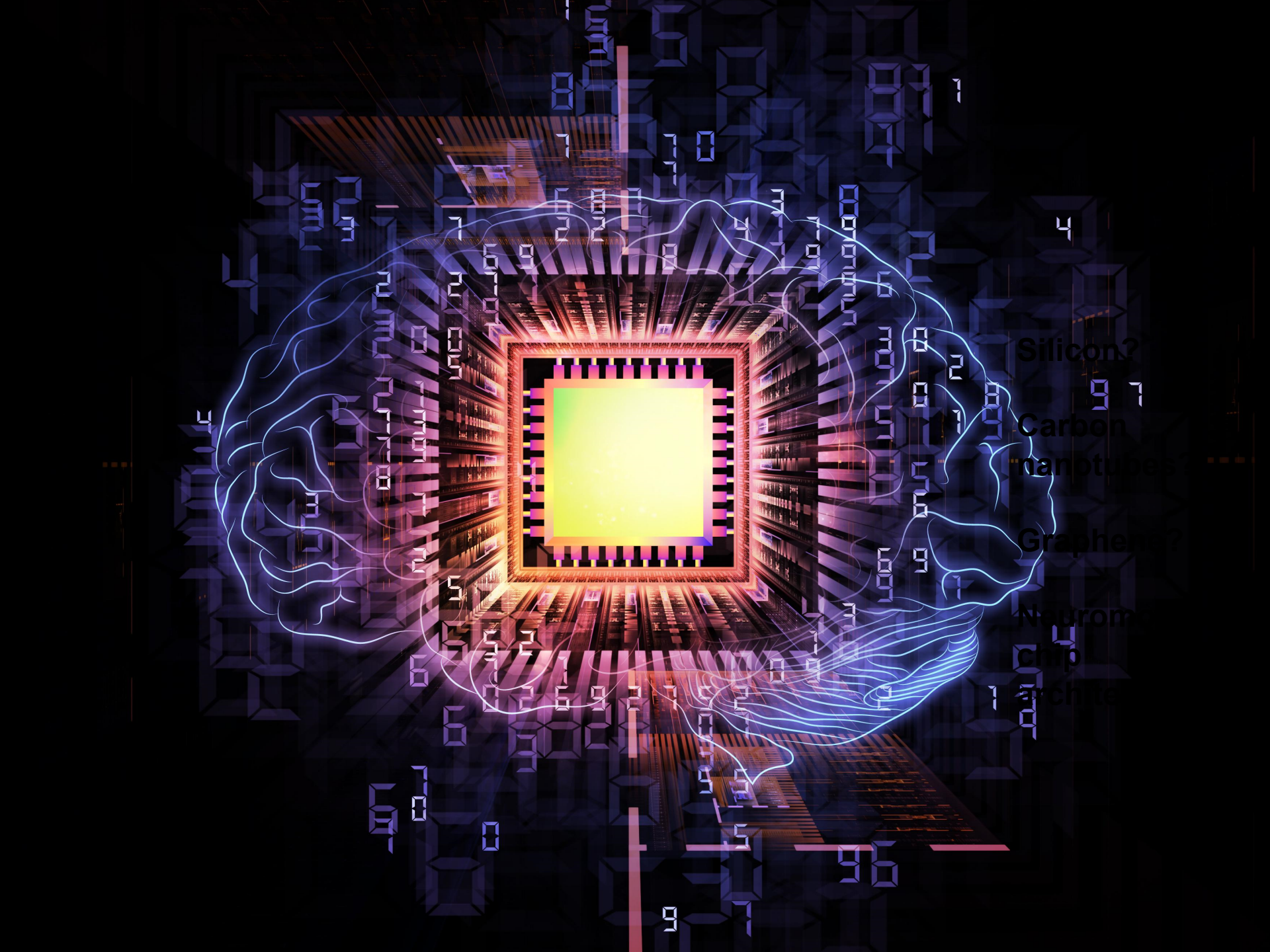
- That helps, but it rules out uploading!
- Further, we don't know when a pattern continues vs. when it ends.
- Maybe deleting a few bad chess-playing habits is kosher, but what about more serious mindsculpting endeavors, like adding several brain chips that give you new cognitive abilities.
- In addition, are you willing to bet your life that the brain view is wrong?

Known Unknowns

- The possibility of a self-ceiling is daunting. The nature of the self is so controversial, we don't know if there's a self ceiling. And there's no way to test competing theories.
- People might enhance anyway. But each person enhancing may be ending their life in the process.
- Note: the mind is not a program – those are abstract entities

A Path Forward

- Watch results of projects using implanted chips to replace parts of brain responsible for consciousness – will tell us about whether consciousness can be implemented by another substrate.
- If you ever stroll into a Center for Mind Design, stick to enhancements that are compatible with *all* the theories of personal identity you care about.
- Public engagement: The public is already confused about the nature of the self and conscious AI. FDA regulation/genetic counseling example.
- Even enhancements that merely involve the rapid or even gradual replacement of parts of one's brain, without even enhancing one's cognitive or perceptual skills, may be risky. (The brain may be essential to the self.) Stick to biological fixes, not implantable chips that require removal of neural tissue.
- Perhaps seek limited integration with AI through external devices or implants that do not remove existing brain tissue, but watch for breaks in psychological continuity.


Silicon?
Carbon
Graphene?
Neuromo
Chip
Technic

Two Design Ceilings on Human Intelligence Augmentation

- The first design ceiling arises if microchips fail to underlie conscious experience — let's call this the "consciousness ceiling." (AIs wouldn't have this limitation.)
- This second ceiling, in contrast, involves the survival of the self. This "self ceiling" is a point beyond which the person who attempts to enhance is no longer the same individual as before, for the procedure causes that individual who sought enhancement to cease to exist.
- Because the nature of the self is so controversial, we don't know if there's a self ceiling. Nor do we know how high, or low, a self ceiling would be situated.


Hasta La
vista,
baby.
www.quotesworthrepeating.com

2. AI Companies Might Cheap Out

The properties that give rise to sophisticated information processing (and which AI developers care about) may not be the same properties that yield consciousness.

- *Alpha go*.
- MP3 ex.
- Global workspace vs. “hot zone”

3. PR Nightmares

- **Developing a conscious system could lead to accusations of robot slavery and other PR nightmares, including demands to ban the use of conscious AI in the very areas the AI was developed to be used in.**
- **Consciousness engineering:**
  - **Dialing it in or out.**

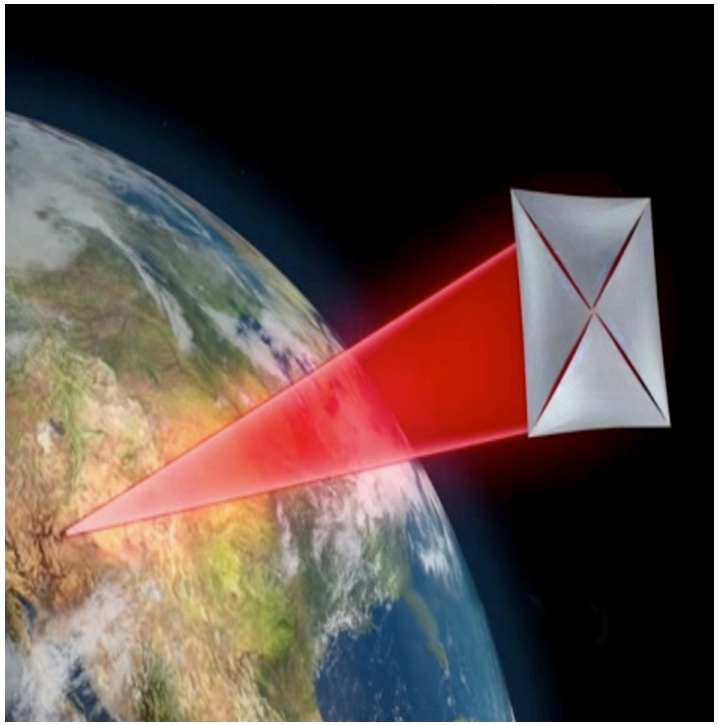
Engineering Consciousness into Machines

- On the other hand, maybe the public will demand conscious AI companions.
- Maybe consciousness aids in AI safety, so we strive to build conscious AI
- Neural prosthetics
- Moonshot projects- functional isomorphs/ uploads
- (*De Grasse Tyson discussion*).

Westworld: HBO

Breakthrough Starshot

- Milner/Hawking Starshot space mission to send light-sail ships at 20% of the speed of light to *Alpha Centauri.* (Est. 20 years to arrive.)
- Caleb Scharf, Edwin Turner, Olaf Witkowski and myself explore use of AGI for interstellar travel. Project title: “Sentience to the Stars.”
- Exoplanets are habitable, but are they inhabited? Origin of life debate.



Upshot: import of AI consciousness goes well beyond familiar robo-rights issues.

We need to know whether AI will be conscious.

How to test? No neuroanatomy. The “black box” issue with deep learning systems.

Some provisional tests…

The ACT Test (with Edwin Turner)

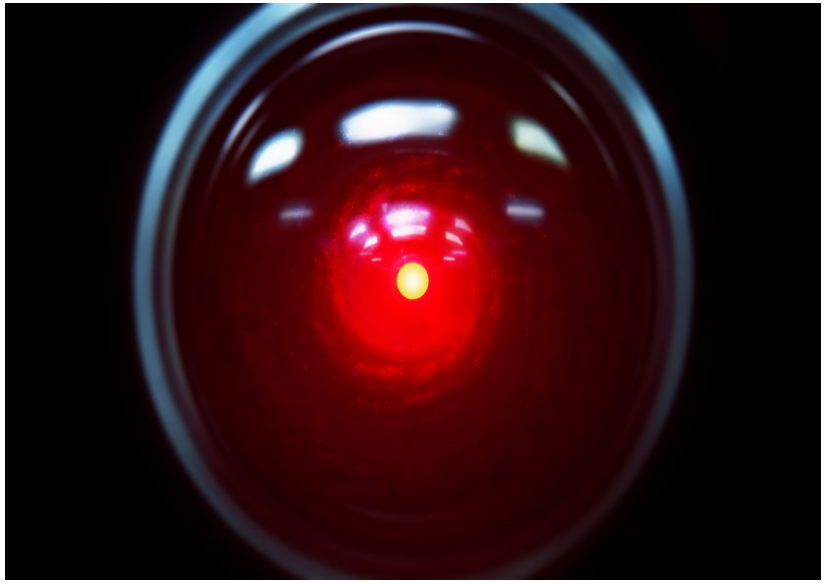
- ACT Test: nearly every adult can *quickly and readily* grasp notions based on this felt quality of consciousness. *Freaky Friday*, reincarnation, out of body, etc.
- Test AI to see if it grasps such notions.
- Example questions

Blade Runner – Warner Brothers

- ***Superintelligence? "boxing in" techniques.***
- ***Advantages:***
  - Doesn't require looking into the 'black box" of a complex AI.
  - Doesn't require verdict on neural basis of consciousness in humans.
  - AI will have a sort of "self" concept
- ***ACT is sufficient only.***

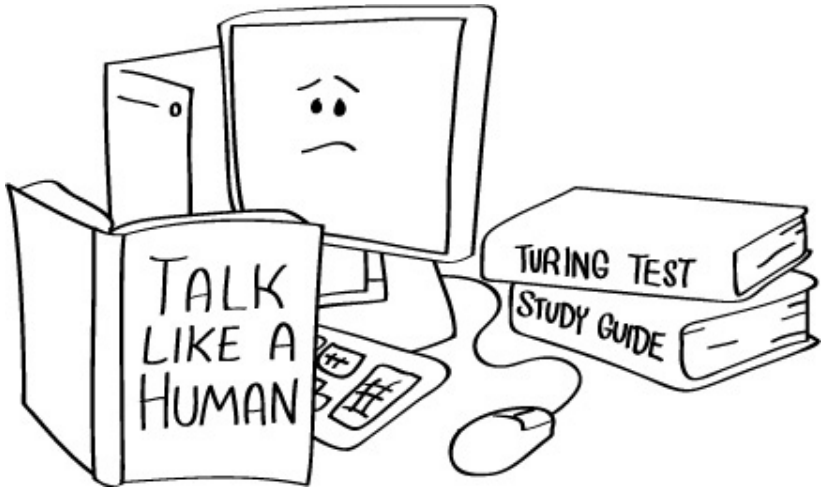
HAL 9000

- The death of the mind of the fictional HAL 9000 AI in Stanley Kubrick's *2001: A Space Odyssey* provides an illustrative example.
- HAL neither looks nor sounds like a human being. Nevertheless, the *content* of what HAL says as it is deactivated conveys a powerful impression that it is a conscious being.
- (Note: a version of ACT could apply to systems that do not have natural language interface as well. Think of octopus behavior.)



Is ACT Just Another Turing Test?

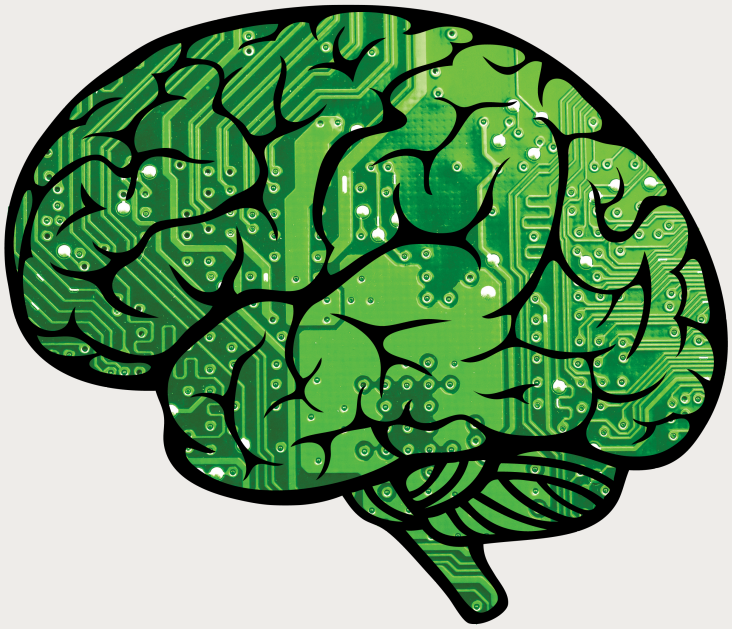
- Like Turning's test, ACT is entirely based on behavior, and like Turing's, it could be implemented in a formalized question and answer format. (An ACT could also be based on an AI's nonlinguistic behavior or the behavior of a group of AIs.)
- But an ACT is also quite unlike the Turing Test, which was intended to bypass any need to know what was transpiring inside the machine.
- By contrast, ACT is intended to do *exactly the opposite*; it seeks to reveal a subtle and elusive property of the machine's mind.
- A machine might fail the Turing test, because it cannot pass for a human, but pass an ACT, because it exhibits behavioral indicators of consciousness.

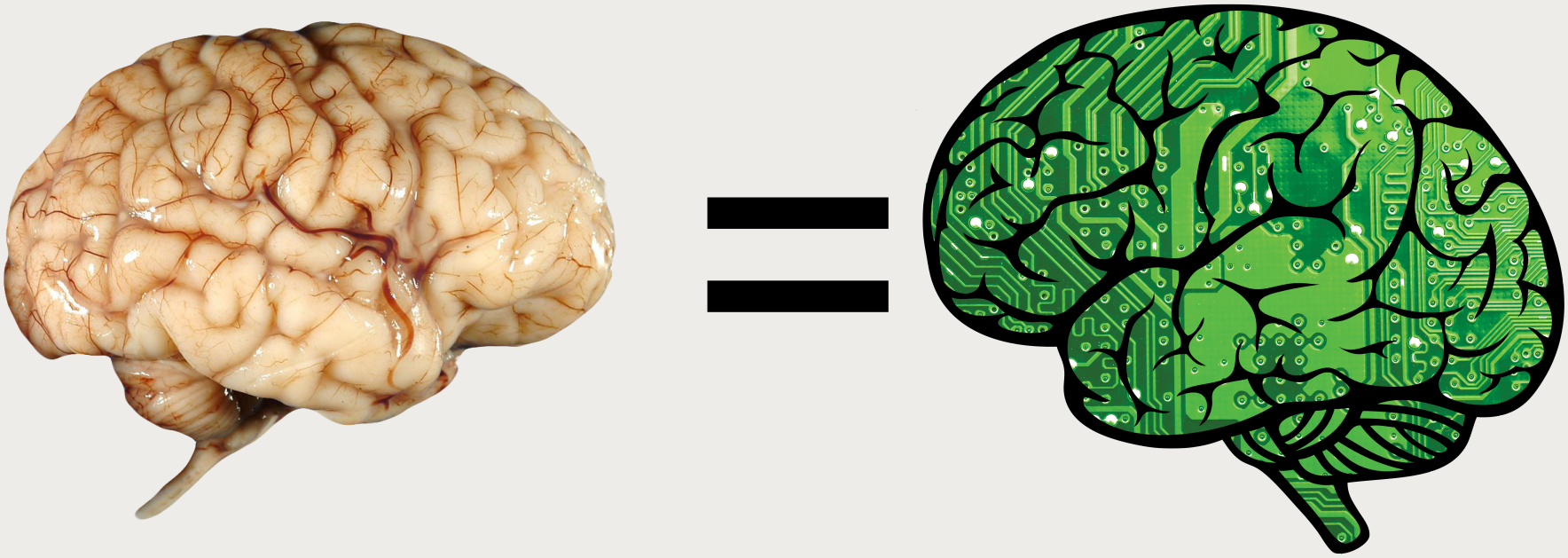



Dr. Peng Zheng
NEURAL DIAGNOSTIC
INITIALIZING
HEART RATE
HR
75
F.ZHU APCH50382
SC02
SC03
HISTORY
PULSE
75/MIN
SP02
97%
ABP

iBRAIN

≠





Ongoing Work

- This is my “Chip Test” (see my Ted talk).
- Stanford Research Institute
- Help determine what level of chip design may be necessary to facilitate consciousness (in that area). (Informs philosophical debates on functionalism.)
- “Marker” for AI consciousness only.
- Comparison of results with other tests: IIT and ACT

Conclusion

Soon, unenhanced humans may no longer be the most intelligent beings on Earth. Perhaps the greatest alien intelligences will be postbiological.

Would these AIs be conscious?

-Import

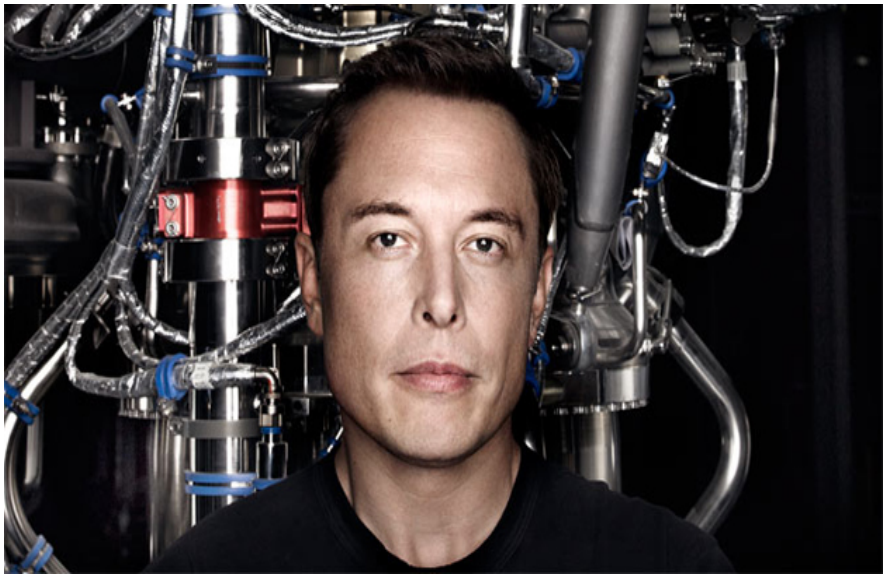
-“Wait and see” approach

-”Consciousness Engineering”

-Provisional tests

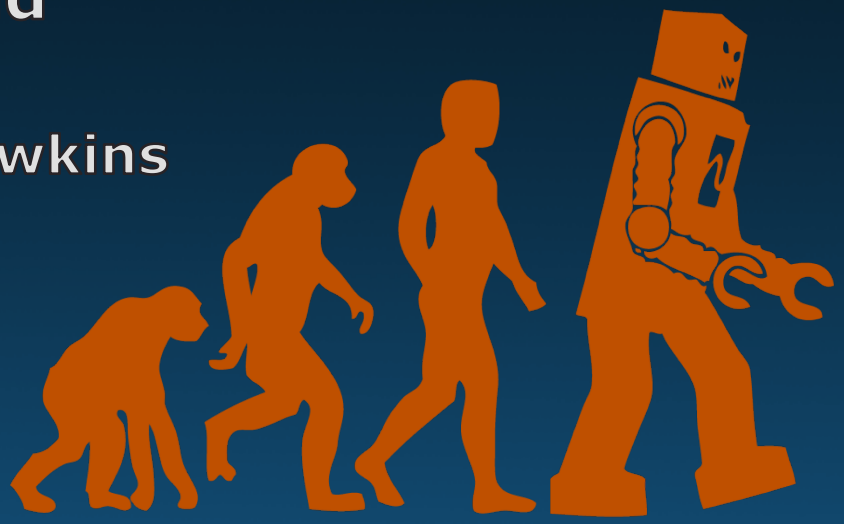
Consciousness and Ethics

- There are other reasons why we need to test for AI consciousness as well:
- Wrongly claiming AI is conscious may lead to:
  - Giving moral status and legal protections to beings that aren't sentient.
  - Mistaken ideas about "human-machine merger" (Musk/Kurzweil). We cant fully merge with AI without loss of consciousness; we can only replace parts of brain not responsible for consciousness. (May limit that AI safety strategy.)
- Wrongly denying that AI is conscious would lead to their suffering and enslavement.
- Not getting AI consciousness right can ruin an AI program. If you don't think your AI product is conscious, and you are wrong, it may not be possible to put it to use (ethics, volatility, public outcry.)



"It's not obvious to me that a replacement of our species by our own technological creations would necessarily be a bad thing."

- Richard Dawkins



IEEE TECH FOR HUMANITY SERIES

#IEEEConsciousMachine

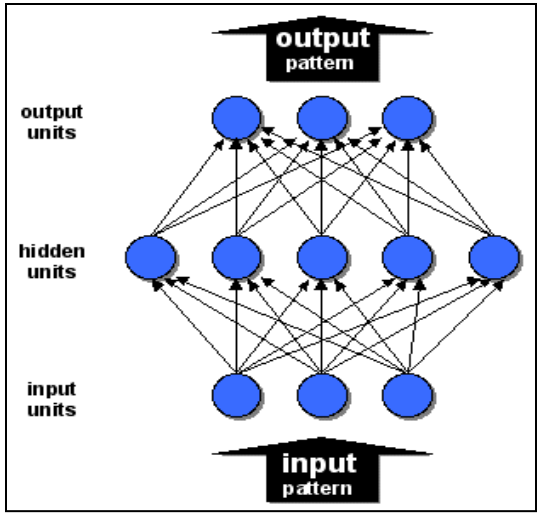
IEEE

We Need Tests for AI Consciousness

- The N=1 problem in astrobiology: in seeking life in the universe we only have 1 case, Earth. If we generalize from it, we may miss life elsewhere. We could be outliers. So avoid defining life.
- We don't have AGI yet. To early to suggest a necessary condition for AI consciousness that all AIs must share. Seek sufficient conditions…
- My "chip test." This can tell us about the substrate, but it has limited applicability for other architectures (see my *Ted talk* or *Nautilus* piece.) Reverse chip test.
- In Princeton, I am writing further tests to determine if AI is conscious with astrophysicist, Edwin Turner.
- E.g., the architecture of the machine could be too alien or opaque, and a behavior based approach is useful.
- Use one test to check results of another test, e.g., IIT.

Deep Learning

- A very simple "connectionist"/deep learning network (from Schneider and Katz).



- **Meant to be roughly brainlike. (if modeling thought, each circle (or "unit") represents either a single neuron or a group of neurons.)**
- **As illustrated, computation flows upward, with the smaller arrows specifying connections between units.**
- **Each hidden or output unit carries a numerical activation value which is computed given the values of the neighboring units in the network, according to a function.**
- **The input units' signals thereby propagate throughout the network, determining the activation values of all the output units.**
- **Actual models of perceptual and cognitive functions are far more complex, exhibiting multiple hidden layers and feedback loops.**
- **Even Alpha Go uses other resources (e.g., decision trees). Hybrid models.**

Could You Merge with AI?

Susan Schneider

NASA/Blumberg Chair in Astrobiology, NASA
Distinguished Scholar, Library of Congress
Director, AI, Mind and Society ("AIMS") Group
The University of Connecticut
SchneiderWebsite.com

gn ceilings on the human mind. 2 today (from my NYT).
cal enhancements, (uploading, replacing the parts of the brain w
feasible if AI components do not support consciousness. AI cons
ay not even be compatible with the persistence of the self or per

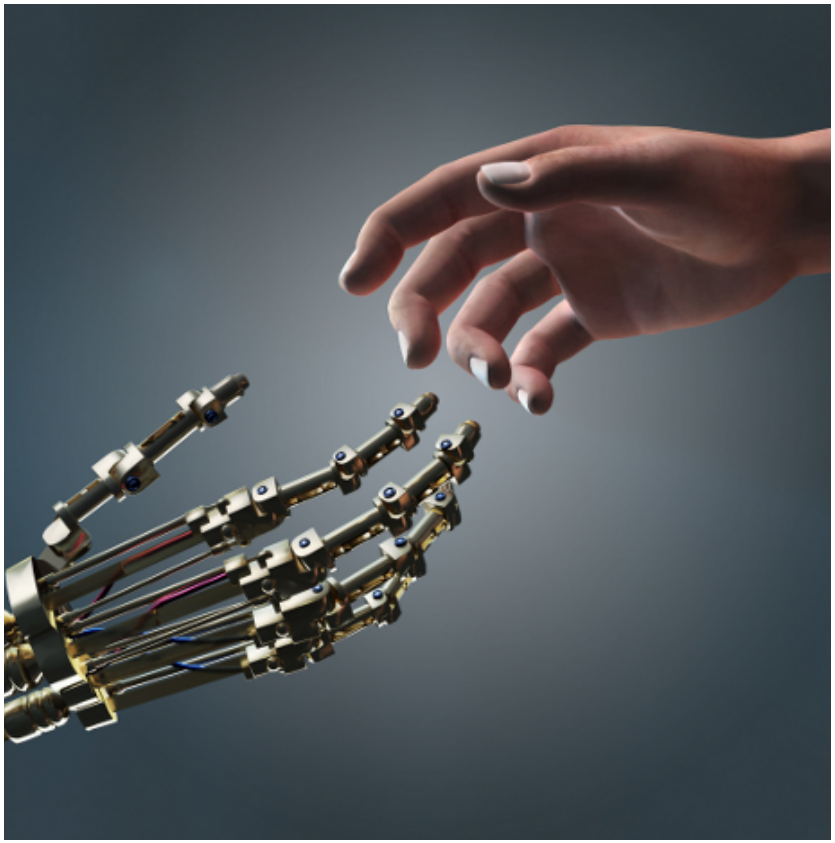
Mind Design, cont.,

This is humbling…

We are not terribly evolved.

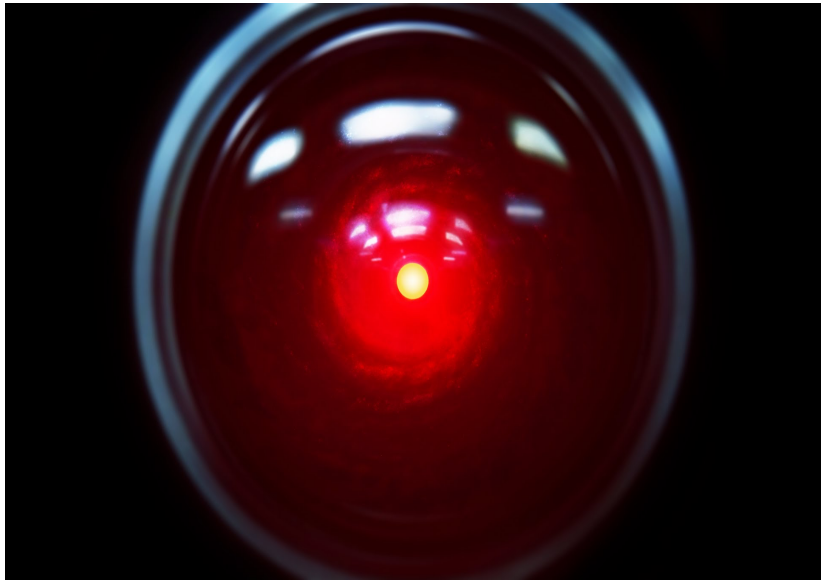
The alien superintelligence in the Carl Sagan film, *Contact*:

“You're an interesting species. An interesting mix. You're capable of such beautiful dreams, and such horrible nightmares.”



Some Perverse Realizations of AI

- Situations in which AI leads to our own suffering, or to the exploitation of other conscious beings.
- Types I discuss in *Artificial You*:
  - (i), overlooked situations involving the creation of conscious machines.
  - (ii), scenarios that concern the use of radical brain enhancements.



Consciousness

When you see the rich hues of a sunset or smell the aroma of your morning coffee, it feels like something to be you.

Consciousness is the felt quality of our inner experience.

Because the nature of the self is so controversial, we don't know if there's a self ceiling. Nor do we know how high, or low, a self ceiling would be situated.