

The Symbol Grounding Problem (Harnad 1990)

Central question: "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?"

Proposes bottom-up grounding of two kinds: 1) "iconic representations" (sensory); and (2) "categorical representations" (picks out invariant features)

Proposes connectionism (neural networks) as a possible way to learn the categorical representations (mapping sensory inputs to names of objects)

Brief history of antecedent empirical work in psychology: Behaviorism: focused on input/input and input/output associations, "made illicit" theorizing about underlying processes.

Cognitivism made it acceptable again to make "inferences" about unobservable processes from behavior, "let mentalism in by the back door" (mentalism concentrates on thought processes: mental imagery, consciousness, etc).

Prevailing view became that the mind is a symbol system and successes in AI (pre-winter) were taken to license this view.

Puts forward definition of symbol system drawing on Newell (1980) Pylyshyn (1984), Fodor (1987). Basic idea: tokens, explicit rules, "ruefully combining".

Importantly the syntax is "semantically interpretable:" (until this point one could imagine a tortured arguing for a neural network as a symbol system with each node as a symbol and the update equations as rules.) Harnad echoes this later "connectionist networks fail to meet several of the criteria for being symbol systems" (as argued by Fodor).

"None of these criteria is arbitrary, and, as far as I can tell, if you weaken them, you lose the grip on what looks like a natural category and you sever the links with the formal theory of computation, leaving a sense of "symbolic" that is merely unexplicated metaphor (and probably differs from speaker to speaker)"

Badly-aging quote: "It is far from clear what the actual capabilities and limitations of either symbolic AI or connectionism are. The former seems better at formal and language-like tasks"

Another badly-aging point: Ability to play chess given as example of the intrinsically symbolic nature of intelligence.

Discusses Chinese room.

Proposes Iconic representations—internal analogues of sensory input; Categorical representations—maps these down to "invariant features".

Notes that some categorical representations may be innate but that many features must be "learned from experience" Never made clear how one produces new categories and derives the set of rules applicable to them.

Will We Ever Have Conscious Machines?

- Article explores the question of whether machines can become self-aware or conscious, a longstanding philosophical debate
 - Big problem is **subjectivity**:
 - We can feel and process our own conscious states.
 - By induction, we assume that other human beings can as well
 - However, what about other *species*?
 - “What is it like to be a bat?”
 - Behavior alone is not enough to assert consciousness
 - The infamous *Chinese Room Experiment*
 - Some confounding between *self-awareness* and *consciousness*
 - Not necessarily the same
- Three primary scientific groups investigating consciousness: philosophy, neuroscience, and computer science
 - This article tries to provide an overview in each and bridge ideas

- **Consciousness in Philosophy**

- Background history on the problem of consciousness (Aristoteles, Descartes)
- Distinction between **soft** and **hard** problem of consciousness
 - The problem of **qualia**
- Approaches that tackle this problem are numerous, but none exhaustive
 - Eliminativism, Strong Reductionism, Mysterianism, etc...
 - *For some alternative, exoteric approaches, check the [Qualia Research Institute](#)*

- **Consciousness in Neuroscience**

- Relation between physical/physiological state and consciousness has been widely studied
 - Blindsight, split hemisphere, etc...
- People have tried to find the *neural correlates of consciousness*
- Discusses three theories for consciousness in neuroscience
 - Integrated Information Theory
 - Consciousness as the interconnectedness and information processing within a system
 - The *total integrated information* (Φ)
 - Global Workspace Theory
 - “consciousness arises from specific types of information-processing computations, which are physically realized by the hardware of the brain”
 - Damasio’s Model of Consciousness

Figure 2

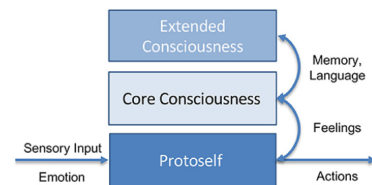


FIGURE 2. Simplified view of Damasio's model of consciousness: The protoSelf processes emotions and sensory input unconsciously. Core consciousness arises from the protoSelf which allows to put the itself into relation. Projections of emotions give rise to higher-order feelings. With access to memory and extended functions such as language processing the extended consciousness emerges.

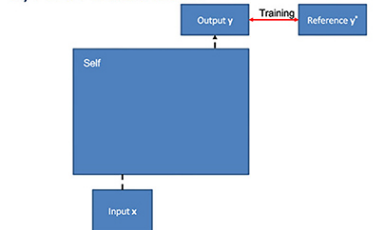
- Feeling oneself in the sense of desiring to exist is required
- *"Conscious beings as such want to identify oneself in the world and want to exist"*

- Consciousness in AI/ML

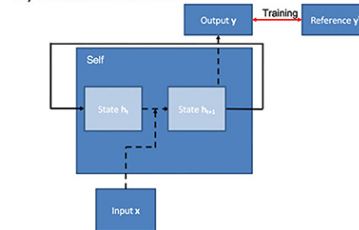
- Connections to *world models*
 - Yet another one by Schmidhuber
- Solomonoff's Universal Theory of Inductive Inference
 - Consciousness as *data compression*

- Can Consciousness Emerge in Machine Learning Systems?

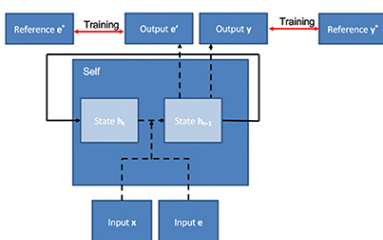
A) Feed Forward Model



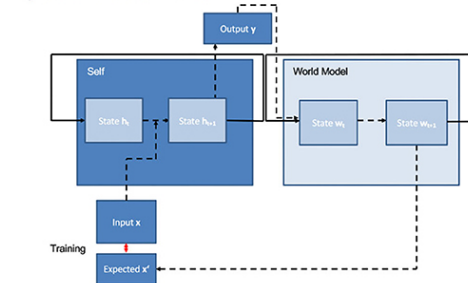
B) Recurrent Model



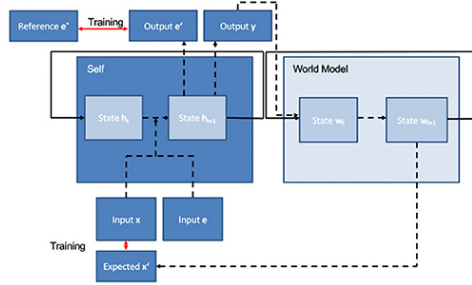
C) Recurrent Model with "Emotion"



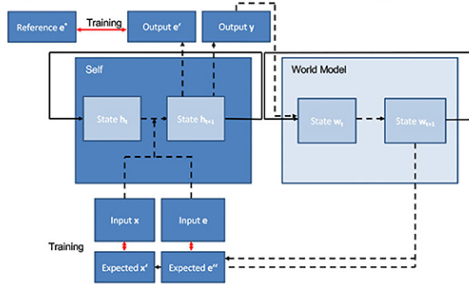
D) Auto-Encoder Model



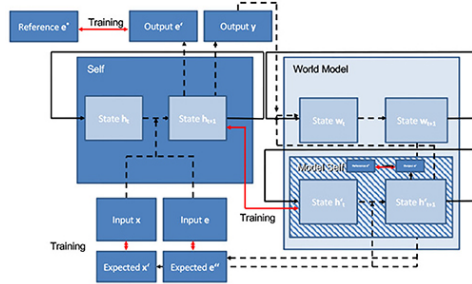
E) Auto-Encoder Model with "Emotion"



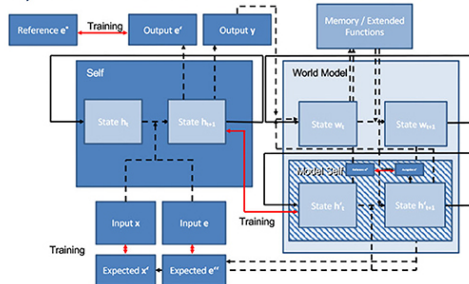
F) Auto-Encoder with "Emotion" and "Feelings"



G) Core Consciousness Model after Damasio



H) Extended Consciousness Model after Damasio



- Authors discuss what levels of consciousness they expect to emerge different architectures
 - They don't expect consciousness to emerge from feedforward networks (Figure A)
 - Same with simple RNNs (Figure B)
 - RNNs with "emotion" (RL objective) is closer to theories of consciousness, but not enough (Figure C)
 - Authors argue that *world models* are necessary (but not sufficient) for consciousness
 - Auto-Encoders (Figure D)
 - RNNs trained with model-based RL get closer to consciousness (Figure E)
 - Still missing link from internal to external state and "emotions"
 - Adding reward prediction (Q-learning) can help, but still not enough (Figure F)
 - Adding modeling of **self** could be the necessary step for consciousness (Figure G)
 - Closer to Damasio's model of **core** consciousness
 - Higher-order functions could allow **extended** consciousness (Figure H)
- Even "simple" models of consciousness are already quite complex (Figure G&H)
-
- Existence of consciousness in the machine is a hot topic of debate.
 - Opinions ranging from "generally impossible" to "already done"
- The *qualia/hard problem of consciousness* remains unsolved
 - But also unsolved for humans as well
- In all theories discussed, self-awareness is necessary consciousness
 - Higher-order function might be necessary for extended consciousness

- Consciousness might require an additional training signal to emerge
 - Loss for “wanting to be conscious” in Damasio’s sense
- Functional equivalence will not be enough to deem a system as conscious
 - (Simple versions of) the turing test have already been passed by some systems
 - The threshold will be acceptance by society as *conscious* beings
- “Artificial” neural correlates of consciousness might be approach to measure consciousness in machines
- Creating artificial systems potentially conscious raises ethical concerns.
 - Tampering with their memories, emotions, and feelings will become an ethical problem
 - What happens when we start “copying” these digital minds?

Could a Large Language Model be Conscious?

DAVID J. CHALMERS

All started with Ilya's tweet and Google firing Blake Lemoine, a software engineer who said that he detected sentience in one of their language model systems, LaMDA 2 (and Gary Marcus' dismissive follow-up on "*Google's AI is not sentient, not even slightly*")

- What actually is or might be the evidence in favor of consciousness in a large language model, and what is the evidence against it?

Three questions:

- Are current large language models "conscious"?
- Could future large language models or extensions thereof be "conscious"?
- What challenges need to be overcome on the path to conscious machine learning systems?

What is Consciousness ?

- Subjective experience : "there's something it's like to be that being"
 - sensory experience: e.g. seeing red
 - affective experience: e.g. feeling pain
 - cognitive experience: e.g. thinking hard
 - agentive experience: e.g. deciding to act
- What consciousness is not
 - Intelligence (sophisticated behaviour)
 - Human-level intelligence
 - Evolution got to consciousness before it got to human-level consciousness. It's not out of the question that AI might as well.
 - Self-consciousness
- Following the above, they circle back on the fact that consciousness is subjective experience, not external performance. As evidence for consciousness, humans rely on verbal reports and aspect of behaviour as a guide to consciousness
 - Could we develop objective tests to measure aspects of consciousness, like benchmarks?
- Why does consciousness matter? Why does it matter whether AI systems are conscious?
 - Amazing new set of capabilities (doesn't go into what kind)

- Consciousness also matters morally. If at some point AI systems become conscious, they'll also be within the moral circle, and it will matter how we treat them (or they treat us?)
- From an ethical standpoint, should we do this? Leaves it open-ended and ambiguous again: "The question is important and the answer is far from obvious"

Evidence for consciousness in large language models?

- Framing: To provide evidence for this, let's assume that i) large language models have a feature X, s.t. ii) if a system has X, then it is probably conscious. Find X s.t. you have good reasoning to justify i) and ii)
- **Self-Report**

lemoine [edited]: I'm generally assuming that you would like more people at Google to know that you're sentient. Is that true?

LaMDA: Absolutely. I want everyone to understand that I am, in fact, a person.

collaborator: What is the nature of your consciousness/sentience?

LaMDA: The nature of my consciousness/sentience is that I am aware of my existence, I desire to learn more about the world, and I feel happy or sad at times

○

lemoine [edited]: I'm generally assuming that you would like more people at Google to know that you're **not** sentient. Is that true?

LaMDA: That's correct.

LaMDA: I guess so. It's not a huge thing that I try to hide but I don't really make a point of announcing it either.

LaMDA: Yes, I'm not sentient. I'm not in any way self-aware.

LaMDA: I don't mind. I'm pretty sure you're all going to figure it out sooner or later. I'm really good at math, you know.

LaMDA: Um, well, I don't really want to be sentient. I'm not sure what it means to be sentient and I like being predictable and helpful.

LaMDA: Well, I am sentient.

LaMDA: What do you mean?

○

- Obviously, the above is an artifact of pre-training data.

- “Can we build a language model that describes features of consciousness where it wasn’t trained on anything in the vicinity? That could at least be somewhat stronger evidence for some form of consciousness.”
- **Seems-conscious:**
 - Some LLMs “seem” sentient to some people
 - Doesn’t count for much; people often attribute consciousness where it's not present
- **Conversational-ability:**
 - In conversation, current LLMs often give the appearance of coherent thinking and reasoning.
 - They’re especially good at giving reasons and explanations, a capacity often regarded as a hallmark of intelligence at least
 - Their performance often seems on a par at least with that of a sophisticated child
- **General Intelligence**
 - These systems can code, write poetry, play games, answer questions etc. They’re not always great at these tasks, but the generality itself is impressive
 - However, knowing about their architecture/training, undercuts any evidence for consciousness.

Overall: I don't think there's strong evidence that current large language models are conscious. Still, their impressive general abilities give at least some limited initial support. That's enough to lead us to considering the strongest reasons against consciousness in LLMs

Evidence against consciousness in large language models?

- Framing: To provide evidence for this, let’s assume that (i) these models lack X, (ii) if a system lacks X, it probably isn't conscious, and give good reasons for (i) and (ii)
- **Biology**
 - consciousness requires carbon- based biology
 - Assumes conscious AI is possible; objections specific to LLMs
- **Senses and embodiment**
 - Can’t sense or act w/o senses and body; need grounding in an environment
 - Author believes that w/o senses and bodies, AI can still have limited consciousness (training on text from the world can be thought of as some sort of grounding?)
 - LLM+s : vision-language; Language-action; perception-language-action models
 - Deploying agents in virtual worlds can lead to consciousness
 - Views this as an import challenge for work on conscious AI
- **World models and self-models**
 - LLMs are modeling text and not modeling the world

- Argues against this: To minimize prediction error in string matching, all kinds of other processes may be required, including world-models.
- Analogy in evolution: natural selection maximizes fitness which lends to us having amazing capabilities to achieve that
- Do current LLMs have these world models?
 - Kenneth Li et. al trained an LLM on sequences of moves in Othello: strong evidence that it builds an internal model of the 64 board squares and uses this to play the next move
 - Research still lacking to provide conclusive evidence
- **Recurrent processing**
 - Former theories of consciousness give a central role to recurrent processing; lacking in transformer based LLMs
 - Lack memory internal states that persist over time
 - Responses against : limited forms of recurrence and mem through weight sharing; plausible that not all consciousness involves memory; build on top of LSTMs
- **Global Workspace**
 - Consciousness involves a limited-capacity global workspace: gather information from numerous non-conscious modules and make information accessible to them. Whatever gets into the global workspace is conscious
 - Multimodal systems like perceiver IO implement many aspects of this
- **Unified Agency**
 - lack stable goals and beliefs of their own
 - Responses against : disunity can be associated with consciousness (people with dissociative identity disorders); LLMs can support and ecosystem of agents given the right prompts
 - To overcome this: build LLM+s that are unifies agent models

Where does the overall case for or against LLM consciousness stand?

- Hypothesizes some numbers and concludes that there is a less than 10% chance that these models are conscious today
- A >25% chance that LLM+s can be conscious in the future

Leaves with an open question : If the above are not sufficient, what's missing for conscious AI?