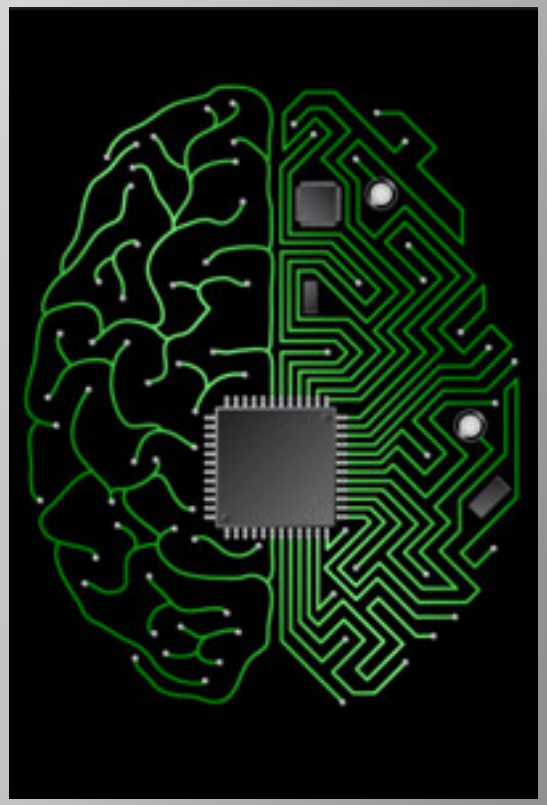
What It Is To Be Conscious: Exploring the Plausibility of Consciousness in Deep Learning Computers



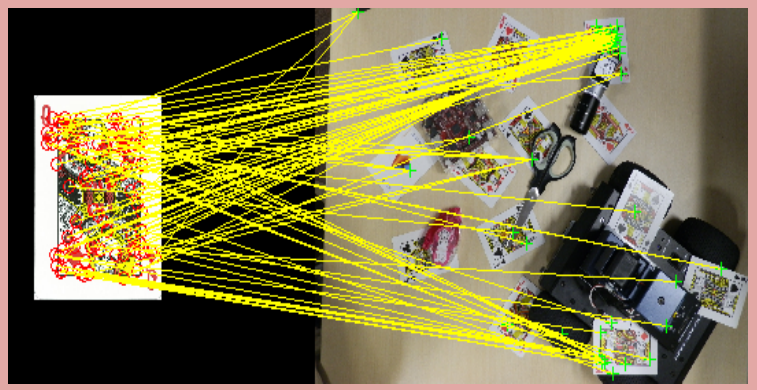
(Peter) Zach Davis

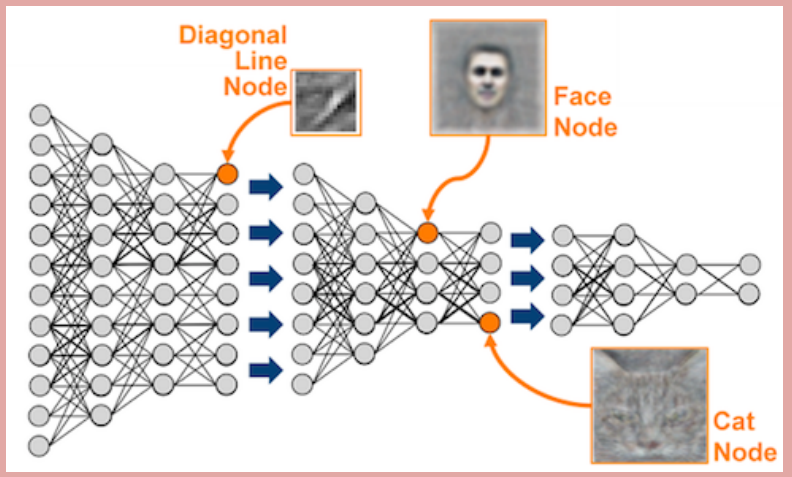
Philosophy & Computer Science ID

Advisors: Kristina Striegnitz and David Barnett

Motivation

Deep learning computers are amazing!



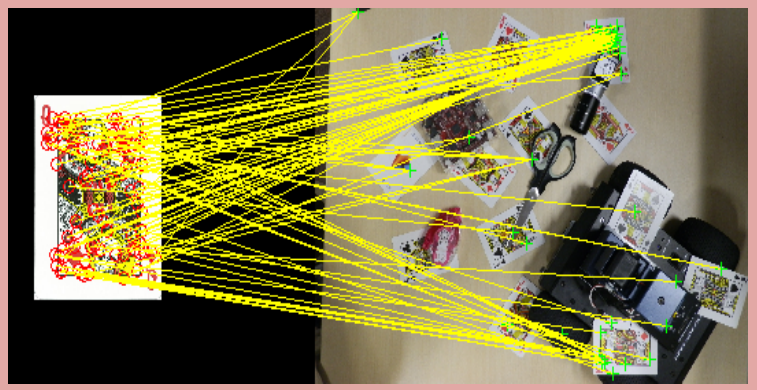


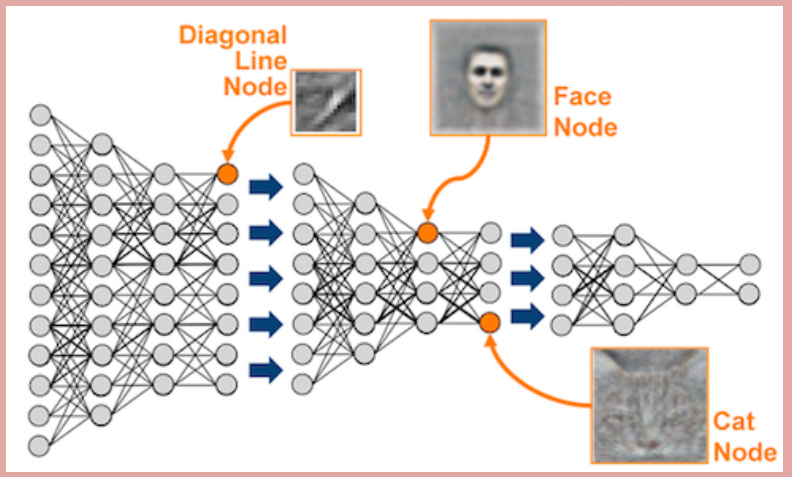
Motivation

Deep learning computers are amazing!

But…

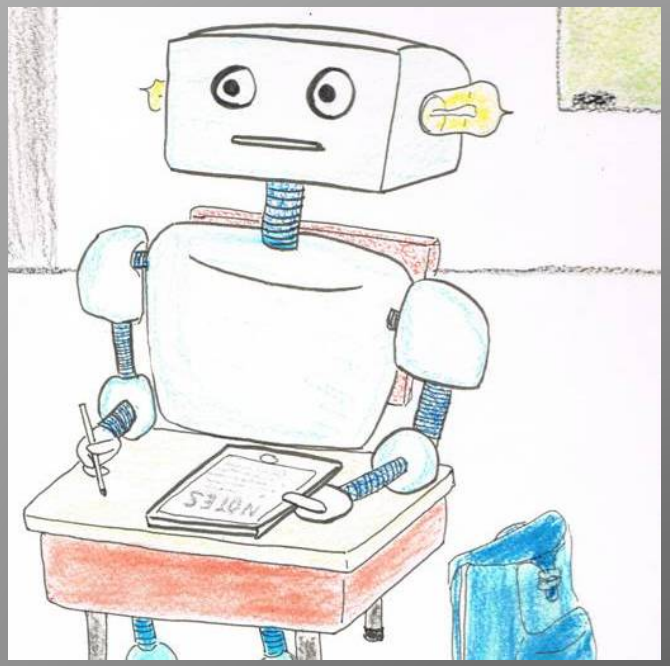
No consensus on their consciousness





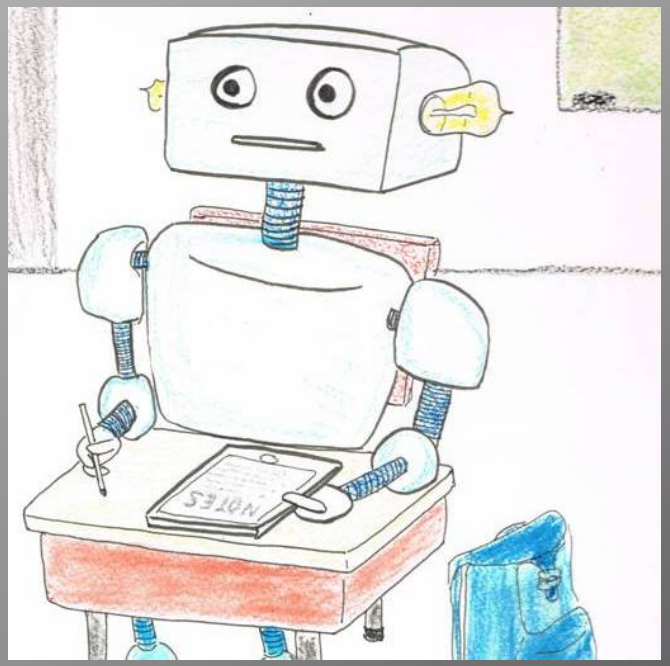
Machine Learning

- Derive generalizations from examples



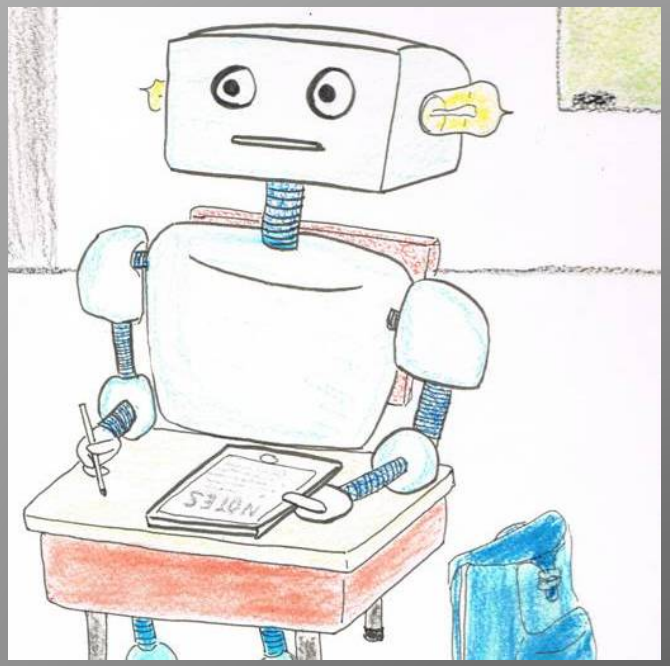
Machine Learning

- Derive generalizations from examples
- Similar to humans



Machine Learning

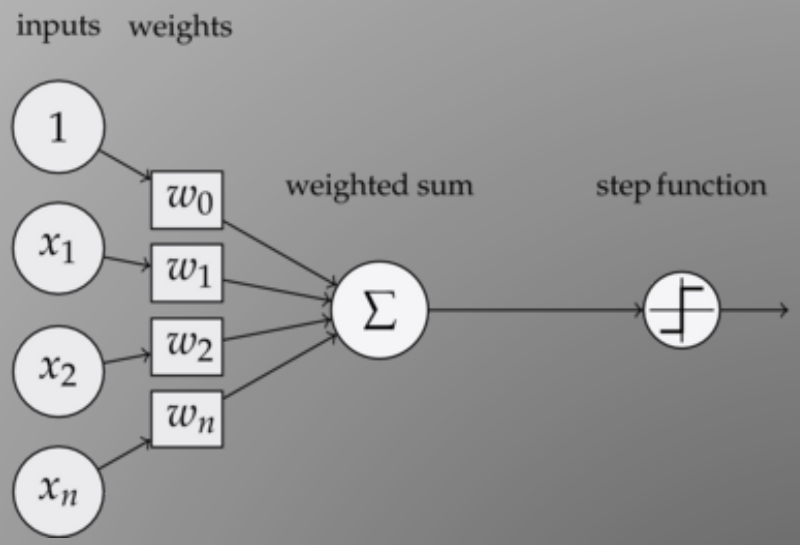
- Derive generalizations from examples
- Similar to humans
- One method uses *artificial neural networks*



Artificial Neural Networks

Single Perceptron

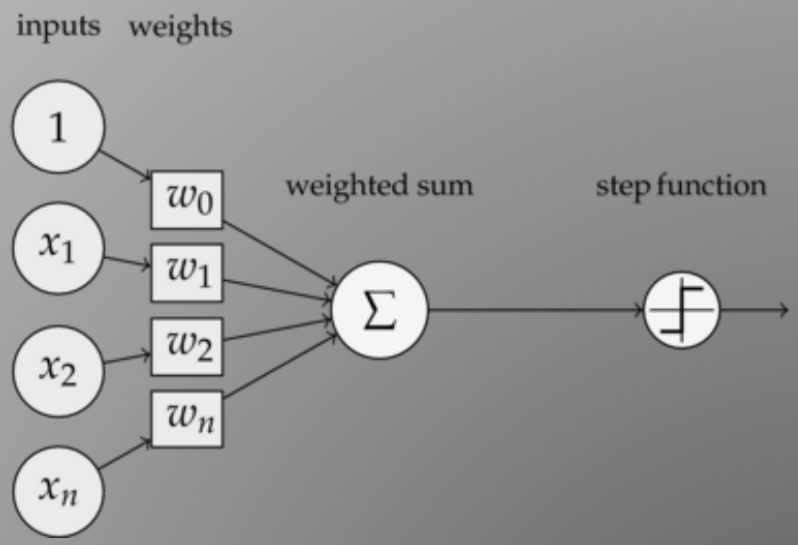
- General model for neuron



Artificial Neural Networks

Single Perceptron

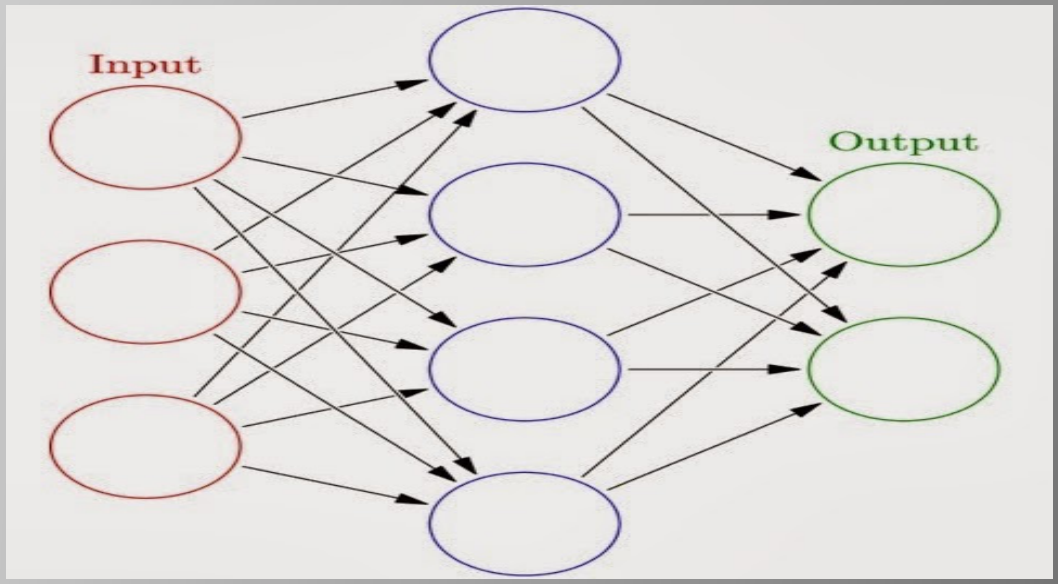
- General model for neuron
- Used in:
  1. Feed-Forward Neural Networks
  2. Recurrent Neural Networks



Artificial Neural Networks

Feed-Forward Networks

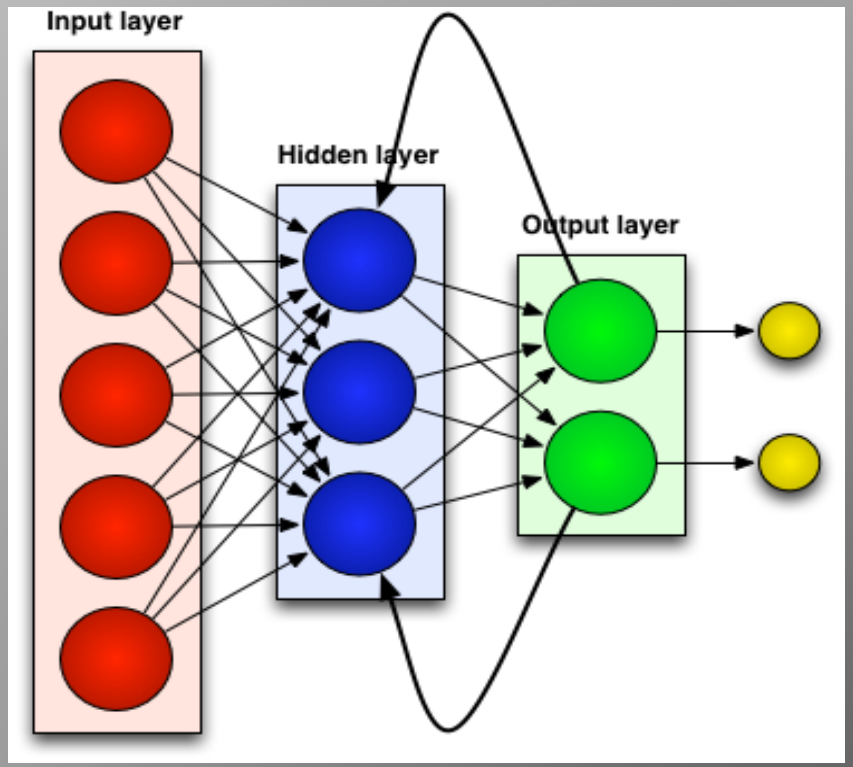
- Most common type
- Neural links only go forward
  - Like an assembly line
- Output becomes input for next layer



Artificial Neural Networks

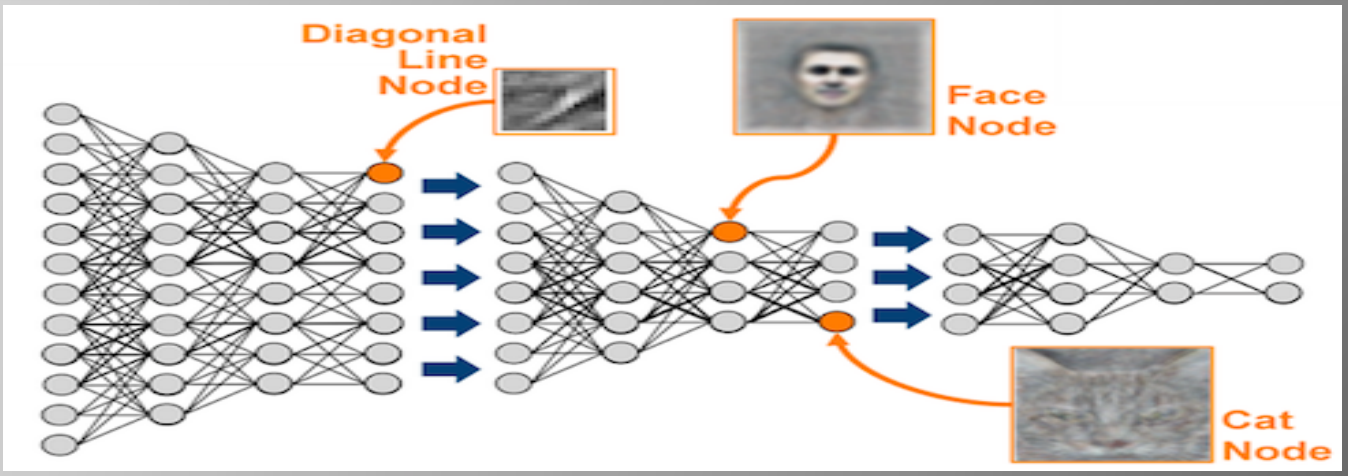
Recurrent Networks

- More complex
- Neural links are bidirectional
- Output can be input for:
  - Next layer
  - Current layer
  - Previous layer
- Support memory



Deep Learning

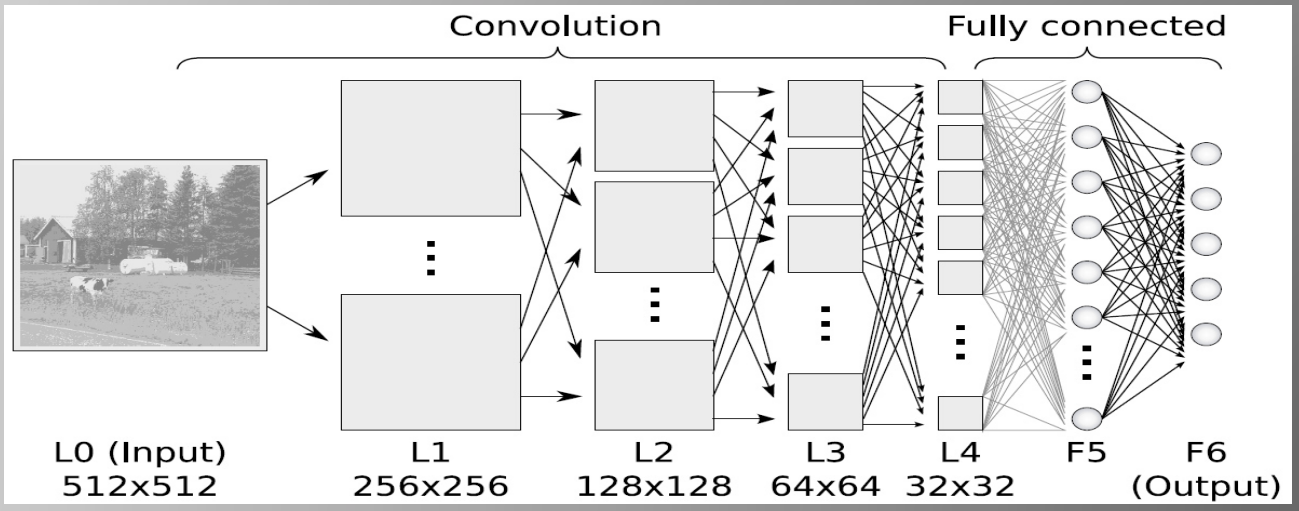
- Type of machine learning
- Specific structure: Deep (lots of) layers of neural networks
- Examples:
  - Convolutional Neural Networks
  - Deep Belief Networks



Deep Learning

Convolutional Neural Networks

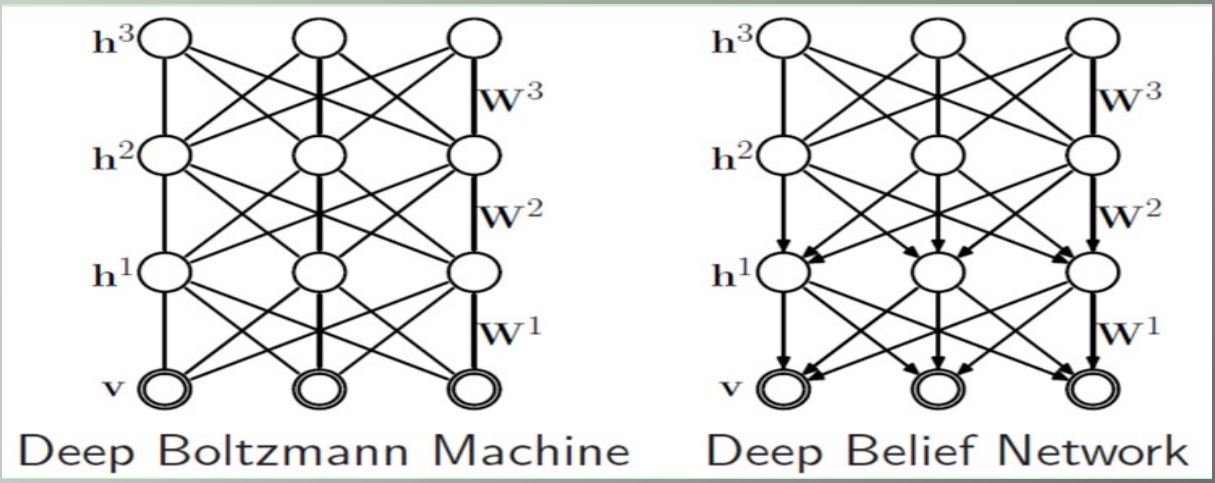
- Feed-forward network
- Neurons correspond to overlapping parts of the image
- Outputs from layers are pooled



Deep Learning

Deep Belief Networks

- Layers learn in top-down approach
  - Layers depend on other layers
- Can reconstruct inputs
- Generative model
  - e.g. generate an image



But are they conscious??

Consciousness

(Functionalism)

Multiple Drafts Model
Daniel Dennett

- Brain activity is parallel
  - Information is continually revisable and accessible

Consciousness

(Functionalism)

Multiple Drafts Model
Daniel Dennett

- Brain activity is parallel
  - Information is continually revisable and accessible
- 'Qualia' don't really exist

Consciousness

(Functionalism)

Multiple Drafts Model
Daniel Dennett

- Brain activity is parallel
  - Information is continually revisable and accessible
- 'Qualia' don't really exist
- Consciousness = the functional effects of judgments

Are Deep Learning Computers Conscious?

Multiple Drafts Model

– Consciousness doesn't "need" qualia

Are Deep Learning Computers Conscious?

Multiple Drafts Model

- Consciousness doesn't "need" qualia
- Deep Learning computers:
  - Function consciously
  - Process information consciously

Are Deep Learning Computers Conscious?

Multiple Drafts Model

- Consciousness doesn't "need" qualia
- Deep Learning computers:
  - Function consciously
  - Process information consciously
- Thus: computers are conscious

Consciousness

(Partial Physicalism)

Hybrid Theory
Ned Block

- Physicalism: Conscious states = Physical states

Consciousness

(Partial Physicalism)

Hybrid Theory
Ned Block

- Physicalism: Conscious states = Physical states
  - Access-consciousness (A-consciousness)
    - states that are available for rational processes
  - Phenomenal-consciousness (P-consciousness)
    - "*what it is like*-ness"

Consciousness

(Partial Physicalism)

Hybrid Theory
Ned Block

- Physicalism: Conscious states = Physical states
  - Access-consciousness (A-consciousness)
    - states that are available for rational processes
  - Phenomenal-consciousness (P-consciousness)
    - "*what it is like*-ness"

  – 'Consciousness' refers to A- and P-states
  – Physical make-up matters!

Are Deep Learning Computers Conscious?

Hybrid Theory

– Consciousness -> both A-states and P-states

Are Deep Learning Computers Conscious?

Hybrid Theory

- Consciousness -> both A-states and P-states
- Deep learning computers aren't P-conscious
  - They don't support P-consciousness

Are Deep Learning Computers Conscious?

Hybrid Theory

- Consciousness -> both A-states and P-states
- Deep learning computers aren't P-conscious
  - They don't support P-consciousness
- Thus: computers are unconscious
  - But they *are* A-conscious

Consciousness

(Modified Functionalism)

Integrated Information Theory
Giulio Tononi

- Consciousness depends on:

Consciousness

(Modified Functionalism)

Integrated Information Theory
Giulio Tononi

- Consciousness depends on:
  - *Information*: number of possible alternative outcomes (based on entropy)
  - *Integration*: interdependency between parts of the system

Consciousness

(Modified Functionalism)

Integrated Information Theory
Giulio Tononi

- Consciousness depends on:
  - *Information*: number of possible alternative outcomes (based on entropy)
  - *Integration*: interdependency between parts of the system
- Amount of consciousness relates to
  1. Amount of *information* in the system
  2. Degree of interdependency in subsystems

Are Deep Learning Computers Conscious?

Integrated Information Theory

– Consciousness = information integration

Are Deep Learning Computers Conscious?

Integrated Information Theory

- Consciousness = information integration
- Feed-back is important

Are Deep Learning Computers Conscious?

Integrated Information Theory

- Consciousness = information integration
- Feed-back is important
- Thus:
  - Feed-forward networks (convolutional networks)
    - not conscious
  - Recurrent networks (deep belief networks)
    - are conscious

*Consciousness varies with design

Where Do We Go From Here?

- Which theory is correct?

Where Do We Go From Here?

- Which theory is correct?
- How do we find out?
  - Philosophical debate
  - Empirical research
    - » Consciousness science
    - » Neural Network Design