Senior Project – Philosophy and Computer Science ID – 2015

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

(Peter) Zachary Davis

UNION COLLEGE

Advisors – Kristina Striegnitz & David Barnett

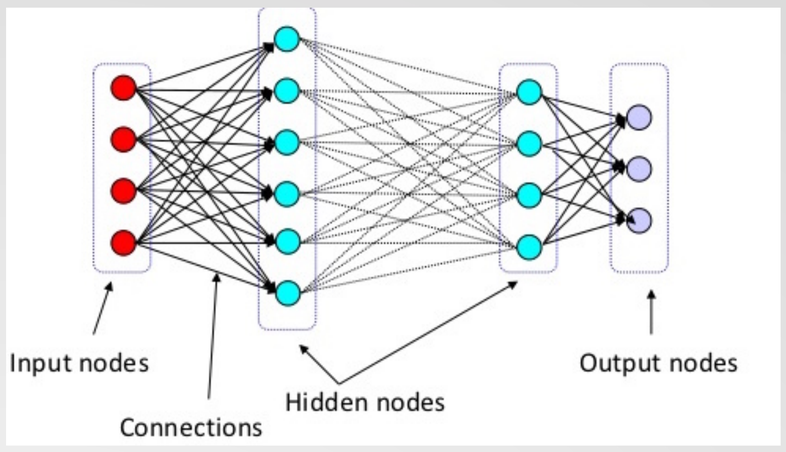
Abstract

Technological advances of recent decades have allowed computer scientists to create robots and computer programs that were previously impossible, thus prompting the question: can a computer be conscious? The answer relies on two key sub-problems. The first is the nature of consciousness: what constitutes a system as conscious, or what properties does consciousness have? Secondly, does the physical composition of the computer matter for consciousness? My aim is to explore these issues with respect to deep-learning computer programs, which use artificial neural networks and learning algorithms to create seemingly intelligent computers that they are roughly comparable to, yet fundamentally different from, the human brain.

Deep Learning

Deep learning is a type of machine learning that uses layers of artificial neural networks (ANN) to complete complex tasks, such as image and language processing. Each neural layer is comprised of *neurons,* which are units that are capable of a simple task. Two types of deep neural networks are commonly used. ANNs aim to replicate brain structures and functions.

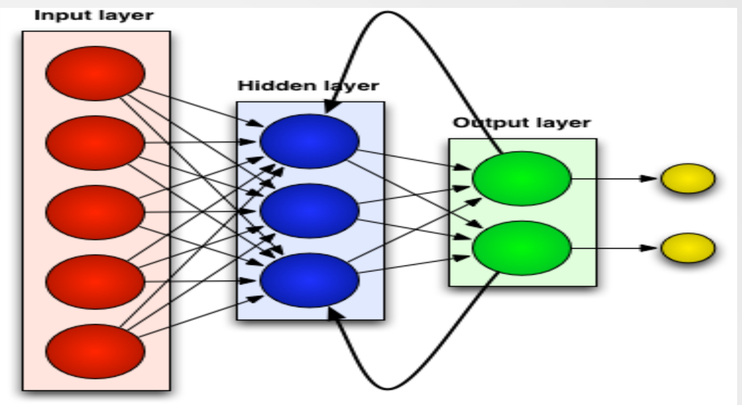
Feed-Forward Neural Networks



Links between neural layers all point in the same direction.

Output from neurons in each layer, except the last layer, acts as input for neurons in the next layer.

Recurrent Neural Networks



Links between layers can go forward *and* backward.

Output from neurons in one layer can act as input for neurons in a previous layer, or in the same layer again.

Conclusions

In order to determine which theory of consciousness is correct, empirical research in the science of consciousness must be conducted. Work should include determining the origins of consciousness, yielding evidence of perceptual experience's role. Neural network research should also be conducted to create recurrent networks that are structured and function as closely as possible to human brains, such as adding neurons and Restricted Boltzmann Machines as connected subsystems.

Consciousness

A vs. P Physicalism (Ned Block)

- Consciousness is a hybrid concept, and refers to:
  - *Access consciousness*: mental states that are available to a system's rational processes
  - *Phenomenal consciousness*: mental states such that there is something it is like to be in them
- Consciousness depends on the realization of the system
- P-consciousness plays a key role in considering whether or not something is conscious

Multiple Drafts Model (Daniel Dennett)

- The brain processes input in a parallel manner, and is continually able to edit and access information
- Consciousness is a product of the collection and availability of information
- Perceptual 'qualia' (P-consciousness) do not exist, only judgments do
- Qualia do not factor into consciousness

Integrated Information Theory (Giulio Tononi)

- The consciousness of a system relies on:
  - *Information*: the number of alternative outcomes that are not the case, based on entropy
  - *Integration*: the structure and use of information
- A system is conscious when information is integrated such that subdivisions of the system are not capable of integrating the same information independently

The Application of the Theories

Judgments about consciousness in deep learning computers differ with each theory

A vs. P Physicalism: The realization of deep learning computers does not support P-conscious states, and therefore, computers are not conscious. However, they *are* A-conscious.

Multiple Drafts Model: Since qualia are not necessary for consciousness, and since deep learning computers both function and process information like humans (conscious beings), computers can be considered conscious

Integrated Information Theory: Some deep learning computers are conscious, but the *amount* of consciousness they have varies by structure and the amount of information in them; feed-back is necessary for high integration. Thus, feed-forward networks are not conscious, while recurrent networks are conscious. If the amount of information and integration is high, they will reach high levels of consciousness.