Chapter 1
Integrated Information Theory (IIT) and Artificial Consciousness

Francis Fallon
St. John's University, USA

ABSTRACT

This chapter aims to evaluate Integrated Information Theory's claims concerning Artificial Consciousness. Integrated Information Theory (IIT) works from premises that claim that certain properties, such as unity, are essential to consciousness, to conclusions regarding the constraints upon physical systems that could realize consciousness. Among these conclusions is the claim that feed-forward systems, and systems that are not largely reentrant, necessarily will fail to generate consciousness (but may simulate it). This chapter will discuss the premises of IIT, which themselves are highly controversial, and will also address IIT's related rejection of functionalism. This analysis will argue that IIT has failed to established good grounds for these positions, and that convincing alternatives remain available. This, in turn, implies that the constraints upon Artificial Consciousness are more generous than IIT would have them be.

DOI: 10.4018/978-1-5225-1947-8.ch001

INTRODUCTION

The functionalist account of mind holds that mental events are functions, rather than fundamental properties attaching to or arising from particular physical arrangements. According to functionalism, it may be the case that only a limited set of physical systems will be practically capable of realizing mental events, such as consciousness. It stresses, at the same time, that these physical systems will realize, e.g., consciousness, in virtue of their functional characteristics. Function may constrain physical form, but function, rather than physical properties themselves, adequately explains consciousness.

IIT departs from functionalism, claiming that consciousness is a fundamental property of certain kinds of physical systems. Systems that integrate information – combine different kinds of information into a unified whole – produce consciousness. The brain does this by reentrant processing. IIT claims that the production of consciousness by integrated information systems is a feature of the physical communication among the parts of the whole, reentrant system. IIT, then, understands consciousness not as a set of functions which can be achieved by any system with the right functional capacity and organization. Rather, IIT identifies consciousness as a fundamental feature only of systems (that integrate information) whose anatomy is physically reentrant. Subjective experience is a property of these systems in the same fundamental way as mass or charge is a property of particles. IIT, then, rules out feed-forward or otherwise insufficiently reentrant systems as having the potential to generate consciousness, which limits the options available to the field of Artificial Consciousness.

This chapter will forego a general evaluation of IIT and focus on its claims concerning the neurophysiological constraints on modelling Artificial Consciousness. This will involve, after laying out the background,

1. Clarifying its anti-functionalist commitments,
2. Tracing the link between those commitments and the conclusion that consciousness emerges as a basic property of reentry systems only,
3. Challenging the anti-functionalist position, and
4. Revisiting the possibilities for Artificial Consciousness architecture, arguing for a more generous interpretation of the parameters for future work (in Solutions and Recommendations).