

A Systematic Approach to Synthetic Phenomenology

David Gamez

Department of Computing and Electronic Systems, University of Essex, UK

Abstract

The term ‘synthetic phenomenology’ has been given a number of different interpretations and I will be using it here to refer to the determination whether artificial systems are capable of conscious states and the description of these states if they occur. This approach to synthetic phenomenology is similar to that put forward by Aleksander and Morton (2007) and it is close to the philosophical tradition of phenomenology, with the word “synthetic” being added to indicate that it is the phenomenology of artificial systems that is being described. Synthetic phenomenology has a number of overlaps with the description of human phenomenology from a third person perspective using measurements of brain activity gathered using techniques, such as fMRI, EEG or electrodes. Good examples of this type of work are Kamitani and Tong (2005) and Kay et al. (2008), who used the patterns of intensity in fMRI voxels to make predictions about the phenomenal states of their subjects.

One of the main objectives of synthetic phenomenology is to establish whether work on machine consciousness is successfully creating artificial conscious states. Systematic techniques for examining systems for representational and phenomenal mental states will also become increasingly necessary as more complex robots are developed that learn through interaction with their environment. With such systems it will be unclear why they are behaving in the way that they do or how they are going to act next, which raises important safety concerns. This type of analysis could also help us to understand the phenomenal states of very young or brain-damaged people, who are incapable of expressing their experiences in language.

My research on synthetic phenomenology is based on an interpretation of consciousness that distinguishes between the phenomenal world of our experiences and the physical world described by science. A correlates-based approach links the phenomenal and the physical worlds and theories of consciousness are used to make predictions about the association between physical and phenomenal states - for example, a physical system that exhibits depiction, volition, emotion, attention and imagination is predicted to be conscious by Aleksander’s (2005) theory. When the potential correlates of consciousness (PCCs) are examined in more detail, Moor’s (1988) and Prinz’s (2003) work on the brain-chip replacement experiment suggests that it is necessary to distinguish between two types of PCCs. Type I PCCs are behaviour-neutral, which makes it impossible to prove their connection with consciousness empirically, whereas type II PCCs do affect behaviour and it can be established whether they are systematically linked to conscious states. This type I/ II distinction can be used to classify different theories of consciousness and it plays an important role in my approach to synthetic phenomenology.

It is impossible to describe the phenomenology of a system that is not *capable* of consciousness, and so the first challenge faced by synthetic phenomenology is to identify the systems that are capable of phenomenal states. Setting aside the problem that some correlates of consciousness may be probabilistic and multifactorial, the behaviour-neutrality of type I PCCs means that we cannot experimentally identify a list of the necessary and sufficient correlates of consciousness. This prevents us from ever knowing for *certain* whether biological neurons, for example, are necessary for consciousness, or if they are just one of the mechanisms by which consciousness happens to be implemented in human beings. Since it is indeterminable whether silicon-based robotic systems are conscious or not, a major obstacle lies in the way of any attempt to describe the *phenomenology* of such systems.

One approach to this problem is to follow Prinz (2003) and suspend judgement about whether artificial systems are capable of phenomenal states. However, one problem with this approach is that many people have a strong intuition that machines built in a similar way to humans are likely to be phenomenally conscious, and so it may be necessary to take the idea that certain types of machines have conscious experiences seriously. Second, as machine consciousness progresses we are likely to start developing machines that exhibit more complex behaviour and spend a lot of time confused and potentially in pain, which has been somewhat dramatically compared by Metzinger (2003, p. 621) to the development of a race of retarded infants for experimentation. To address these ethical worries without stifling research a way

needs to be found to evaluate the likelihood that a robot is capable of phenomenal states. A third problem with suspending judgement is that as more sophisticated robots emerge, people are inevitably going to attribute more and more consciousness to them. People are already prepared to attribute emotions to robots as simple as Braitenberg's vehicles (Dautenhahn 2007), and a systematic way of evaluating phenomenal states in a system needs to be in place before this becomes a live public issue. The general public is very interested in the question whether something is *really* conscious and it would be helpful if the machine consciousness community could formulate some kind of answer, even if this is based on analogy with human beings. To address these issues and provide a framework within which the more detailed work of synthetic phenomenology can proceed, I have developed an ordinal machine consciousness scale that orders machines according to the degree to which their type I PCCs match human type I PCCs.

Once the question about type I PCCs has been dealt with it, type II theories of consciousness can be used to generate predictions about machines' phenomenal states. The methodology that I have developed is based around definitions of a mental state and a representational mental state, which can be identified by exposing a system to different test stimuli and measuring its response. Type II theories of consciousness, such as Tononi (2004), can then be used to predict whether the system's mental states are associated with phenomenal states. There are a number of reasons why human language is unsuitable for the final phenomenological description, and I have developed an alternative approach that uses a markup language to combine human and physical descriptions with other information about the system.

To illustrate and test this approach to synthetic phenomenology, a spiking neural network was developed and analyzed for consciousness using Tononi's (2004), Aleksander's (2005) and Metzinger's (2003) theories. To identify the representational mental states in the network, noise was injected into the input and output layers and mutual information was used to identify the internal neurons that responded to the external stimulation. The network was then examined for information integration (Tononi and Sporns 2003), which was used to analyze the network according to Tononi's theory of consciousness, to support the analysis for Metzinger's theory of consciousness and to evaluate the integration between neurons in the network. The final part of the analysis was the generation of files containing a description of the predicted phenomenology of the network at each time step, and the predicted distribution of consciousness was plotted for Tononi's, Aleksander's and Metzinger's theories. These results showed that different parts of the network were predicted to be conscious according to these three theories, but it was not possible to predict the absolute amount of consciousness because the measures had not been calibrated on normal waking human subjects.

Full details about this work can be found in my recent PhD thesis, which is available at: www.davidgamez.eu/mc-thesis.

References

- Aleksander, I. (2005). *The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in People, Animals and Machines*, Imprint Academic, Exeter.
- Aleksander, I. and Morton, H. (2007). Depictive Architectures for Synthetic Phenomenology, in A. Chella and R. Manzotti (eds.), *Artificial Consciousness*, Imprint Academic, Exeter.
- Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction, *Phil. Trans. R. Soc. B* 362: 679–704.
- Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain, *Nature Neuroscience* 8(5) 679-685.
- Kay, K.N., Naselaris, T., Prenger, R.J. and Gallant, J.L. (2008). Identifying natural images from human brain activity, *Nature advance online publication*, doi:10.1038/nature06713.
- Metzinger, T. (2003). *Being No One*, The MIT Press, Cambridge, Massachusetts.
- Moor, J.H. (1988). Testing robots for qualia. In H.R. Otto and J.A. Tuedio (eds.), *Perspectives on Mind*, D. Reidel Publishing Company, Dordrecht/ Boston/ Lancaster/ Tokyo.
- Prinz, J.J. (2003). Level-Headed Mysterianism and Artificial Experience. In O. Holland (ed.), *Machine Consciousness*, Imprint Academic, Exeter.
- Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience* 5:42.
- Tononi, G. and Sporns, O. (2003). Measuring information integration. *BMC Neuroscience* 4:31.