Available online at www.sciencedirect.com

ScienceDirect

Consciousness and Cognition 18 (2009) 187–192

Consciousness and Cognition

www.elsevier.com/locate/concog

Establishing consciousness in non-communicative patients: A modern-day version of the Turing test

John F. Stins

Research Institute MOVE, Faculty of Human Movement Sciences, VU University Amsterdam, van der Boechorststraat 9, 1081 BT, Amsterdam, The Netherlands

Received 12 October 2007
Available online 8 February 2008

Abstract

In a recent study of a patient in a persistent vegetative state, [Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science, 313*, 1402] claimed that they had demonstrated the presence of consciousness in this patient. This bold conclusion was based on the isomorphy between brain activity in this patient and a set of conscious control subjects, obtained in various imagery tasks. However, establishing consciousness in unresponsive patients is fraught with methodological and conceptual difficulties. The aim of this paper is to demonstrate that the current debate surrounding consciousness in VS patients has parallels in the artificial intelligence (AI) debate as to whether machines can think. Basically, (Owen et al., 2006) used a method analogous to the Turing test to reveal the presence of consciousness, whereas their adversaries adopted a line of reasoning akin to Searle's Chinese room argument. Highlighting the correspondence between these two debates can help to clarify the issues surrounding consciousness in non-communicative agents.

Keywords: Vegetative state; Turing test; Chinese room; Machine consciousness; Neurology

1. Introduction

Extensive brain damage can cause a complete and permanent loss of cognitive and behavioral capabilities. In some cases, brain damaged patients enter a so-called vegetative state (VS), with poor prospects of recovery. Patients in VS have intact sleep–wake cycles, and may on occasion move their limbs or facial muscles, which sometimes gives the appearance of willed behavior. Yet, the consensus is that these patients lack awareness of themselves and their surroundings. Despite the neurological damage, a large amount of neural tissue is often still structurally and functionally intact. It is quite common for physicians to probe the extent of residual neural functioning for diagnostic purposes, e.g., to differentiate between VS and the minimally conscious state (MCS; e.g., Di et al., 2007). Also, brain damage may sometimes result in locked-in syndrome (LIS), which resembles VS in several regards (Kobylarz & Schiff, 2004). LIS is usually caused by severe pontine lesioning,

E-mail address: j.stins@fbw.vu.nl

1053-8100/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.concog.2007.12.005

and is characterized by complete paralysis of the body, although some control over eye movements still exists. Crucially, LIS patients are fully aware of themselves and their surroundings, and they are able to engage in acts of communication through a system of eye movements, e.g., one blink signals "yes", and two blinks signal "no" (Laureys et al., 2005). As such, LIS is not considered a disorder of consciousness (DOC), whereas VS is.

Most researchers and physicians are reluctant to take brain activity in VS as evidence for the presence of consciousness in their sets of patients. Professionals are aware of the methodological and ethical pitfalls involved in attributing consciousness to seemingly intact brain functioning, and they tend to interpret it as residual brain activity that operates on a more or less automatic level, rather than evidence of a conscious self. A notable exception to this self-imposed ban on talking of consciousness is the recent study of Owen et al. (2006, 2007a, 2007b) who, using a novel and clever fMRI paradigm, found elaborate brain activity in a patient who had been in VS for more than 5 months. Owen et al. (2006) concluded that the most parsimonious interpretation of this brain activity was that the patient was truly conscious, but that the extent of her brain damage apparently prevented her from expressing herself, which is often a defining characteristic of consciousness. Owen et al.'s (2006) radical interpretation did not go unnoticed, and their study was challenged on methodological grounds (Greenberg, 2007; Nachev & Husain, 2007), to which Owen et al. (2007b) wrote a reply, defending their original position.

At the heart of the controversy lies, I submit, a fundamental question that has parallels in the artificial intelligence (AI) debate as to whether machines can think. In essence, Owen et al. (2006) used a method analogous to the Turing test to reveal the presence of consciousness (see also Naccache, 2006), whereas their adversaries adopted a line of reasoning akin to the Chinese room argument (Searle, 1980), and which also seems to form the core of Greenberg's (2007) critique. In this paper, I will argue the position that the epistemological question whether (at least some) VS patients are conscious is analogous to the question whether machines can think.

2. The Owen et al. (2006) study: a real-life Turing test

In 2006 Owen et al. published a high profile paper in Science. Owen et al. (2006) performed an fMRI scan on a woman who was diagnosed as VS following a traffic accident, and who had been in that condition for more than 5 months. During the scan they asked the patient (who was completely unresponsive) to engage in one of two mental activities: she was sometimes asked to imagine herself playing tennis, and other times to imagine herself walking through her house. This mental activity had to be maintained for the duration of the scan (30 s). Astonishingly, the differential pattern of brain activity was indistinguishable from the brain activity recorded with (conscious) control subjects. Apparently, not only did this patient 'understand' the verbal request to produce some mental activity, she also 'complied' with the request to do so. The pattern of results was so compelling to Owen et al. (2006) that they drew the far-reaching conclusion that this patient must have been conscious. Owen et al. (2006) claimed explicitly that the patient "*understand[s]* spoken commands", "*[decided]* to cooperate", and demonstrated an "*act of intention*" (p. 1402; italics added).

However, as others, such as Greenberg (2007) and Nachev and Husain (2007) were quick to point out, there is an alternative explanation to the results: the verbal content in the requests could have automatically triggered corresponding brain activity. For example, the word 'tennis' could have triggered, via an intact lexical module, brain activity related to motor activity, such as in the supplementary motor area. In response, Owen et al. (2007b) ran a control condition with a healthy volunteer, who showed no activation in the relevant brain areas when presented with isolated words, such as 'tennis' and 'house'. Are these results then sufficient evidence for the presence of consciousness in the VS patient? Owen et al. (2007b) remained convinced they are, and they reiterated that "this patient was consciously aware and purposefully following the instructions given to her" (p. 1221). However, a critic could maintain that the observed brain activity is still no evidence for the presence of subjective experience: it could be the case that the patient has a rich and complex neural life, yet is completely devoid of consciousness.

In my opinion the question boils down to what it takes for an outside observer to be convinced that an entity (a machine such as a computer, or a VS patient) is conscious. With respect to testing the possibility of consciousness in a machine, Alan Turing (1950) proposed that the machine should pass the so-called Turing test in order to be considered conscious. In a nutshell, if the machine manages to convince a human interrogator in a question-and-answer session that s/he is conversing with a human being, then the machine must be intelligent (for an overview of some past and present issues surrounding the Turing test, see French, 2000)[1].

At the heart of the Turing test is the aspect of communication; we can only know if other entities such as human beings can think and have feelings similar to ours, because we can somehow engage in an act of conversation with them, via whatever means available. This point was also raised by Owen et al. (2007a) when they state that "our ability to know unequivocally that another being is consciously aware is ultimately determined not by whether or not he or she is aware but instead by his or her ability to communicate that fact through a recognized behavioral response" (p. 1099). Other human beings share their inner most thoughts and experiences with us in word and gesture (or in the form of the blink of an eye, in the case of LIS), and based on these utterances we are led to conclude that our fellow human beings are indeed conscious and that they experience the world in ways similar to ourselves. The Turing test can be considered an attempt to spell out the conditions under which consciousness can be established within another agent, using a communication protocol.

The test used by Owen et al. (2006) and by others to assess consciousness in VS is analogous to the Turing test in three regards. First, some sort of external input is applied to the computer or the patient. In the case of the Turing test the input consists of a set of questions directed at the computer in some language that the computer can process. In most accounts of the Turing test the questions are designed so as to maximize the chances of correctly distinguishing a conscious agent (human) from a non-conscious agent (the computer). In the case of studies with VS patients the stimuli can be delivered to different sensory modalities and come in different degrees of complexity. Some of the stimuli that have been used are (a) direct physical stimuli, such as a loud noise or strong noxious stimulation (e.g., Kassubek et al., 2003), (b) highly salient stimuli such as hearing one's own name (e.g., Laureys, Perrin, & Brédart, 2007; Perrin et al., 2006), (c) natural language sentences with some interesting property such as a semantic anomaly (e.g., Schoenle & Witzke, 2004), and (d) complex instructions asking the patient to mentally engage in some sort of visualization, as in Owen et al. (2006, 2007b). This so-called "hierarchical approach" is based on the assumption that increases in task complexity may index degrees of residual cognitive capabilities (Owen & Coleman, 2007; Owen et al., 2005).

Second, the system under scrutiny must generate some form of output. The output in the case of the Turing test is usually thought of as consisting of written text in response to the written input. Not only the content of the text messages, but also other "cues", such as the speed at which the message is delivered, can be taken into account by the interrogator. For example, if the input consists of a difficult maths question, and the correct answer is produced in a split second, we can be reasonable certain we are not dealing with a human being (at least not an ordinary one). In the case of VS patients the output consists of bodily reactions in response to the stimulus and not of verbal reactions, due to the very nature of VS. The reactions are sometimes cardiovascular (e.g., elevated heart rate in response to noxious stimulation; Kassubek et al., 2003), but are most often of a functional neuroimaging nature, such as an EEG pattern or the BOLD signal. For example, Perrin et al. (2006) found a clear P3 auditory evoked potential in response to hearing one's own name in LIS patients, a delayed P3 in MCS patients, and a delayed or even completely absent P3 in VS patients. In the absence of verbal or motoric behavior, these neural reactions may provide a glimpse of the patient's subjective life.

Third, and most importantly, the output has to be interpreted by the interrogator or the researcher. Turing claims that if the interrogator is unable to distinguish, based on an extensive question-and-answer session, whether s/he is having a conversation with a computer or a human being, then the computer must be intelligent. What about interpreting bodily signals emanating from VS patients? Here, the situation is more problematic because there is still no gold standard against which to compare the brain signals for the presence or absence of consciousness. What is interesting about the Owen et al. (2006) study (and their reply to criticisms) is that they made the bold claim that their patient was conscious, precisely because she "retained the ability to understand spoken commands and to respond to them through her brain activity, rather than through speech or movement" (Owen & Coleman, 2007, p. 635). In essence, the VS patient had passed the Turing test, at least as far as Owen et al. (2006) are concerned.

[1] In the original formulation the focus was on machine intelligence, whereas nowadays the emphasis has shifted to the possession of consciousness. However, in both instances we are dealing with a unique human mental capability, arguably not (yet) shared by machines.

3. Objections to the Turing test: Searle's Chinese room

Now, what should we conclude from the observation that the pattern of neural activity in a given VS patient, such as the one studied by Owen et al. (2006), has all the properties of a conscious, yet unresponsive, agent? Are we to conclude that the patient is conscious but somehow not capable of willful expression? Most researchers would be hesitant to ascribe consciousness to such a patient on the grounds that what we are witnessing may as well be isolated patches of intact neural activity that in no way resembles our own subjective experience. For example, the fact that the auditory cortex becomes active some 90 ms after presentation of a tone is a necessary but not a sufficient condition for the conscious registration of a particular sound. Most researchers would argue that something 'more' is needed for conscious awareness of the sound. For example, Kobylarz and Schiff (2004) stated that "we cannot confirm awareness simply on the basis of imaging findings without some reliable communication from the patient." (p. 1358). A criticism along similar lines was raised by Nachev and Husain (2007) in response to Owen et al.'s (2006) study when they argued that "the presence of brain activation is not sufficient evidence for the associated behavior—here, supposedly consciously mediated behavior" (p. 1221). Finally, Boly et al. (2007) wrote that "in the absence of a full understanding of the neural correlates of consciousness, even a near-to-normal activation in response to passive sensory stimulation cannot be considered as a proof of the presence of awareness in these [VS] patients." (p. 980).

These criticisms are reminiscent of the famous Chinese room argument formulated by Searle (1980) in response to the Turing test.[2] In a nutshell, Searle argued that it is possible to think of a system that can pass the Turing test, yet without being conscious. As a thought experiment, Searle envisioned a situation of a (conscious) human being who does not understand a word of Chinese and who is locked up in a room. On occasion the human receives a piece of paper with Chinese characters through a hole in the wall. The human has access to a look-up table on how to convert certain strings of Chinese symbols (which he does not understand) to other strings of Chinese symbols (which he does not understand either), allowing him to pass the written output through a hole to a person outside the room. To an outside observer endowed with Chinese language capabilities, the human inside the room may appear to understand Chinese. After all, meaningful linguistic input is consistently transformed into meaningful linguistic output. Thus, the observer at some point would have to conclude that the person inside the room has passed the Turing test with respect to understanding Chinese language. However, the person inside the room blindly follows a set of rules without a true understanding of the linguistic material. The crux of the argument is that a system that follows a set of syntactic rules may (to an outside observer) give the appearance of a system endowed with semantic capabilities. Does the VS patient in Owen et al.'s (2006) study truly understand and follow the instructions? Yes, Owen et al. (2006) claim. Not necessarily, Greenberg (2007) responds. The nervous system of the patient may as well reside in a dormant syntactic mode, as it were: capable of processing but not capable of understanding. Computation without comprehension. After all, the brains of sleeping persons may also show signs of cognition (e.g., the brain clearly responds to its own name; Bastuji, Perrin, & Garcia-Larrea, 2002), yet the sleeper is completely unaware of the stimulus. More often than not, stimuli that do not reach the consciousness threshold can be processed to a full semantic extent, as is evidenced by masking studies. Interestingly, Schoenle and Witzke (2004) found intact semantic processing in some VS patients. They employed a paradigm using semantically anomalous sentences that were presented auditorily to a group of VS patients and a group of controls. Inspection of the N400 ERP component revealed that the brains of (at least some) patients were "semantically intact" (Schoenle & Witzke, 2004, p. 331). However, semantics in this context refers to a linguistic capability (apparently intact in some VS patients), and not the meaningfulness that characterizes phenomenal subjective experience.

4. Owen's position

As Searle has repeatedly emphasized, ultimately we can only witness *behavior*, and we have no access to other people's first-person subjective experiences. Furthermore, even though a conscious experience may give

[2] There are many criticisms of the Turing test in its original formulation, but for present purposes I focus on Searle's argument.

rise to a certain behavior, we can not deduce from our observations of acts of behavior the presence of a mental life. How then are we to know that our fellow human beings are conscious? At the very best we can *infer* from other people's behaviors that they are conscious. If throughout our lifetimes we consistently see that people exhibit the same responses to environmental events as ourselves, we must conclude that their experiences must also be very similar to ours. So in a sense, we are continuously trying to read other people's minds, based on their behavior and verbal output. But this is exactly what Owen et al. (2006, 2007a) are doing when they resort to interpreting patterns of brain activation, which they refer to as "rudimentary mind reading" (p. 1101; 2007a). In essence, they take as their starting point the patient's neural behavior, and use these observations to draw inferences about the contents of their mental life. This approach is motivated by the spectacular amount of knowledge accumulated in the field of cognitive neuroscience. This field of research can be considered an attempt at identifying how neural states map onto mental states and/or behavior states (e.g., Gazzaniga, Ivry, & Mangun, 2002). The current state of the field of cognitive neuroscience allows us nowadays to tell with a reasonable amount of certainty what is going on in someone's mind, based solely on inspection of his or her neural activity. In addition, the emphasis in the field of neurology on the neural basis of consciousness in intact and damaged brains fits well with Searle's emphasis on the biological foundation of consciousness. Throughout his career Searle has defended the claim that brains can give rise to consciousness, whereas complex systems that are not composed of neural tissue such as computers have no consciousness and never will. According to Searle, consciousness is a higher level property of the brain, just as a higher level property of water is its liquidity (e.g., Searle, 1992).

But there is one aspect in Owen et al.'s (2006) logic that is radically different from other studies in the field of cognitive neuroscience, and that has to do with the ontological status of the brain signals that are being recorded. The consensus in the neuroscience community is that changes in brain states cause changes in a particular behavior and/or changes in mental content. Simply put, behavior is ultimately caused by the brain. But what Owen et al. (2006) did was treat the brain activity *itself* as an act of behavior. For them, brain activity in their VS patient was a vehicle for behavioral expression, similar to verbal report or bodily movements. The patient in Owen et al.'s (2006) study demonstrated her conscious self to the researchers by displaying her ability to understand the requests given to her, and by displaying her willingness to cooperate. She did this by modulating her brain activity. The brain thus serves a dual purpose in Owen et al.'s (2006) study: it not only creates consciousness (similar to Searle's position), it also *communicates* consciousness, and is therefore in principle amenable to a Turing test-like procedure.

The Turing test and the Chinese room argument were originally designed to tackle the philosophical question of consciousness in computers, but they also have relevance for the very real issue of whether consciousness exists in unresponsive patients. Although highlighting these view points can not be used to establish the presence or absence of consciousness for an individual case such as the patient studied by Owen et al. (2006), it can be used to clarify some of the issues surrounding the neural correlates of (un)consciousness.

Acknowledgments

I express my gratitude to Peter J. Beek and Neelke Troost for their helpful discussions on this topic.

References

Bastuji, H., Perrin, F., & Garcia-Larrea, L. (2002). Semantic analysis of auditory input during sleep: Studies with event related potentials. *International Journal of Psychophysiology, 46*, 243–255.

Boly, M., Coleman, M. R., Davis, M. H., Hampshire, A., Bor, D., Moonen, G., et al. (2007). When thoughts become action: An fMRI paradigm to study volitional brain activity in non-communicative brain injured patients. *NeuroImage, 36*, 979–992.

Di, H. B., Yu, S. M., Weng, X. C., Laureys, S., Yu, D., Li, J. Q., et al. (2007). Cerebral response to patient's own name in the vegetative and minimally conscious states. *Neurology, 68*, 895–899.

French, R. M. (2000). The Turing test: The first 50 years. *Trends in Cognitive Sciences, 4*, 115–122.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2002). *Cognitive neuroscience*. Norton.

Greenberg, D. L. (2007). Comment on "Detecting awareness in the vegetative state". *Science, 315*, 1221b.

Kassubek, J., Juengling, F. D., Els, T., Spreer, J., Herpers, M., Krause, T., et al. (2003). Activation of residual cortical network during painful stimulation in long-term postanoxic vegetative state: A $^{15}O–H_2O$ PET study. *Journal of the Neurological Sciences, 212*, 85–91.

Kobylarz, E. J., & Schiff, N. D. (2004). Functional imaging of severely brain-injured patients. *Archives of Neurology, 61*, 1357–1360.

Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., et al. (2005). The locked-in syndrome: What is it like to be conscious but paralyzed and voiceless? *Progress in Brain Research, 150*, 495–511.

Laureys, S., Perrin, F., & Brédart, S. (2007). Self-consciousness in non-communicative patients. *Consciousness and Cognition, 16*, 722–741.

Naccache, L. (2006). Is she conscious? *Science, 313*, 1395–1396.

Nachev, P., & Husain, M. (2007). Comment on "Detecting awaraness in the vegetative state". *Science, 315*, 1221a.

Owen, A. M., & Coleman, M. R. (2007). Functional MRI in disorders of consciousness: Advantages and limitations. *Current Opinion in Neurology, 20*, 632–637.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science, 313*, 1402.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2007a). Using functional magnetic resonance imaging to detect covert awareness in the vegetative state. *Archives of Neurology, 64*, 1098–1102.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2007b). Response to comments on "Detecting awareness in the vegetative state". *Science, 315*, 1221c.

Owen, A. M., Coleman, M. R., Menon, D. K., Berry, E. L., Johnsrude, I. S., Rodd, J. M., et al. (2005). Using a hierarchical approach to investigate residual auditory cognition in persistent vegetative state. *Progress in Brain Research, 150*, 457–471.

Perrin, F., Schnakers, C., Schabus, M., Degueldre, C., Goldman, S., Brédart, S., et al. (2006). Brain response to one's own name in vegetative state, minimally conscious state, and locked-in syndrome. *Archives of Neurology, 63*, 562–569.

Schoenle, P. W., & Witzke, W. (2004). How vegetative is the vegetative state? Preserved semantic processing in VS patients—Evidence from N 400 event-related potentials. *NeuroRehabilitation, 19*, 329–334.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*, 417–424.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.