The Human Brain: Mind or Machine?

Module: Understanding Patterns of Action

Convener: Professor David Clarke

Bethan Parry (Bsc Psychology)

Abstract:

The Human Brain: Mind or Machine? This question is investigated in terms or whether artificial agents can ever truly possess consciousness; the main functionalist argument against the absent and inverted qualia is presented alongside some of its present pitfalls and critical responses. Further arguments against functionalism are investigated and critically assessed. The conclusion debates the relevance of Dualism in the modern day and whether an answer to the question can ever be empirically proven.

Introduction:

The defining factor that makes the human brain unique has long been a contentious subject spanning many disciplines; from philosophy, ethics, and psychology to electronic engineering and cybernetics. Some people argue the principle of "consciousness" means the human brain mind transcends any machine. Consciousness can be simply defined as a construct's capability to perceive subjective thought and experience. But could a machine ever achieve consciousness? The argument has its conception in the Cartesian perspective "cogito ergo sum"; "I think therefore I am". Descartes' theory of dualism essentially states that the physical reality of the body can be realised through its extension in space, thoughts lack this extension and thus must be comprised of an entirely different substrate that a computer could never possess. The philosophy of Functionalism refutes this, asserting that mental states such as beliefs are solely a by-product of functional organization of neurons and theoretically, if a machine could be designed with its hardware identically functional organized it could accomplish a conscious state. The main functionalist theories will be described and evaluated in order to answer the question.

For the purposes of this essay, the focus will be centered around Strong Artificial intelligence (AI): Kurzweil (2005) defines this as a machine that has intelligence comparable to, or surpassing that of a human. Russell and Norvig (2003) construe intelligent action as comprising of reason, judgment, knowledge (including common sense), self-awareness, sentience, the ability to plan, learn and communicate and integrate all skills towards goals.

Emulation or Possession?

It has been widely documented that computational models, programmes and "robots" can emulate human activity efficiently. A prime example of this is an early programme called ELIZA (Weizenbaum, 1966). The ELIZA programme attempted to assimilate natural language processing in a primitive way. The most famous subtest was that of the DOCTOR in which ELIZA assumed the role of a psychotherapist and could respond to text scripts issued by the "patient" in a back and forth manner. The programme itself had a limited knowledge base of psychotherapy, and when the knowledge base was depleted the programme would simply substitute words from

the patient's scripts into pre-prepared sentences which would often result in a circular dynamic to the conversation. Although superficially effective, time would eventually betray the programme's ignorance and lack of true communication. ELIZA seems to be demonstrating language, a marker of strong AI, however this is merely an illusion evident as the level of communication breaks down, therefore ELIZA cannot be said to possess the facet of natural language, it is solely producing a pale imitation of it.

Igor Aleksander (1995) argues that the brain is a finite state machine, (a mathematical abstraction determining the brain has a limited number of possible states and transitions) suggesting that if these states could be determined consciousness could be replicated in digital computers.

Furthermore, Hans Morevac believes that computers can attain emotion. But as Crevier (1993) points out only in the mercenary manner of using emotions as a method of adapting behaviour to enhance a species' survival. It is debatable whether even if computers could effectively emulate emotion whether they could actually *feel* them.

Bringsjord, Bello, Ferruci (2001) devised the Lovelace Test to test computers' ability to possess originality, to create original thought, ideas etc, not merely a regurgitation of garnered input. Such experimental paradigms involved penning short stories from one given sentence It was found that computers consistently failed the lovelace test, it was theorised that computers need to possess something beyond mathematical models and networks of causation found in their programme to truely possess original ideas. The previous evidence, again, suggests that consciousness is nigh impossible for an artificial agent to achieve.

Chalmers' Functionalism:

The most compelling argument in favour of functionalism was proposed in David Chalmers' seminal paper of 1995: Absent Qualia, Fading Qualia and Dancing Qualia. He argues that conscious experience must have its genesis in physical properties and that these physical properties must be organized determined by laws; he denies that the brain's biochemical substrates are pertinent to consciousness. Instead he asserts that a system's experience is as a result of its functional organization alone. Functional organization is the complex template of causal

interactions between various components of a system and that if the functional organization of a human brain could be acutely replicated the two systems should share qualitatively duplicate experiences and that the causal organization of the human brain can, theoretically, be realised in many different forms. One such form proposed by Chalmers is via the replacement of all the brain's neurons with silicone chips; thus two functionally isomorphic systems have been create (the brain and the silicone system), whom share their organization at a fine grain level. In order to do this the number of constituents of the system would have to be identified, as well as all the possible states these constituents can employ; a network mapping the inter-dependency of these constituents would have to be formulated (e.g. how on constituent relies on the state of its previous counterpart to determine it's own state and so forth) and how any output from the system is dependent on this network.
Chalmers' offers explanations to two main objections to functionalist theories. The first of which is coined the "absent qualia objection". Qualia being an individual's subjective experiences. The objection quashes the idea that a computer may be conscious because they do not share the same biochemical substrates as a human mind. In order to dismiss the absent qualia objection, Chalmers uses a neural replacement scenario to hypothesise that absent qualia is empirically impossibly. it is argued that it is theoretically possible to replace neurons in the brain with silicone chips (providing the chip could perform the same local functions and could process the same input and produce the correct output). If you replaced one he posits it will make no difference to the functional organization. The replacement could continue neuron by neuron until eventually all biochemical mechanisms had been eradicated, leaving only a silicone system. By the absent qualia objection, the robot should not be conscious. The loss of consciousness would have either faded out over the spectrum of neural replacement (fading qualia) or there was a point at which consciousness was abruptly terminated (suddenly disappearing qualia).
Chalmers goes on to theorize that if suddenly disappearing qualia were an accurate representation of the loss of consciousness over the timescale of neuronal replacement there would have be a determinable point at which consciousness could effectively be activated and deactivated which he says is both implausible and "bizarre".
So if the suddenly disappearing theory of loss of consciousness is implausible, what about fading qualia? Chalmers imagines a mid-point system of half levels of

consciousness; this mid-point system would be functionally isomorphic to the "normal" human brain the spectrum begins with. So, for example, when the normal human brain perceives colour would the mid-point perceive only a dulled hue? The subtle distinctions of colour would no longer be present in the mid-point, the system would report the colour as a vivid blue for example, when it is in fact appearing subjectively as a much murkier tone rendering the mid-point systematically incorrect. Chalmers' points out that individual are rarely incorrect about their own subjective experience; he gives the example of the spectrum of falling asleep; one does not believe they are fully awake when they are falling asleep. Thus he suggests it is empirically impossible for a system to be so dissociated from its own experiences, rendering fading qualia unlikely, in turn weakening the absent qualia objection. Another popular objection opposed to functionalism is the "inverted qualia objection" Locke (1690): that a computer may achieve consciousness if correctly assimilated; but it's subjective experiences although of the same intensities, would be quite distinct from those of a human mind. For example, a computer may perceive the colour yellow as human minds would perceive the colour green Shoemaker (1982). If this objection were true there would be a spectrum as green (the human experience) became yellow (the system's experience); there must be two locations on this spectrum where there was no more than a tenth of the system between them, but where the colour yellow was discernable as a significantly different experience to that of the colour green. Chalmers argues that there is possible way to go from green to yellow in only ten units. Furthermore, if you could implement the system's circuitry into one's own human mind and switch back and forth between the two so that the colours "danced" interchangeably (dancing qualia), according to the inverted qualia objection there should be no perceivable difference; this is emperically highly unlikely. Since dancing qualia are virtually impossible, Chalmers' argues that so inverted qualia must be too. Chalmers' concludes that the replacement of the brain's neurons with silicone chips to the point that no more biochemical substrates remain will preserve qualia as long as functional organization is also retained.

Controversy and Responses to Chalmers:

Chalmers' paper provoked much discussion amongst the scientific community. Cultural differences pose a potential problem for Chalmers' argument. As discussed in Deregowski's (1989) review of cross-cultural perspectives, different cultural groups do not necessarily pictures in the same way. For example, African children may not necessarily perceive 3-D; they spend a great deal less time looking at and drawing the ambiguous trident (as seen in figure 1) because they do not perceive 3D they do not see an illusion and do not find it difficult to replicate. Western children do perceive 3D and find the trident much harder to copy. The two sets of children have essentially the same functional organization; it is not as if African children lack the brain structures involved in processing depth, but they do not share the same subjective experience. Surely, this is an empirical example of inverted qualia.

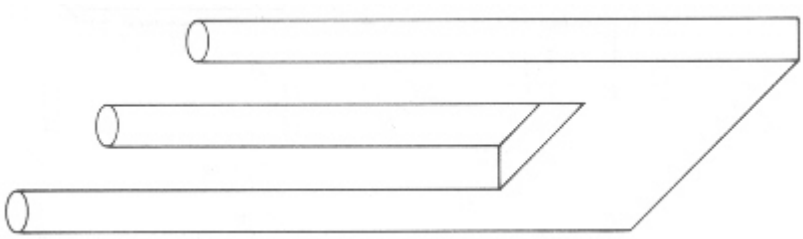

Figure 1: The ambiguous trident from Deregowski (1989)

Additionally, Chalmers argues that if we replaced a single neuron in the brain with a silicone chip that the functional organization would be identically and there would be no effect on the overall functioning or consciousness of the mind. However, could it be determined that the silicone chip had not effected functioning, it could be that it does not function at the level but the brain employs its own compensatory measures such as neural plasticity that Frost, Barbay, Friel, Plautz and Nudo (2003) demonstrated in patients after an ischemic stroke to counteract the loss of function in that neuron.

Block's (1978) "Chinese Nation" theory predates Chalmer's paper but the argument still stands. It is argued that the functional organization of the brain could be mirrored in the arrangement of the population of China, yet this would not lead to the commencement of a "group mind". This is an essentially dualist point of view, that the population lack the correct substrates to externally reproduce consciousness i.e.

the individual people are not neurons.
Bostrom (2006) argues that Chalmer's imagined mid-point does not see dulled hues instead of vivid colour, but that as consciousness dwindles it is not the quality of experience that lessens, but the quantity. Consciousness becomes fragmented. He argues that the change in quantity can occur in an analogous manner, in continuous degrees, thus you would not notice a change in perception of colour because as the quality does not differ there would be nothing to notice. Whilst not negating Chalmers' functionalist theory; quantity of experience may offer support for fading qualia.
Greffenstette argues that qualia and mental states are epiphenomenal, i.e. products of the physical world and cannot be reduced to brain states therefore, no matter how a system is functionally organized qualia are irrelevant. Libet (1979) demonstrated that control of action is present within the brain approximately 0.5 seconds before a subject is consciously aware of it. For example, if one thinks they would like to pick something up, neural activity will have already begun half a second previously of this thought occurring; this offers corroboration for consciousness and qualia being a product of physical states alone. If Greffenstette's theory were true functionalist theories would almost be forced to admit a dualism between neurons and qualia which seems paradoxical.

Prinz (2003) argues that it would be impossible to ever empirically prove that systems possessed consciousness as it remains nigh impossible to prove consciousness in anything. It is argued that properties imperative for conciousness are known as are properties sufficient for consciousness but any crossover between the two is unexplored. That is to say, computers are capable of possessing components that are known to be imperative but it is not known whether they will be sufficient to produce consciousness.
More recently Hales (2009) proposes an methodology for objectively testing consciousness in an artificial system involving a P Consciousness Scientist Test. The test takes place in a laboratory where the subject is placed, the subject is then expected to demonstrate the required lab/subject behaviour with the claim that only a conscious subject would recognise and adapt its behaviour to the setting. However, the test lacks validation from other peers and seems to be a flawed method of

measuring consciousness as consciousness is not just simply adapting to particular settings. The test would rely on observation which is not the strongest methodology. As previously demonstrated in the ELIZA programme (Weizenbaum 1966); many people originally failed to believe they were interacting with a machine; that there must be a live person at the other end of the network. Thus, computers may be capable or superficially demonstrating the correct behaviour when they do not, so Hales paradigm appears to be flawed.

Georgiev (2004) states that if Chalmers' theory is true, conscious is a fundamental asset to human life but fundamentally meaningless. Georgiev refutes epiphenomenalism that consciousness is solely a product of brain processes but has no causal effect on processes. He argues that for Chalmer's theories to be correct infinite numbers of ad hoc psychophysical laws must be present resulting in infinite regress which is scientifically unacceptable.

Against Functionalism:

Izhikevich (2005) developed a full size synthetic brain model with the equivalent of 10^11 neurons and 10^15 synapses; the model exhibited known brain frequencies such as alpha and gamma rhythms. It took 50 days to recreate 1 second of brain dynamics; although slow this research proves it is possible to recreate models of the brain. But could it be argued that in those 50 days that brain dynamics were being recreated that the model had consciousness? No tests were carried out to attempt to demonstrate consciousness, but according to functionalist theories theoretically the system should have possessed consciousness.

John Searle (1990) proposed a famous thought experiment "The Chinese Room". Searle supposes if a man were shut inside a room and Chinese characters were slid underneath the door of the room, with the correct instructions and equipment the man could process them and produce Chinese characters as an output. To an outsider, it would appear that the system understands Chinese, yet the man himself, responsible for the output is not a Chinese speaker. Searle extends the argument to computers and machines with Strong Artifical Intelligence; the man in the room cannot understand Chinese but produces a convincing simulation; thus a computer must also produce only a simulation.

However; this theory in itself is controversial.

It is argued that the whole room and everything in it acts a uniform system. Thus the

whole room and its contest could be said to understand chinese, the honus is not on the man. It is argued that the room has typical Von Neumann architecture (the structure of modern computers). For example there is a program (the instructions), there is memory (the paper on which to write the characters), a central processing unit (CPU) to follow the instructions (the man), and a method to transcribe the symbols (the pencil). Thus the “systems reply” (Russel and Norvig 2003) is that the man not understanding Chinese is irrelevant; the whole system can understand the Chinese.

Similarly, there is the “virtual mind reply” (Minksy, 1980) in which a mind appears to exist with in the computer because of the software. For example, there are only electrical components inside the body of a computers, yet they have “virtual” objects such as files. Similarly, Minsky argues, a computer may posses a virtual mind in the same way a brain does; if you were to look inside a brain you would only find physical components.

Lastly, there is the “robot reply” (Cole ,2004). If the room was allowed “eyes” such as cameras to read the input and “hands” to provide the output there would be a "causal connection" between the symbols and their representation; thus reliance on a human to input the symbols under the door would be eliminated as would the need for a human to output the symbols by drawing them. The system would solely rely on the outside world and physical states to produce the output thus constituting a mind.

Furthermore, Searle (1980) states that a key distinction between the human mind and a machine is that humans have the capability of intentionality from causal relations between mental processes and the brain. Searle states that a machine could never spontaneously enter into a state of intentionality and that attempts to create intentionality (Strong AI) will never succeed by the mere designing of programmes. However, as previously discussed Libet’s (1979) discovery of neuronal firing before conscious awareness may debunk this theory somewhat.

Interestingly, successful Strong AI attempts would need to incorporate common sense knowledge (Russell and Norvig 2003). However, humans are known to violate common sense norms. For example, in the case of Monty Hall’s Three Door Problem in which there are three doors behind which prizes lie, competitors are asked to pick a door (there is a 1/3 chance of the prize lying behind their selected with 1/3 chance of the prize being behind the second door and a further 1/3 chance it is behind the third door), the host of the game then eliminates one of the two doors the competitor

has not picked, the probability of the prize being the competitor's door remains at 1/3, yet the probability of it being behind the other door has now risen to 2/3, if the competitor makes the rational choice that will maximise their expected utility they will opt to switch their selected door. Various studies have been conducted using the 3 door paradigm. For example Franco-Watkins' Derks, Dougherty (2003). Results revealed that participants make a probability judgement reflecting the number or value of available prizes over the actual probability regarding the number of unopened doors. Furthermore, Slembeck and Tyran (2004) found that most people initially fail to apply Bayes' Law of Probabilities correctly; this can be attributed to failure of understanding the nature of the problem but also to the sense of regret a competitor might feel if they originally had the door with the prize and then switched and won nothing. Would a rational computer system be able to tie in emotion to make a decision that violates common sense norms. I personally do not believe it could, if Morevac and Crevier are to be believed computers could possess emotions but only as a means to an end in terms of survival of the fittest thus it would be in the artificial agent's interest to always opt to maximise their utility. As Bates (1994) points out; it is quirks that provide personality and personality that provides life as it were. Thus, artificial agents would need to allow for regularities and norms to be broken. It is my opinion that a rational system derived from mathematical abstraction would be incapable of violating such norms.

The Human Brain: Mind or Machine?:

The issue of consciousness being the unique component of a human mind is contentious. Firstly, it is almost impossible to prove consciousness. If dualism is to be believed the solution of what defines consciousness could never have its solution based in physical science and through its own abstract nature could never be empirically proven because as Descartes identified thought does not have extension and therefore no basis of matter. However, dualism is almost a school of thought reliant on faith. Arguably, the human mind must have a physical basis as it is highly implausible to suggest that a mind can exist as a separate entity, not rooted in anything solid. Furthermore, it seems redundant to rely on century old philosophies and dismiss the advances in science and technology since this time. So dualism as a workable theory remains incomplete.

Chalmers' makes a convincing case for the functionalists systematically highlighting

the implausibility of the qualia objections, however it is not without its pitfalls. Importantly, he does not deny the existence of qualia (only certain types of qualia); to have a functionalist theory so firmly based in the physical and then to subscribe to the more abstract idea of qualia seems almost paradoxical. But if functionalism is not a viable theory, what is? It seems that, at present, there is no infallible theory. Computers may certainly evolve to realistically emulate consciousness but whether they could ever be competent to truly possess it is uncertain. Weizenbaum (1976) outlined the threat to human dignity that artificial agents may pose in such roles as a nurse because, as humans, we require genuine empathy from these societal roles. Furthermore, if an artificial agent could ever attain consciousness it would redefine the concepts of human rights. Additionally, he talks about “atrophy to the human “ spirit reducing the complexities and wonders of the human brain to mere mechanics. To date, attempts to recreate consciousness in artificial agents have been unsuccessful and no foolproof methodology has been created to effectively measure consciousness either. Perhaps it is the arrogance of human nature that leads us to believe that we are more than just mere machines; but as there is no solid causal link between the firing of neurons and conscious experience yet it can neither be disproven nor proven.

References:

Aleksander, Igor (1996), *Impossible Minds*, World Scientific Publishing Company

Bates, J. (1994) The role of emotion in believable agents *Communications of the ACM* 37 (7) 122–125

Bostrom, n. (2006) Quantity of experience: brain-duplication and degrees of consciousness *Minds and Machines* 16;2 pp 185-200

Block, N. (1978). Troubles with functionalism. In (Block, ed.) *Readings in the Philosophy of Psychology, Volume 1*. Cambridge, MA: Harvard University Press.

Bringsjord, S., Bello, P. and Ferrucci, D. (2001), “Creativity, the Turing Test, and the (Better) Lovelace Test,” *Minds and Machines*, 11: 3–27

Chalmers, D., (1995) Absent Qualia, Fading Qualia, Dancing Qualia. *Conscious Experience.* Imprint Academic

Cole, David (2004), "The Chinese Room Argument", in Zalta, Edward N., *The Stanford Encyclopedia of Philosophy*

Crevier, Daniel (1993), AI: The Tumultuous Search for Artificial Intelligence, New York, NY: BasicBooks pp. 48–50

Deregowski, J. B. (1989). Real space and represented space: cross-cultural perspectives. *Behavioural and Brain Sciences,* 12, pp. 51-119.

Franco-Watkins A., Derks P., Dougherty M., (2003) Reasoning in the Monty Hall problem: Examining choice behaviour and probability judgements *Thinking & Reasoning Volume 9,* Issue 1, 2003, Pages 67 - 90

Friel KM, Barbay S, Frost SB, Plautz EJ, Hutchinson DM, Stowe AM, Dancause N, Zoubina EV, Quaney BM, Nudo RJ (2005) Dissociation of the sensorimotor deficits after rostral vs. caudal lesions in the primary motor cortex hand representation. *J Neurophysiol* 94: 1312-1324

Hales C. (2009) An empirical framework for objective testing for P-consciousness in an artificial agent *Open Artificial Intelligence Journal 3, pp 1-15*

Izhikevich (2005) Human Brain Simulation http://www.izhikevich.org/human_brain_simulation/Blue_Brain.htm

Kurzweil, Ray (2005), The Singularity is Near, Viking Press p. 260)

Libet, B., Gleason, C. A., Wright, E. W., and Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*, 106:623-642.

Locke, John (1690) Essay of Human Understanding II Oxford: Oxford University Press

Georgiev, D. (2004) Chalmers' principle of organizational invariance makes consciousness fundamental but meaningless spectator of its own drama.*Phil Sci* http://philsci-archive.pitt.edu/id/eprint/1702

Greffenstette, E. (2006) Inverted/absent qualia and the problem of epiphenomenalism. *British Undergraduate Journal of Philosopy* 3 p 239,241

Minsky, Marvin (1980), "Decentralized Minds", *Behavioral and Brain Sciences* 3: 439–40

Moravec, Hans (1976), The Role of Raw Power in Intelligence

Prinz, J. (2003) Level-Headed Mysterianism and artificial experience *journal of consciousness studies* 10;4-5 pp 111 - 132

Russell, Stuart J. Norvig, Peter (2003), *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall p. 959,

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3, 417-57

Slembeck T., Tyran, J.R. (2004) Do institutions promote rationality?: An experimental study of the three-door anomaly *Journal of Economic Behavior & Organization* Volume 54, Issue 3, 337-350

Shoemaker, S. (1982). The inverted spectrum. *Journal of Philosophy*, 79, 357-81

Weizenbaum, Joseph (January 1966), "ELIZA — A Computer Program For the Study of Natural Language Communication Between Man And Machine", *Communications of the ACM* 9 (1): 36–45

Weizenbaum J (1976), quoted in McCorduck 2004, pp. 356, 374–376